

**Models of Flexible Workforces in Stochastic
Service Environments, the One-Job Case**

E. J. Pinker and R. C. Larson

OR 313-95

September 1995

Models of Flexible Workforces in Stochastic Service Environments, the One-Job Case

Edieal J. Pinker*

Richard C. Larson†

September 27, 1995

Abstract

In this paper we develop a framework for modeling the management of flexible workforces in services with stochastic work arrivals and absenteeism. We present a classification scheme for a family of workforce sizing models that take into account absenteeism, overtime, backlogging, working time and functional flexibility, and the timing of work arrival information. The working time flexibility feature of the models is represented in a general way by constrained overtime and call-in workers. We explicitly formulate models for work environments with only one job type and therefore ignore functional flexibility. The various models, in this paper, are formulated as optimization problems that determine the combined labor and backlog cost minimizing pool sizes for call-in and regular workers. Embedded within this optimization problem is a dynamic programming problem of making optimal daily staffing decisions with respect to the utilization of call-in and overtime resources. The models not only determine optimal pool sizes but provide managers with a tool for making optimal dynamic staffing decisions on a daily basis. We implement several of the models and demonstrate, with numerical examples, that there is a strong link between the benefits of different types of flexibility, the stochasticity of the work environment, and the availability of information for decision making.

*Operations Research Center, Massachusetts Institute of Technology

†Operations Research Center, Massachusetts Institute of Technology

It is now a well established fact that the U.S. economy has become dominated by services, 91.3 million civilian employment [Sta95]. The special characteristics of services, reviewed in [BL93b], and the increasing acceptance of novel work arrangements (see [Fie94], [Fod95]), demand a modern and innovative set of models for managing workforces. Specifically, the lack of inventories and the relatively short time scales involved in services mean that stochasticity in work arrivals and absenteeism require firms to find flexible approaches to utilizing workforces. Failure to do so will result in either high labor costs or poor service.

Within the labor economics literature several forms of flexibility are mentioned. We will follow the classification found in [Tre92] of numerical, working time, functional, and pay flexibility. Numerical flexibility governs the employers options in altering the size of their workforce and allows a firm to better match their workforce to workload, when there is workload variability over a time scale of months. Working time flexibility governs the scheduling of work hours and the number of hours worked, and allows a firm to better match its workforce to workload when there is workload variability from day to day or within days. Functional flexibility governs many organizational issues such as job definition, supervisory hierarchies, and internal mobility. This flexibility can increase the range of activities a workforce is capable of performing and therefore make it more adaptable to work demand variability. It can also be a necessary condition for a firm to be able to apply different operational techniques such as JIT and continuous improvement. Pay flexibility governs a firm's ability to tie wages more closely to the firm's economic performance and/or the worker's individual performance. Labor advocates have traditionally criticized "flexible" work arrangements as exploitative, portraying them as means to cheapen labor unfairly [HB90]. Most complaints focus on issues of job security, benefits and safety [Reb95] for flexible workers. In short, there is evidence that flexible workers, and particularly working time flexible workers are treated as second tier employees. This paper does not address these issues, and we make no assumptions about employer-employee relations. The flexible work arrangements we model are compatible with well compensated, long-term, and safe employment. Similarly, they are compatible with unfavorable labor relationships as well.

We are interested in the problem of *matching workforce size to workload in service environments with variability in workloads*, and focus our attention on those forms of labor

flexibility that directly relate to this problem, namely numerical, working time, and functional flexibility. The variability in a workload can be characterized by the degree to which it is stochastic and the time scale of the variability. We consider the following time scales: months, weeks, days, and hours. For each of these time scales there are examples of work environments in which there is deterministic or seasonal workload variability, examples in which there is stochastic variability, and examples of combinations of both. We use the term stochastic broadly to describe work arrivals that have a random component. We do not preclude correlation between work arrivals at different times. Each form of variability may benefit from a different form of labor flexibility.

Monthly variability in workload can be a result of unexpected changes in demand for a firm's services, a change in the type of services required or a seasonal effect. In all cases a firm may need to expand or contract its workforce for a period of one or more months but usually not more than 6 months¹. The problem of deciding how to adapt the workforce size in such situations was considered for manufacturers in [HM60]. The same approach applies for services without the ability to build inventory. In [HM60] the only option for adjusting workforce sizes is to hire or fire workers. Today there are other options commonly used to achieve the same effect such as, subcontracting excess work and contracting contingent workers². The ability to contract contingent workers is a form of numerical flexibility. When a firm faces a shift in demand from one mix of services to another it may be aided by functional as well as numerical flexibility. Crosstrained (or multiskilled) workers can be shifted, from a low demand job to a higher demand job, thereby reducing the firm's need to lay-off workers, from the low demand job, and its need to hire new workers, for the high demand job, both actions involving significant costs³. In these situations firms also use working time flexibility in the form of overtime and short-time to smooth their labor

¹We assume that the firm has already taken measures such as scheduling vacations and low priority tasks for low workload times of year.

²The term 'contingent worker' has been used in many different ways. We will use the definition in [Pol89]: "Any job in which an individual does not have an explicit or implicit contract for long-term employment or one in which the minimum hours worked can vary in a non-systematic manner."

³Crosstraining also involves significant costs and many other complications such as, assigning supervisory responsibility to workers who perform a variety of tasks, setting pay incentives, paying for training, and maintaining skill levels [Kle94].

requirements [Mic87].

When workload varies on a weekly, daily or hourly basis working time and functional are the most applicable forms of flexibility. It is not feasible to hire and fire on a regular basis from week to week, therefore numerical flexibility is not relevant in any significant scale⁴. When workloads vary deterministically working time flexibility is manifested in days-off scheduling and the use of part-time work. The benefits of such flexibility are demonstrated in [Bak73, BLP94, MR73]. The benefits of crosstraining (functional flexibility) for within-day variability is demonstrated in [BLP94, War72].

When workload varies stochastically within a day there is a practical limit to working time adjustments. Once a scheduled worker has arrived to work it is difficult to reassign them to different hours or to send them home. The most common working hour adjustment is to offer overtime hours. In mail processing plants managers can move the start time of a worker's shift up to two hours, but must pay an overtime premium for those shifted hours and must be able to make the adjustment before the worker has arrived. The most common phenomenon is for mail, the workload, to arrive in the predicted quantities, but at later than planned times. In these situations a large amount of overtime is usually utilized. Short-timing is a less common option and is used when workload is light. Functional flexibility can be beneficial for within day variability if the variability shifts work from one job to another. This occurs in stochastic flow shops and the benefits of crosstraining for these environments has been shown in [Tre89].

When workload varies stochastically from day to day or week to week there is an opportunity to benefit from functional flexibility, as in the previous case, but there is also a significant opportunity to benefit from working time flexibility. The primary vehicle for this benefit is the call-in worker. Call-in workers are workers who are *called in* to work on short notice when there is a need for them. A firm employing call-in workers would draw upon two sources of labor. The primary source, is the pool of regular staff who have fixed schedules and are permanent employees of the firm. The second source would be call-in workers who do not have preset schedules, i.e. workers with flexible working time. These

⁴It is feasible, and common, to use temporary workers to cover for absences and unfilled positions.

workers could be permanent employees or contracted from an external temporary employment agency. In either case the firm would be required to guarantee some minimum pay over the employment period to each call-in worker. If the workers are called in more than the guaranteed amount they would be compensated accordingly. An aggregate model of this approach, with no backlogging of work from day-to-day, is presented in [BL93a] and [BL94].

An interesting hierarchical model for planning the use of long-, medium-, and short-term flexibilities is presented in [WR93]. In this paper an attempt is made to model the use of short-timing, temporary workers, overtime, and crosstrained floaters in a service environment with stochastic workloads and absenteeism. The classification of the model as service based is used to justify the lack of inventories *and/or* backlog. Furthermore it is assumed that there is an endless supply of temporary workers. These assumptions together lead to a formulation in which the daily staffing decisions are independent of one another. This approach is successful in modeling the impact of long-term planning decisions on the availability of resources for short-term staffing decisions but fails to accurately represent the short-term system behavior.

There are many work environments that would fall under the category of services with day-to-day workload variability. For example, in the transaction processing center of a mutual fund company the amount of work that arrives in the mail each day is highly variable. This workload is subject to variability in mail service, economic trends, market events, and response to marketing promotions. Customer's requests for transactions must be processed very quickly, preferably on the day of receipt, and obviously cannot be processed before they arrive. Staffing to the average workload leads to the accumulation of backlogged work and poor customer service. Staffing at a level that is sufficient, with high probability, is very costly and involves a low utilization of the workforce. Another example is a hospital. Patients arrive to an emergency room randomly each day and, as many medical services are distributed to clinics and centers, a higher proportion of hospital workloads are emergency cases and therefore less predictable and manageable. The hospital cannot afford to be understaffed since patients must receive their treatments in a timely manner. On the other hand, medical personnel are highly trained and expensive, therefore overstaffing can be a

great financial burden.

The use of call-in workers has many potential benefits. For the employer, call-in workers may provide an expanded labor capacity with a lower cost than an equivalent labor capacity composed of entirely full-time regular workers. Call-in workers also provide a consistent source of expanded capacity making it easier for a firm to maintain consistent workforce performance standards. For employees, the use of call-in workers will reduce the amount of overtime worked and should reduce the overall amount of wage received per worker employed by the firm. On the other hand the call-in worker arrangement suggests a long term commitment of employment. Furthermore, the reduction in wages paid come with an increase in leisure time for regular workers who work less overtime and for call-in workers who do not work full-time hours. Although the uncertainty in schedule is a negative factor, overall a call-in arrangement offers a stable source of income for workers who do not want to work full-time.

Some employers try to reap these benefits today but in very adhoc ways. The USPS has a class of employees called casual employees that are not required by union contract to have regular schedules. These workers are utilized to fill staffing gaps in operations caused by absenteeism or vacations etc. [Ser94]. In some financial services companies part-time data entry clerks are informally promised 16 hours of work each week and are sent home early without pay if workload is light, this is an example of short-timing⁵. In neither case has a systematic analysis been done to determine staffing needs in light of the existing scheduling flexibility⁶.

The financial services and hospital examples described previously have several characteristics in common that are not modelled comprehensively in any of the existing literature:

- Workload is stochastic.

⁵The limitation of this approach is that short-timed workers have already come in to work and must be compensated for the inconvenience.

⁶In practice employers commonly rely on temporary personnel services to provide call-in workers. However, from our perspective this is just one extreme of the flexibility spectrum. Such temporary workers are call-in workers who receive no commitment from their employer beyond the single continuous period of utilization.

- Workload can be backlogged.
- There is no 'finished goods' inventory.
- The time scale does not allow for hiring/firing of workers.

The purpose of this paper is to fill this void in the literature by fulfilling the following goals:

- Formulate a new family of workforce management models that address the major trade-offs involved in determining staffing policies in a service firm with variable workloads.
- Provide managers with tools for setting and testing staffing policies.
- Demonstrate the potential utility of call-in workforce arrangements and illuminate their dynamics.

The remainder of this paper is organized as follows. In section 1 we present a classification scheme for distinguishing different work environments and thereby define a family of new modern workforce management models. In section 2 we formulate a representative group of one-job models. In section 3 we perform some analysis of these models. In section 4 we present some numerical results demonstrating the use of these models and the potential benefits of workforce flexibility. Finally in section 5 we present some extensions and discuss conclusions.

1 Definitions and Problem Classification

We develop a general set of models for workplaces similar to the financial services and hospital examples described previously. We assume that all of these workplaces follow operating policies with the following general structure: The firm processes all work that arrives each day with only regular staff. If there is excess work a decision is made to draw upon non-regular labor capacity or not. Excess capacity is drawn from the call-in

workforce and overtime hours. If excess work remains, it is backlogged. Different workplaces are distinguished from one another by the following characteristics: notification, backlog tolerances, workload distributions, absenteeism, call-in worker contracts, crosstraining, and their cost structures. In this section we define what we mean by the above distinguishing characteristics of workplaces:

- **Period** - We model the problem using a general time unit called a period. This can be considered to be a day, a week, etc. On the level of a period we make decisions about drawing upon call-in workers or overtime and how much backlog to allow. The day is the time unit for which the problem is most naturally defined but we do not restrict ourselves to this scale.
- **Planning Horizon** - A planning horizon is some number of periods for which we make the decision of how to size the workforce.
- **Notification** - The information available when call-in workers are informed, that they will be needed in any particular period, is strongly dependent upon the stochastic process governing the exogenous work arrivals. If a manager has good knowledge⁷ in period 1, of what the workload will be in period 2, she can call in workers in period 1 for period 2. In this situation we assume it is reasonable to expect the call-in workers to be able to come in. If a manager's information about the period 2 workload is incomplete until the start of period 2 it is less likely that she can make the call-in worker utilization decision for period 2 with full workload information. Based upon this reasoning we will consider two different notification scenarios. The first scenario occurs when the manager has workload information early enough to contact call-in workers for the same period that workload is expected. The second scenario occurs when the manager only receives complete workload information when it is too late to call in workers for the same period as the workload arrives. In this scenario call-in worker's utilization will be decided in one period for the next-period with incomplete workload information. To summarize we will consider two notification schemes: **same period**, and **next period**.

⁷In reality there will always be uncertainty in the workload so we consider knowledge of workload to be a low variance estimate of workload.

- Backlog and Backlog Tolerances - Backlog is the work that is carried from one period to the next. Any firm may have specific tolerances for backlog which we encode in the backlog penalty function.
- Workload Statistics - We characterize new work arrival distributions by considering the following two questions: 1) Are the workloads in each period identically distributed, and 2) are the period workloads independent?
- Absenteeism - Another source of variability in a workplace is absenteeism. We consider the cases when there is no absenteeism and when there is absenteeism in one or both of the regular and call-in staffs.
- Call-in worker contracts - We view call-in workers as being contracted for a planning horizon. These workers could be regular staff that are temporarily placed into this scheduling category, permanent call-in workers, or externally contracted temporary workers. In any event we assume that call-in workers are guaranteed a minimum number of payed work periods for the contracted planning-period. If call-in workers are needed more than the guaranteed number of periods then they must receive additional compensation.
- Crosstraining - In firms where there are multiple jobs requiring different worker qualifications, crosstraining can be used to increase worker management flexibility. If the workload of one job is not strongly correlated with the workload at another job, it is possible that having some workers who are crosstrained to perform each job will reduce labor costs.
- Shift - A shift is the number of hours worked within a period without requiring overtime pay.
- Cost Structures - In formulating the cost functions for the models we take into account the following sources of labor costs: Benefits costs for regular and call-in workers, hourly wages for all workers and overtime wage premiums. We also assess a penalty for carrying backlog.

While all the above characteristics may distinguish work environments we focus on a subset that most strongly defines the structures of the models we develop in this paper. These characteristics are: Notification and Crosstraining, and Backlog policy. We use these characteristics to define a family of workforce management problems. This family is outlined in the form of a tree in Figure 1.

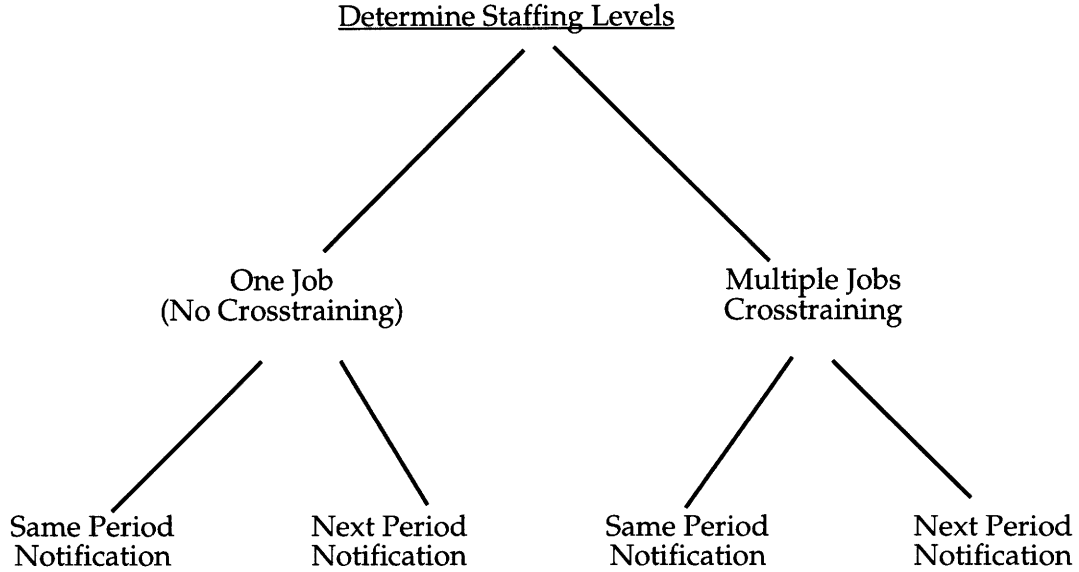


Figure 1: Tree of Workforce Problems

On the most general level, the problem we want to solve is: *What are the optimal staffing levels that would minimize the sum of labor and backlog penalty costs?* On the next level, we model one job and multiple jobs with no crosstraining as a single problem type. Environments with multiple jobs and crosstraining require a distinct model. On the lowest branching level, on the problem tree, we distinguish between the two different notification regimes, same-period and next-period. Within the context of each of the problems, on the tree, we will have to consider the effects of absenteeism and non-homogeneous and/or dependent work arrival distributions.

2 Formulations

In this paper we are focusing on the one-job branch of the problem tree. In this situation the manager must decide:

Problem 2.1 *How many regular workers N and call-in workers M to staff for a planning horizon of length V time periods, so that the expected labor and backlog costs are minimized over the entire planning horizon.*

We initially assume that there is no absenteeism and that the new work that arrives each period is independent and identically distributed to all other periods and formulate the same-period and next-period notification problems. We then formulate both problems with absenteeism, in doing this we show how absenteeism can be manifested in several different ways requiring distinct formulations.

The problem formulations all address problem 2.1 and assume that in each period managers have two decisions to make, how many call-in shifts to utilize and how many overtime shifts to utilize. These decisions are assumed to be made optimally and dynamically. The major difference between the different problems formulated in this chapter is the information available about workloads and absenteeism when these two decisions must be made. This means that for each problem in this branch of the problem tree we formulate a different dynamic program for the period by period decision making.

In all the formulations we define x_t to be either the amount of work in the system in period t , or the amount of work in excess of the staffing available the period. This work has two components, new work that arrived at the beginning of period t , which we call d_t and work left in the system at the end of period $t - 1$. For any given fixed staffing policy we view $\{x_t\}$ as a Markov process. The transitions from state to state are driven by the iid. work arrival process $\{d_t\}$ and staffing decisions made in each period.

In each period the amount of work that can be processed is determined by the number of regular workers present, the number of call-in workers utilized and the number of overtime shifts utilized, any excess is backlogged to the next period. In the different problems the

timing of the call-in and overtime decisions may be different but the trade-offs involved in these decisions are the same. Let's consider these trade-offs here.

If we do not utilize call-in or overtime work and the workload in a period t does not exceed the available regular staff's capacity there are no backlog costs and no new staffing costs, since the cost of the N regular workers is a sunk cost. If the workload exceeds the regular staff's capacity we have backlog costs unless a sufficient number of call-in and/or overtime shifts are utilized. I.e. we must tradeoff the cost of backlog with the cost of overtime and/or call-in worker utilization. The cost of overtime is a linear function of the number of hours utilized in period t . The cost of call-in workers is dependent upon the cumulative use of call-in workers up until period t ; because, we contract M call-in workers for the planning horizon with a guarantee of a fraction G , of V , paid periods of work in the planning horizon. The payment for GV per call-in worker periods of work is a sunk cost and any periods worked in excess of GV periods per call-in worker incurs extra costs. This means that the overtime/call-in decision, in each period t , is a dynamic decision based upon cumulative call-in utilization, expected future utilization, and future backlogs.

To make these various costs more tangible we now define the cost parameters that are used in all the problem formulations. First, we recall our definition of the *shift* as the number of hours a worker works within a time period at their ordinary wage⁸. Second, we only define labor costs over the course of a single planning horizon (composed of an arbitrary number of time periods). Third, we assume that each worker has three components to their compensation: A salary received for each shift worked, a benefits component for each shift worked, and a fixed component for being part of the workforce during the planning period in question. The cost parameters we use in the formulations are as follows:

C_f = fixed cost for each worker, call-in or regular that is a member of the workforce for the planning period. It includes fixed component of compensation and fixed costs per worker for the firm. E.g. human resource department costs, services available to all employess regardless of status, etc.

C_{rw} = per-shift cost of a regular worker that combines the benefits and salary.

⁸Later we introduce the constant π as the number of units of work processed per shift of work.

C_{cw} = per-shift cost of a call-in worker that combines the benefits and salary.

C_{ot} = the premium paid per-worker per shifts worth of overtime worked.

C_b = the per-time period penalty incurred by the firm for every unit of work backlogged.

C_B = the penalty incurred by the firm for every unit of work backlogged in the final period of the planning period.

We have assumed linear costs of C_{ot} for overtime shifts worked and C_{cw} for call-in worker shifts paid for. We have specified linear backlog penalties C_b and C_B . Note: We could formulate the models with an arbitrary functional form for backlog penalties. We use linear penalties here to be consistent with the numerical examples and analysis sections. Having understood these trade-offs and cost parameters we can now state the problem more completely as:

Problem 2.2 *What staffing level $S = (N, M)$ should a firm contract, over a planning horizon of length V , to minimize the expected labor and backlog cost incurred when optimal call-in/overtime decisions are made dynamically each period, if the call-in workers are guaranteed at least a fraction G , of V , paid periods of work per planning horizon.*

The implication of this problem statement is that for each problem in the one-job branch of the problem tree the expected cost of a staffing level S is the expected cost of the optimal solution to a finite horizon, labor allocation, dynamic problem. We formulate problem 2.2 as a mathematical program:

P1

$$\min_S C_{rw}NV + C_{cw}MGV + C_f(N + M) + f_1^S(x_1, \kappa_1)$$

subject to:

$$S \geq (0, 0)$$

$$S \text{ integer}$$

where C_f is a fixed cost for each worker, call-in or regular, and C_{rw} and C_{cw} are respectively per-period costs for employing regular and call-in workers that include hourly wages and pro-rated benefits. The expression $f_1^S(x_1, \kappa_1)$ is the expected cost of making optimal dynamic staffing decisions over the planning horizon given a staffing level S , with x_1 units of work in the system, and κ_1 guaranteed shifts of call-in workers unused in period 1.

2.1 Same-period notification, no absenteeism

Definitions

The problem has V stages that are equivalent to periods.

The state of the system at each stage t is given by the vector (κ_t, x_t) .

κ_t is unused portion of the total call-in worker guarantee MGV at the end of stage $t-1$.

x_t is defined to be the workload in the system at the start of stage t . This work is composed of new work that can be thought of arriving between stages $t-1$ and t which we will call d_t , and work left in the system at the end of stage $t-1$.

Decisions

u_t is the number of call-in workers utilized in stage t .

ω_t the number of overtime shifts utilized in stage t .

Constraints

$$u_t \leq M$$

$$\omega_t \leq OT_{max}(N, u_t)$$

Where $OT_{max}(N, u_t)$ is a function representing the maximum amount of overtime that can be performed by a complement of workers (N, u_t) . I.e. N regular workers are always available and u_t call-in workers are on site in stage t .

Transitions

$$\kappa_t = [\kappa_{t-1} - u_{t-1}]^+$$

$$x_t = [x_{t-1} - \pi(N + u_{t-1} + \omega_{t-1})]^+ + d_t$$

where π is a constant that converts units of people-shifts to units of work. The geometry of the state space is depicted in figure 2.

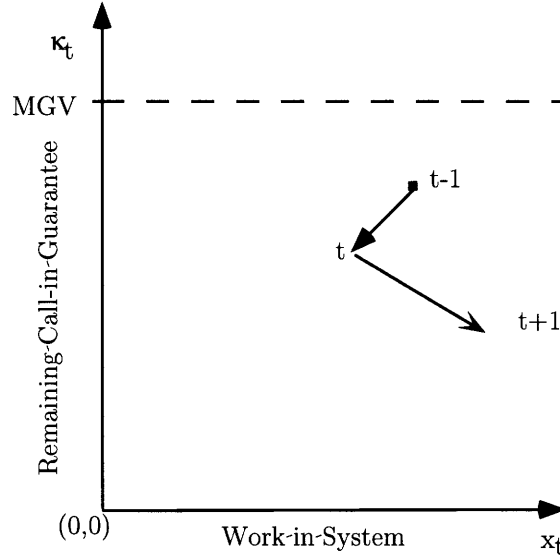


Figure 2: State Space Schematic

Costs The costs incurred in stage t are the costs of backlog, the cost of overtime and a cost for call-in worker usage in excess of the guarantee. The cost at stage t is given by:

$$C_{cw}[u_t - \kappa_t]^+ + C_{ot}\omega_t + C_b(x_t - \pi(N + u_t + \omega_t))$$

To save space we define: $b_t = \text{Max}(x_t - \pi(N + u_t + \omega_t), 0)$. The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at

the end of the planning period and is defined as:

$$C_V(b_V) = C_B(b_V)$$

The cost to go function in stage V is:

$$f_V(\kappa_V, x_V) = \text{Min}_{u_t, \omega_t} \{C_{cw}[u_t - \kappa_t]^+ + C_{ot}\omega_V + C_V(b_V)\}$$

and in stage t:

$$f_t(\kappa_t, x_t) = \text{Min}_{u_t, \omega_t} \{C_{cw}[u_t - \kappa_t]^+ + C_{ot}\omega_t + C_b(b_t) + E_{d_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

The interpretation of this equation⁹ is: If you are in state (κ_t, x_t) in stage t, the expected cost you incur from stage t until the end of the planning period is the minimum expected cost over all decisions (u_t, ω_t) of the overtime and backlog cost in stage t, plus the cost that is incurred from stage $t + 1$ until the end of the planning period, when following an optimal policy. When we solve this dynamic program we start with $f_V()$ and work back to $f_1()$.

2.2 Next-period notification, no absenteeism

In the previous section we have assumed that all staffing decisions (i.e. how many call-in workers and overtime shifts to utilize), for a period, have been made with perfect information about the workload in that period. In this section we consider the scenario where we do not have perfect information about the workload when we make the decision to call-in workers. This situation can arise because, for example, we must notify call-in workers by 5pm on Tuesday if we want them to work on Wednesday. It can also occur, for example, when we can notify call-in workers by 7am Wednesday morning that we want them to work that day, but do not have perfect knowledge of the workload for Wed. by that time. However, next-period notification describes all situations in which call-in worker decisions are made for a time period with incomplete knowledge of workload for that period. In this paper we

⁹The notation used in this formulation is based upon the conventions of [Ber87].

only consider next-period notification scenarios in which exogenous work arrivals in each period are independent and identically distributed. Therefore, information about workload is unchanged until a new work arrival occurs.

Within this scenario we can consider two subcases, relating to when we make overtime decisions. Subcase (i): Overtime must also be allocated before we have perfect workload information. Subcase (ii): Overtime may be allocated after information has become available.

I.e. In subcase (i) overtime and call-in decisions are made at the same time and in subcase (ii) the overtime decision may be made later. These two cases arise in practice in the sense that in many workplaces managers may not require overtime and therefore must seek volunteers. These volunteers are more likely to be found if sought out earlier as represented in subcase (i). On the other hand there are workplaces in which employers may require overtime (within some limits) and therefore do not have provide notice to the employees as is represented in subcase (ii). In this paper we only present the formulation of the second subcase.

In this subcase we make the assumption that the timing of the call-in decision for a period and the overtime decision for that period are separated by the arrival of information about the workload (see Figure 3). To accommodate this we redefine the start and end of a period. A period will begin with the exogenous arrival of new work followed by the overtime decision for that period and call-in decision for the next-period (see Figure 4).

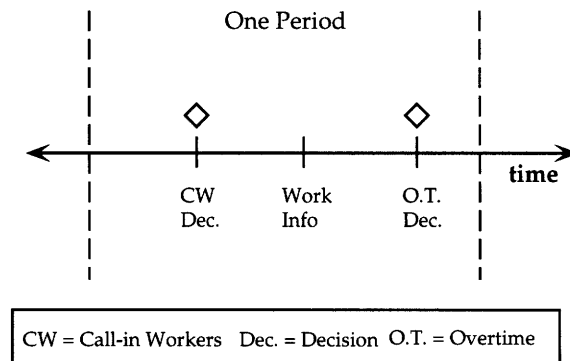


Figure 3: Order of events with next-period notification

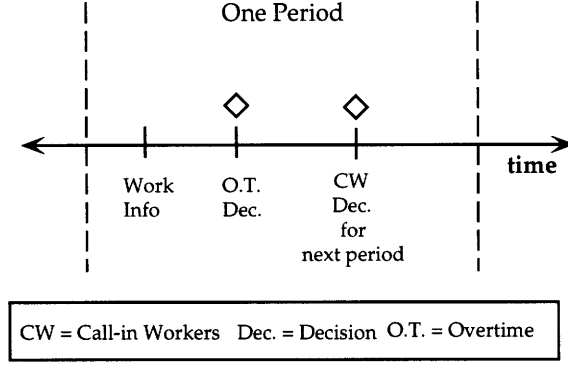


Figure 4: Order of events for next-period notification with redefined periods

Definitions We define a two dimensional state space: (x_t, κ_t) where:

κ_t is unused portion of the total call-in worker guarantee MGV at the end of stage $t - 1$.

x_t is defined to be the workload in the system, in excess of the call-in workers on hand, at the start of stage t . x_t may be negative.

We define a single random variable d_t for each stage t as, the amount of exogenous work arriving to the system at the start of stage t .

Decisions We have two decision variables: u_t , the number of call-in workers utilized in stage t , and ω_t the number of overtime shifts utilized in stage t .

Constraints

$$u_t \leq M$$

$$\omega_t \leq OT_{max}(N, M)$$

Where $OT_{max}(N, M)$ is a function representing the maximum amount of overtime that can be performed by a complement of workers (N, M) . I.e a standard limit on overtime has

been set based upon the total staffing level¹⁰.

State Transitions

$$x_t = [x_{t-1} - \pi(N + \omega_t)]^+ + d_t - \pi u_{t-1}$$

$$\kappa_t = [\kappa_{t-1} - u_{t-1}]^+$$

Cost Functions The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_V(x_V - \pi\omega_V) = C_B(x_V - \pi\omega_V)$$

The cost-to-go function in stage V is:

$$f_V(\kappa_V, x_V) = \min_{\omega_V} \{C_{ot}\omega_V + C_V(x_V - \pi\omega_V)\}$$

in stages t :

$$f_t(\kappa_t, x_t) = \min_{\omega_t, u_t} \{C_{ot}\omega_t + C_b(x_t - \pi\omega_t) + C_{cw}[u_t - \kappa_t]^+ + E_{d_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

2.3 Same-period notification with absenteeism

We now consider the staffing problem when there is absenteeism among the regular and call-in worker pools. There are several different scenarios for how absenteeism can affect the decision making process. We characterize these scenarios by the relative timing of absence information and staffing decisions. In all scenarios we assume that absences among regular workers are independent of absences among call-in workers and that absences are independent of workload. As in the no absenteeism case, we always have two staffing decisions to make for each period, namely how many call-in workers to use and how many

¹⁰In theory basing the overtime limit on the total staffing level allows for situations in which overtime is assigned in amounts that require more call-in workers to be present than are actually utilized that day. In practice this anomaly is only significant for extreme problem parameter choices.

overtime workers to use. We model the multiple decision points by splitting each exogenous work arrival cycle, or time period, into two stages for the DP staffing engine. Therefore, while we considered a V stage problem in the no absenteeism case we now consider the problem to have $2V$ stages with stage 1 being the first stage.

Case (i) In this case we assume that we make the call-in decision when we know the workload and regular worker absenteeism. We then assume that we make the overtime decision when we know the call-in worker absenteeism. In this case we are assuming that regular workers are giving some notice about their absence and that we know this information when we start to solicit call-in workers. We then solicit call-in workers until we find the amount we want or have exhausted the available ones. At this point in time we know how many regular workers and call-in workers are actually at work and how much work there is and make the overtime decision. The order of events is depicted in figure 5.

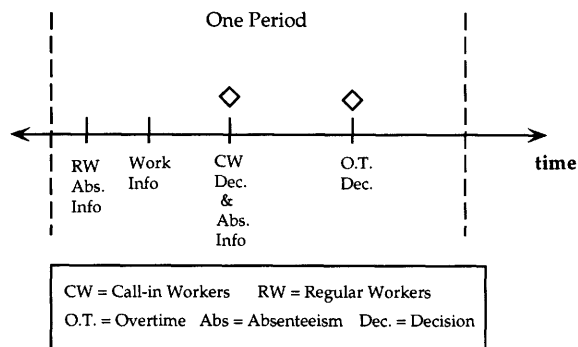


Figure 5: Order of events for same-period notification with absenteeism case (i)

Definitions We define a two dimensional state space: (x_t, κ_t) where:

κ_t is unused portion of the total call-in worker guarantee MGV at the end of stage $t - 1$.

x_t is defined to be the workload in the system at the start of stage t .

We define the following random variables in each odd numbered stage:

n_t is the number of regular workers who are present in stage t .

d_t is the amount of exogenous work arriving to the system in stage t .

We define $s_t = d_t - \pi n_t$ to be the newly arrived staffing 'shortage' for stage t . I.e. if s_t is positive it means that the new work arriving to the system exceeds the regular staff. If s_t is negative it means that there are more regular workers present than needed for the new work and therefore the excess staff can work on the backlog, if any, from the previous day.

For the even-numbered stages we define the random disturbance: m_t as the number of call-in workers who are available in stage t .

Decisions As with the random variables we have different decisions variables for odd and even stages.

If t is odd we make the decision: u_t , the number of call-in workers utilized in stage t .

If t is even we make the decision: ω_t the number of overtime shifts utilized in stage t .

Constraints

$$u_t \leq M$$

$$\omega_t \leq OT_{max}(\bar{n}_t, \bar{m}_t)$$

Where $OT_{max}(\bar{n}_t, \bar{m}_t)$ is a function representing the maximum amount of overtime that can be performed by a complement of workers (n_t, m_t) . I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available¹¹.

¹¹This means that on days with a lot of absenteeism we could ask workers to perform more overtime per person than on days with less absenteeism as long as the overtime does not exceed a limit based upon the expected number of regular workers present. This situation is fairly realistic.

State Transitions For odd stages $t + 1$:

$$x_{t+1} = [[x_t - \pi\omega_t]^+ + s_{t+1}]^+$$

$$\kappa_{t+1} = \kappa_t$$

Note: We are defining the workload at the start of an odd stage as the work beyond the processing capability of the regular staff.

For even stages $t + 1$:

$$x_{t+1} = [x_t - \pi \min[u_t, m_t]]^+$$

$$\kappa_{t+1} = [\kappa_t - \min[u_t, m_t]]^+$$

Cost Functions The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_{2V}(x_{2V} - \omega_{2V}) = C_B(x_{2V} - \omega_{2V})$$

The cost-to-go function in stage 2V is:

$$f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \{C_{ot}\omega_{2V} + C_{2V}(x_{2V} - \omega_{2V})\}$$

and in even stages t :

$$f_t(\kappa_t, x_t) = \min_{\omega_t} \{C_{ot}\omega_t + C_b(x_t - \pi\omega_t) + E_{s_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

and in odd stages t :

$$f_t(\kappa_t, x_t) = \min_{u_t} \{E_{m_{t+1}}[C_{cw}[\min[u_t, m_t] - \kappa_t]^+ + f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

Case (ii) In this case we assume that we make the call-in decision without knowing the regular worker absenteeism, only the workload. We then assume that we make the overtime decision with complete information about call-ins, regulars, and the workload. In this case we are assuming that the regular workers do not give notice of absences and just don't show

up. I.e. when we find out how many regular workers are present it is too late to solicit more call-in workers.

We can see from the timeline in figure 6 that as in case (i) there are two decision points, one for call-in workers and one for overtime workers. We again model this by splitting each exogenous workload arrival cycle into two stages for the DP staffing engine. We define the same two dimensional state space as in case(i): (x_t, κ_t) .

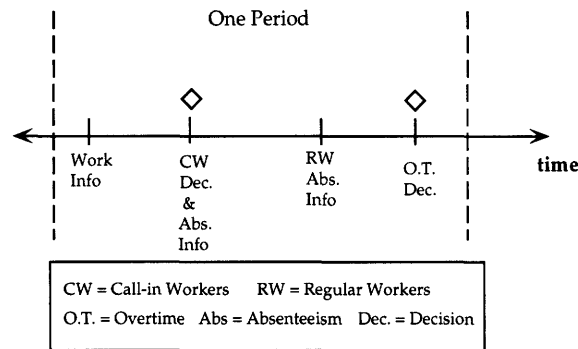


Figure 6: Order of events for same-period notification with absenteeism case (ii)

Case (iii) In this case we assume that we make the call-in decision without knowing the regular worker absenteeism and without knowing the call-in worker absenteeism but make the overtime decision with complete information. The implication here is that the regular workers do not give notice of their absences and that the call-in workers who are absent are ones that have said they would come in to work but do not show up.

We can see from the timeline in figure 7 that as in case (i) there are two decision points, one for call-in workers and one for overtime workers. We again model this by splitting each exogenous workload arrival cycle into two stages for the DP staffing engine.

2.4 Next-period notification with absenteeism

In this case we assume that we make the call-in decision when we know the regular worker absenteeism but before we know the new work arrival. We then assume that we make the overtime decision when we know the call-in worker absenteeism and the workload situation.

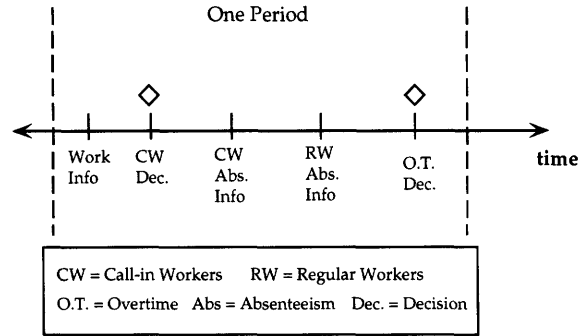


Figure 7: Order of events for same-period notification with absenteeism case (iii)

In this case we are assuming that regular workers are giving some notice about their absence and that we know this information when we start to solicit, call-in workers. We solicit call-in workers until we find the amount we want or have exhausted the available ones. At this point in time we know how many regular workers and call-in workers will actually be at work and how much work there is and make the overtime decision. The order of events is depicted in figure 8.

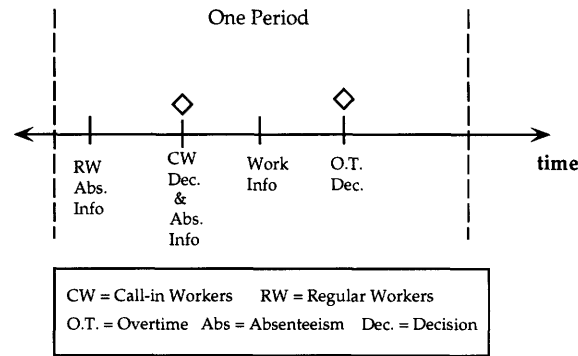


Figure 8: Order of events for next-period notification with absenteeism case (i)

We define a two dimensional state space: (x_t, κ_t) where:

κ_t is unused portion of the total call-in worker guarantee MGV at the end of stage $t - 1$.

x_t is defined to be the workload in the system at the start of stage t .

We define the following random variables in each odd numbered stage:

n_t is the number of regular workers who are present in stage t .

d_t is the amount of exogenous work arriving to the system in stage t .

For the even-numbered stages we define the random variable: m_t as the number of call-in workers who are available in stage t .

Decisions As with the random variables we have different decisions for odd and even stages.

If t is an odd stage we make the decision: u_t , the number of call-in workers utilized in stage t .

If t is an even stage we make the decision: ω_t the number of overtime shifts utilized in stage t .

We place the following constraints on the decisions:

$$u_t \leq M$$

$$\omega_t \leq OT_{max}(\bar{n}_t, \bar{m}_t).$$

Where $OT_{max}(\bar{n}_t, \bar{m}_t)$ is a function representing the maximum amount of overtime that can be performed by a complement of workers (n_t, m_t) . I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available¹². To simplify notation we define $a_t = \min[u_t, m_t]$ as the actual number of call-in workers utilized in stage t .

State Transitions For odd stages $t + 1$:

$$x_{t+1} = [x_t - \pi\omega_t]^+ - n_t$$

¹²This means that on days with a lot of absenteeism we could ask workers to perform more overtime per person than on days with less absenteeism as long as the overtime does not exceed a limit based upon the expected number of regular workers present. This situation is fairly realistic.

$$\kappa_{t+1} = \kappa_t$$

Note: x_t may take on negative values for odd t . If x_t is positive it means that the backlogged work in the system exceeds the regular staff processing capacity. If x_t is negative it means that there are more regular workers present than needed for the backlog currently in the system and therefore the excess staff can work on the new work yet to arrive that period.

For even stages $t + 1$:

$$\begin{aligned} x_{t+1} &= [x_t - \pi a_t + d_t]^+ \\ \kappa_{t+1} &= [\kappa_t - a_t]^+ \end{aligned}$$

Cost Functions The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_{2V}(x_{2V} - \pi\omega_{2V}) = C_B(x_{2V} - \pi\omega_{2V})$$

The cost-to-go function in stage $2V$ is:

$$f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \{C_{ot}\omega_{2V} + C_B(x_{2V} - \pi\omega_{2V})\}$$

and in even stages t :

$$f_t(\kappa_t, x_t) = \min_{\omega_t} \{C_{ot}\omega_t + C_b(x_t - \pi\omega_t) + E_{n_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

and in odd stages t :

$$f_t(\kappa_t, x_t) = \min_{u_t} \{E_{m_{t+1}, d_{t+1}}[C_{cw}[a_t - \kappa_t]^+ + f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

The assumption that the systems starts with no backlogged work requires some discussion. The formulation clearly does not require this to be the case and we could also draw the starting work in the system from some probability distribution representing the backlog in the system at the end of V periods. We could generate such a distribution with numerical experiments. Alternatively we could set the terminal stage backlog cost very high to drive the system to empty itself. In the numerical results chapter we will discuss this issue more.

3 Analysis

In this section we analyze the optimization problems formulated in the previous section. In the analysis we attempt to achieve two goals. The first is to characterize the real-world work arrangements that can be represented by some of the special cases of the various problems. The second goal is to derive methods to simplify the computational burdens of solving the problems.

3.1 Special Cases

The models we have formulated are very general in many senses. We have not assumed any special probability distributions for the random variables in the model. We also have created a structure that can represent many of the realistic work force management arrangements that exist today. These arrangements are modelled by choosing special values for some of the different parameters in the models. We now demonstrate how this can be done.

Traditional workplace In the traditional, inflexible workplace there are only regular workers and overtime available to the managers. The problem of determining how many regular workers to staff and how to utilize overtime can be represented in our models by setting $M = 0$. The outer optimization becomes a problem of choosing the cost minimizing value for N and the dynamic decisions made within the inner optimizations involve managing the use of overtime to reduce backlog.

Temporary workers Today many workplaces rely on temporary agencies to supply them with labor when the regular workforce is insufficient. These temporary workers are only paid for the days they work. This arrangement can be represented in our models by setting $G = 0$. This means that the call-in workers are not guaranteed any days of work. This implies that there is no need to select a value of M . It is assumed that there is always some employment agency that can provide temporary help on short notice.

Short-timing/Flexitime In some workplaces (more commonly in Europe than the United States) management works out arrangements with workers in which the hours of regular fulltime employees are set over longer time spans than usual. For example, a fulltime worker is defined to be any worker that works 40 hours per five-day work week. This definition does not require that each worker work eight hours per day. Although, there might be a restriction that within any day a worker may only be required to work up to 10 hours. This form of flexibility can be represented by our models as well. If we consider call-in workers to be the flexible workers we can set $G = 1$ and only charge an overtime premium for work done in excess of the guarantee MGV .

Everything in between While we have shown the way the most common workforce management practices can be represented by our models it is important to note that these special cases are extremes. This suggests that what is done in practice is only a very limited representation of the arrangements available to a firm. Our models provide a framework for making decisions about a wide spectrum of work arrangements that form compromises between the special cases described above.

3.2 Computational Analysis

The way we have formulated the problem we have two nested optimization problems. The inner problem is the optimal, in an expected value sense, dynamic allocation of preset staffing resources over a finite planning horizon. The outer problem is the determination of the staffing resources to have available to minimize the cost of the inner problem plus the cost of maintaining a workforce. In this section we first analyze the outer optimization and then analyze the dynamic optimization problem in order to find ways to reduce the computational intensiveness of solving these problems. For purposes of this discussions we will assume that workload has been scaled by π , so without loss of generality, we only use units of worker shifts.

Outer Optimization As stated before the solution of problem 2.2 involves solving the problem $P1$. The objective function can be decomposed into a deterministic part:

$$C_{rw}NV + C_{cw}MGV + C_f(N + M)$$

and a stochastic part based upon the dynamic program:

$$f_1^S(x_1, \kappa_1)$$

The deterministic part is just a combination of linear functions in N and M . The stochastic part is not linear but it is non-increasing in N and M and therefore quasiconvex. This means that the objective function of problem 2.2 is a combination of linear and quasiconvex functions of the staffing level S and therefore it is a quasiconvex function. We prove later that if we relax the integrality constraints, the objective function of the outer optimization is convex.

Therefore, the outer optimization problem is the minimization of a convex function on a two dimensional lattice. One way to solve this problem would be to constrain our search to the non-negative values of M and N such that $M + N \leq W_{up}$ that would depend upon the workload distribution and absenteeism rates. We can then search for the optimal solution using a two dimensional binary search that would require $O((\log W_{up})^2 DP(W_{up}))$ time to solve, where $DP(W_{up})$ is the time to solve the dynamic program.

3.3 Inner Dynamic Optimization Analysis

In analyzing the inner, dynamic optimization problem we limit the discussion to the same-period notification with no absenteeism problem and the next-period notification with absenteeism problem. These two problems are representative of the main computational issues of the family of models. In this section we summarize the important observations about the control policies and leave their proofs to the appendix. Some of the results are general and some are specific to a continuous approximation of the real problem.

3.3.1 Same-period notification with no absenteeism

If we consider different relative cost structures we can characterize the control decisions made each period. We assume that the backlog costs in each period including the final period are linear and, equal. We also assume that exogenous work arrivals are iid. Within the dynamic programming problem there are then three cost parameters, C_{cw} , C_{ot} , and C_b . Their relative values will determine the control decisions made each period. In this analysis we will assume all the state and control variables to be continuous. Those results that are independent of the continuous approximation will be noted.

$C_{cw} \leq C_{ot} \leq C_b$ In this case in each period in which x_t exceeds N it is optimal to use as many call-in workers and overtime workers as are available to eliminate backlog. Furthermore we always use all the call-in workers available before utilizing any overtime. This means in each such period t , $u_t = \text{Min}[M, x_t - N]$ and $\omega_t = \text{Min}[OT_{max}(N, M), x_t - N - u_t]$. We notice that this cost structure results in our control decisions only being dependent upon the state variable x_t .

$C_{cw} \leq C_b \leq C_{ot}$ In this case, in each period in which x_t exceeds N it is optimal to use as many call-in workers as are available to eliminate backlog. This means in each such period t , $u_t = \text{Min}[M, x_t - N]$. The determination of the overtime utilization policy requires a little more work. We summarize our observations about the optimal policies here. 1) We do not use overtime shifts before using all available call-in workers. 2) In the final stage we do not use overtime shifts at all. If we assume that all variables are continuous we can prove that: 3) $f_t(x_t, \kappa_t)$ is convex in (x_t, κ_t) for all t . 4) There exists a backlog tolerance $\beta_t(\kappa_t)$ that is the maximum amount of backlog we will tolerate before we utilize overtime shifts.

The backlog tolerance dictates the use of overtime as follows: When $x_t > N + M$ we define $y_t = x_t - N - M$. The optimal overtime usage is:

$$\omega_{V-1}^* = y_{V-1} - b_{V-1}^*$$

Where, b_{V-1}^* is define by the following:

$$b_{V-1}^* = \begin{cases} y_{V-1} & \text{if } \beta_{V-1}(\kappa_{V-1}) \geq y_{V-1} \\ \beta_{V-1}(\kappa_{V-1}) & \text{if } [y - OT_{max}(N, M)]^+ \leq \beta_{V-1}(\kappa_{V-1}) < y_{V-1} \\ [y - OT_{max}(N, M)]^+ & \text{if } \beta_{V-1}(\kappa_{V-1}) < [y - OT_{max}(N, M)]^+ \end{cases}$$

It is clear that $\beta_t(\kappa_t)$ is a non-decreasing function of κ_t since the more free call-in hours are available the easier it will be to deal with future backlog. It should also be clear that when $C_B = C_b$, $\beta_t(\kappa_t)$ increases with t . I.e. as we approach the end of the planning horizon we become more tolerant of backlog.

$C_b \leq C_{cw} \leq C_{ot}$ In this case we make the following observations: 1) We do not hoard “free” call-in workers. This means that as long as $\kappa_t > 0$ we utilize call-in workers whenever the workload exceeds N . 2) As a result of observation 1, we will not utilize overtime shifts before utilizing all available call-in workers. Assuming continuity we can show that: 3) Results similar to the previous cost structure hold with respect to backlog tolerances and the convexity of the cost-to-go functions.

Based upon the convexity results mentioned above, we prove that the objective function of problem $P1$ is actually convex.

3.3.2 Analysis of Next-period notification with Absenteeism Problem

We now consider the analysis of the dynamic programming models for the next-period notification scenario with absenteeism. We recall from section 3 that we distinguish between even and odd stages. In the even stages we make decisions regarding overtime usage. In the odd stages we make decisions regarding call-in worker utilization. We will focus on the case where $C_{ot} > C_{cw} > C_b = C_B$.

Unlike the same-period problems we cannot make any simple statements about the call-in utilization policy even when $\kappa_t > 0$ because the decision is made before we know the new exogenous work arrival. This means that we cannot be certain that call-in workers that we utilize are actually be needed. However, if we analyze the continuous approximation

to the problem in a similar way as the same-period problem we can show: 1) In the odd stages, for every κ_t there is a $U_t(\kappa_t)$ such that if $x_t \geq U_t(\kappa_t)$ we call-in workers. 2) As in the previous analysis of the same-period notification problem we can also characterize the overtime decisions in the even stages by a backlog tolerance function $\beta_t(\kappa_t)$. 3) $f_t(x_t, \kappa_t)$ is convex in (x_t, κ_t) . 4) The objective function of problem $P1$ is convex.

4 Numerical Results

In this section we review some numerical results that demonstrate the effects of different levels of workforce flexibility on the performance of the work system. These results show the linkages between stochasticity of the work environment, information, and flexibility. We also make observations about the effects of using a finite horizon in the model.

4.1 Examples

In these results we assume that the system starts with no backlog and that new work arrives each period in independent and identically distributed amounts. We perform tests for two work arrival distributions, $W1$ and $W2$, depicted in figure 9. The coefficient of variation for $W1$ and $W2$ are respectively .56 and .2. $W1$ is a scaled representation of the work arrival to a new accounts processing area of a large mutual fund company in Boston. $W2$ was chosen arbitrarily for purposes of comparison, as a distribution with lower coefficient of variations and higher mean. In all the runs we assume that the fixed costs C_f of each employee is zero, that the number of periods in the planning horizon is $V = 20$, and all workload quantities are in terms of person-shifts.

The non-zero worker cost parameters are: $C_{rw} = 1$, $C_{cw} = 1.2$, $C_{ot} = 1.5$. We also assume a linear cost for backlog each period with the final period having the same backlog cost as all other periods, namely: $C_b = C_B = 1$. In the first set of runs we fix the overtime limit at $.25(N + M)$ and compare the relative costs of call-in worker guarantees of $G = .6, .4, .2$. In each run we set the cost of operating with no call-in workers to be 100% and show the cost benefits of increased call-in worker flexibility as G is decreased. We also

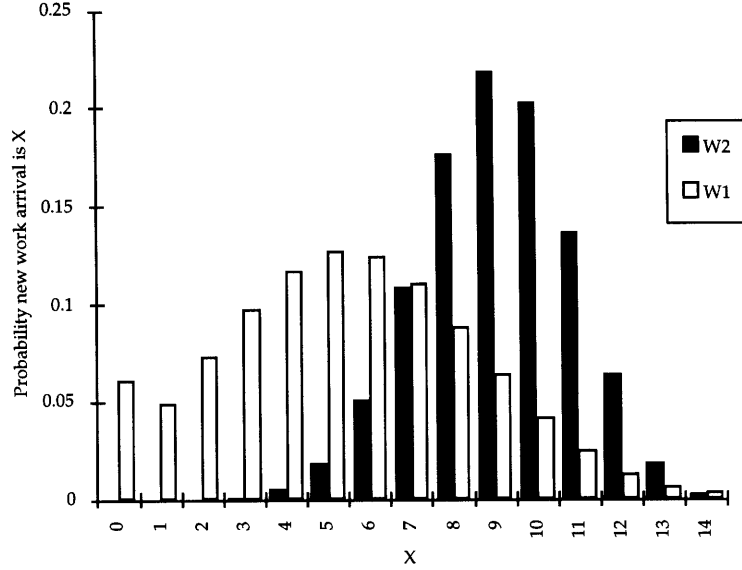


Figure 9: Probability distributions for work arrival pattern W1 and W2

display a lower bound generated by assuming that every period we had the exact number of regular workers to process all the newly arrived work. For every different set of values for the problem parameters we are finding the optimal staffing S and comparing these optimal objective values. E.g. when $G = .6$ the objective value we use is the expected value of the dynamic program with $S = (7, 3)$ when $G = .2$ we use the expected value of the dynamic program with $S = (7, 6)$.

In figure 10 we have the results for the case of same-period notification, with no absenteeism, for work arrivals $W1$ and $W2$. We can observe two phenomena in this comparison. First, there are diminishing returns to increased call-in flexibility, and second, there is a greater benefit from flexibility in the scenario with greater stochasticity in the work loads, i.e. $W1$. The system with work arrivals $W2$ is more predictable and therefore we see that staffing with just regular workers yields a solution that is only about 10% greater than the lower bound. We also notice that the benefit of flexibility relative to the lower bound is greater in the $W1$ case than the $W2$ case.

We next compare two cases in which we have next-period notification with high variability work arrivals, $W1$. In the first case we assume that there is no absenteeism and in

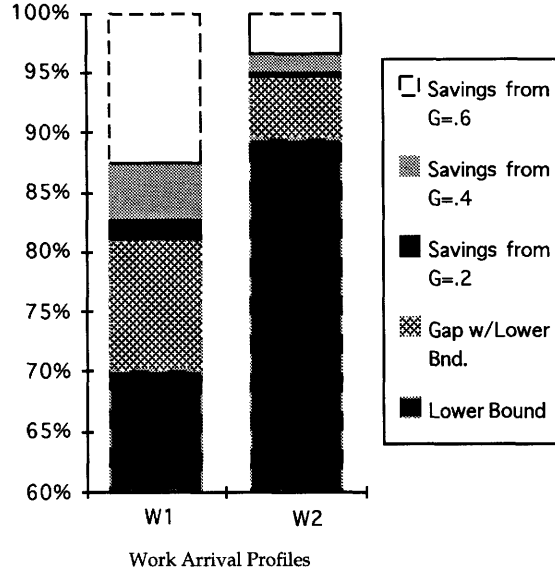


Figure 10: Comparison of the benefits of different levels of Call-in worker flexibility for two work arrival scenarios with same-period notification.

the second that the absenteeism rate for regular workers is .05 and for call-in workers it is .1. By these rates we mean that the probability that a particular regular worker will be absent is .05 and is independent of all other workers. For call-in workers the probability that a particular worker will not be able to come when we call them is .1. The results are displayed in figure 11.

In this figure we can see that when there is no absenteeism there is almost no benefit from the use of call-in workers and only when the guarantee is very low, $G = .2$. This can be understood in terms of information. In the next-period notification problem the manager must decide to call-in workers before she is aware of the new work arrivals. This means that her only information about the system state is the amount of backlogged work. Since the backlog penalty is relatively high the system tends to maintain a very low level of backlog. This means that when the call-in decision is made there is very little new state information beyond what is known at the beginning of the planning period. Since call-in workers are more expensive than regular workers, per shift of work, they are not useful.

This situation changes when we introduce absenteeism. Even though the absenteeism

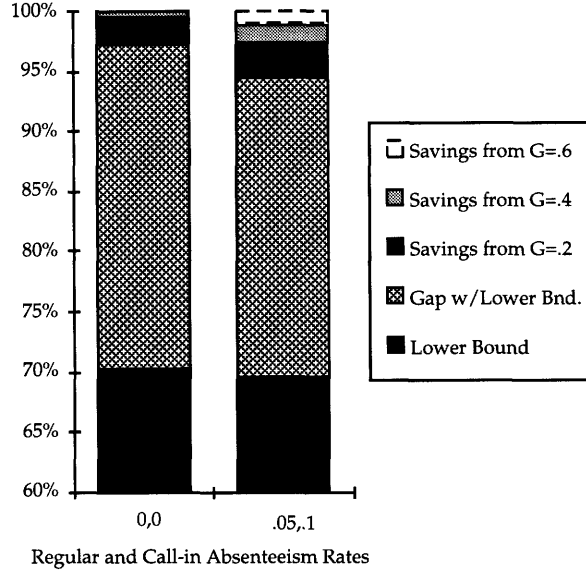


Figure 11: Comparison of the benefits of different levels of call-in worker flexibility for two absenteeism scenarios with work arrival profile W1 and next-period notification.

rates used in the example are biased against call-in workers, we find that they become more useful than in the no-absenteeism case. This phenomenon can be explained by information as well. We have assumed that the manager is aware of regular worker absenteeism before making the call-in decision. This means that she has relatively more system information when she makes the call-in decision than in the no-absenteeism case. This results in a more effective utilization of the flexibility provided by call-in workers.

In all notification cases we always make overtime decisions after we know how much work is in the system and how much absenteeism there is. This means that the overtime decisions is always the most informed. In the next-period notification problems this gives overtime a preference over call-in flexibility. In figure 12 we compare the benefits of increasing the overtime limit versus decreasing the call-in guarantee.

To make these comparisons we calculated the optimal solution values when $G = .2, .4, .6$ with no overtime. We then calculated optimal solution values when $OT_{max} = .25, .5, .75$ with no call-in workers. We also calculated a base case solution value with no call-in workers or overtime. The results of these calculations in table 1.

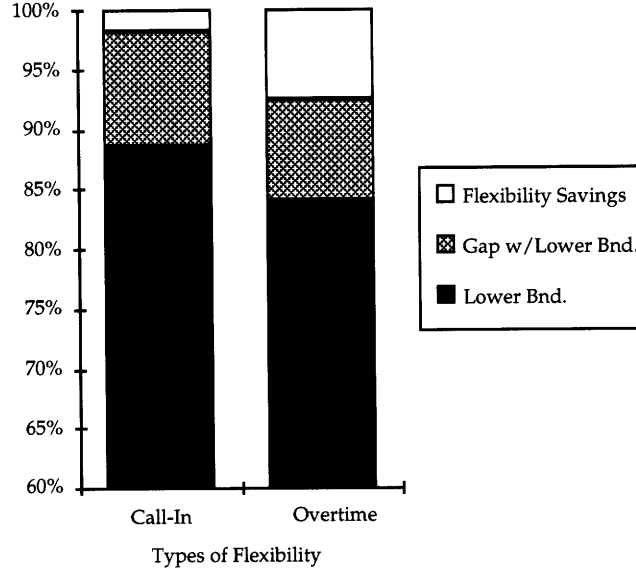


Figure 12: Comparison of different types of flexibility for next-period notification with work profile W2

In each column of the graph we compare the maximum benefits of the type of flexibility relative to the base case of no flexibility and show the gap with a lower bound based on perfect information and staffing.

This can be compared to the same-period notification cases in which the benefits of call-in flexibility bring greater benefits than overtime as can be seen in figure 13. In this figure we compare runs in which there is absenteeism and the work arrivals follow the W2 distribution. The results of the various runs appear in table 2. Figure 13 is organized similarly to figure 12.

To summarize, the numerical examples demonstrate that there are strong links between the benefits of workforce flexibility, the stochasticity of the work environment, and the information available to the decision maker. In general flexibility is more beneficial in environments with more stochasticity, than less. However the benefits may only be realized when there is sufficient information available at decision points. In the numerical examples we considered, call-in workers only provided a significant benefit when we had a lot of system information when the call-in decision was made. Similarly, overtime flexibility became more

Call-In	Obj. Value	Obj. Value	Overtime
No-CW	230	230	No-OT
G=.6	228	225	$OT_{max} = .25$
G=.4	225	211	$OT_{max} = .5$
G=.2	222	211	$OT_{max} = .75$
Lower Bnd.	192	192	Lower Bnd.

Table 1: Notification: Next-period, Work arrivals: W3, Absenteeism: .05 regular, .1 call-in

Call-In	Obj. Value	Obj. Value	Overtime
No-CW	230	230	No-OT
G=.6	211	216	$OT_{max} = .25$
G=.4	203	211	$OT_{max} = .5$
G=.2	200	211	$OT_{max} = .75$
Lower Bnd.	192	192	Lower Bnd.

Table 2: Notification: same-period, Work arrivals: W3, Absenteeism: .05 regular, .1 call-in

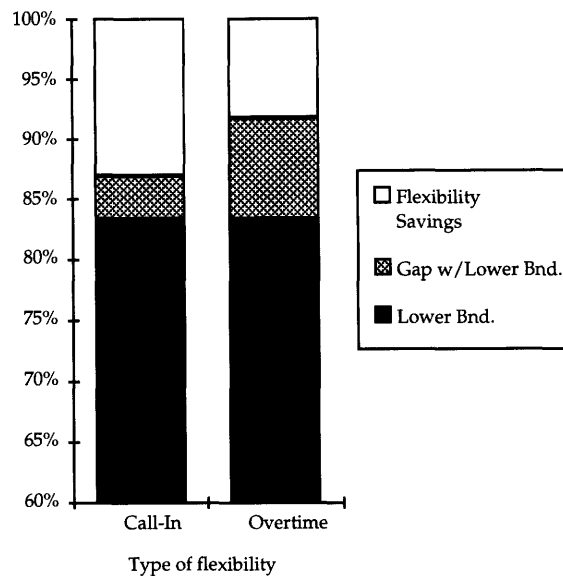


Figure 13: Comparison of overtime and call-in flexibility for same-period notification with overtime and work arrivals W2

useful than call-in workers in the next-period notification cases.

4.2 Call-in Usage Trajectories

We now consider what the pattern of call-in utilization looks like. In figures 14 and 15 we look at the period by period expected call-in worker usage and its coefficient of variation for next-period notification with absenteeism, work arrivals W1, and staffing level $S = (5, 2)$ for different values of G . (This staffing level is optimal for the $G = .6$ case.) We make the following observations about figure 14:

Obs1: In each period the expected call-in usage increases with G .

Obs2: For each G there is a peak late in the planning horizon.

Obs3: For each G the call-in usage decreases sharply toward 0 after the peak.

Obs4: Call-in usage always increases from period 1 to period 2.

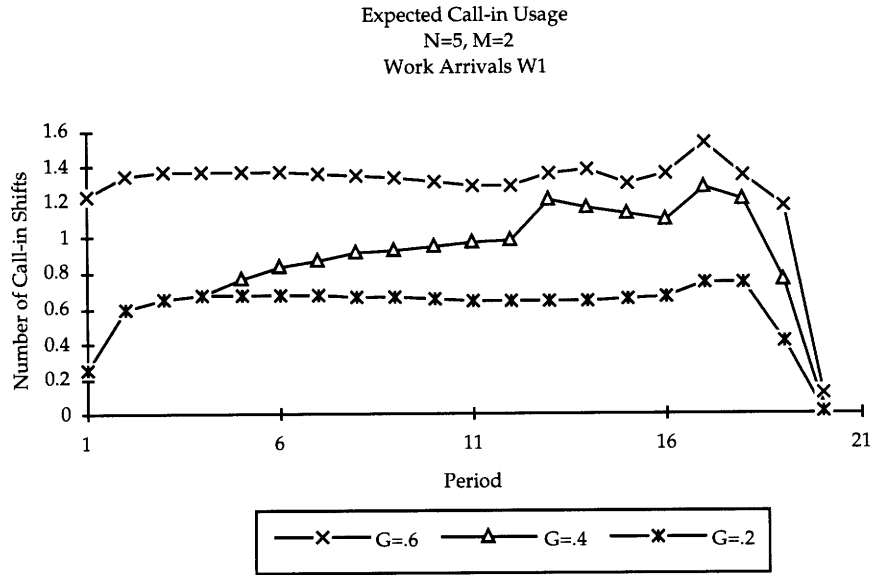


Figure 14:

Obs1 is explained by the fact that for higher values of G more call-in worker shifts have been guaranteed and therefore, within the context of period by period operations, there are more free call-in shifts to use. Obs2 is explained by the call-in guarantee as well. We know that if toward the end of the planning horizon we have shifts remaining in the call-in guarantee it is worthwhile to use them up in a less conservative manner than in earlier periods, this causes the observed peak. Obs3 is expected in this case since call-in worker costs per shift are greater than backlog costs. At the end of the planning period it is most likely that the guaranteed shifts will have been used up; furthermore, the finite horizon makes the system more tolerant of backlog. These three factors make call-in utilization less attractive. Obs4 is explained by the initial conditions of the system. We assumed that the system starts out empty which means that it will probably have more work in the second period than the first since there might be some backlog after the first period. Therefore, in expected value, there will be a greater need for workers in period 2 than in period 1. We make the following observations about figure 15:

Obs1a: Variation decreases with G .

Obs2a: Variation increases dramatically in the last periods.

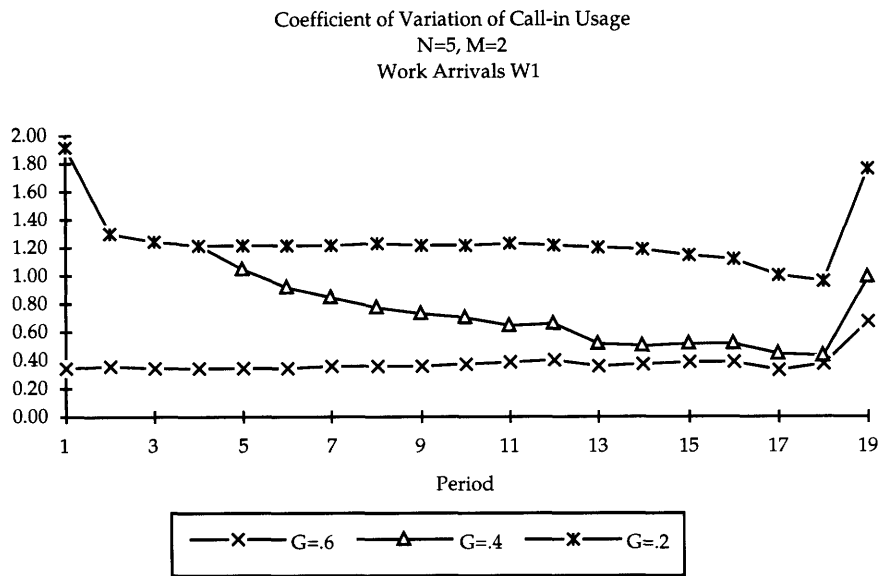


Figure 15:

Obs3a: Initial coefficient of variation for $G = .2, .4$ is very high.

Obs4a: For $G = .2$ and $G = .4$ variation seems fairly steady during the body of the planning horizon, while during this time it is decreasing for $G = .4$.

Obs1a is closely related to Obs1. When we change G we don't change the randomness in work arrivals and absenteeism and since expected call-in usage is higher when G is higher it makes sense that the coefficient of variation decreases with G . However, this relationship is even stronger because the reduced usage of call-in workers brought about by lowering G results in an increase in backlog levels and the variability of backlog. Obs2a can be explained by the sharp drop in call-in worker usage. Obs3a can be explained by the low call-in usage in the first periods caused by the initial conditions. Obs4a seems to suggest that when the call-in guarantee is substantial, $G = .6$, or small, $G = .2$, there is little difference between system behavior in the body of the planning horizon, i.e. some form of steady-state. When $G = .4$ the behavior of the system seems to start like a $G = .2$ system and then converge to a $G = .6$ system. This behavior shows that for intermediate values of G the system is very sensitive to the remaining call-in guarantee and therefore never achieves any sort of steady state.

To get a sense of the behavior of the system in the same-period notification case we compare it with the behavior of a next-period notification system with the same parameters. In figures 16 and 17, we have the expected value and coefficient of variation of call-in utilization by period for $W1$ work arrivals, no absenteeism, $G = .4$ and $S = (2, 7)^{13}$.

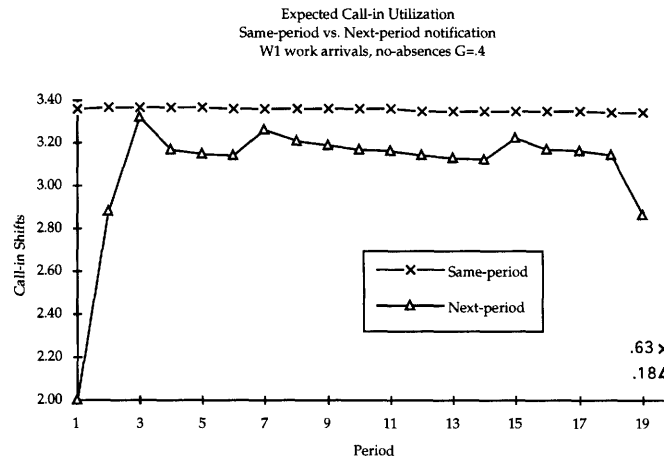


Figure 16:

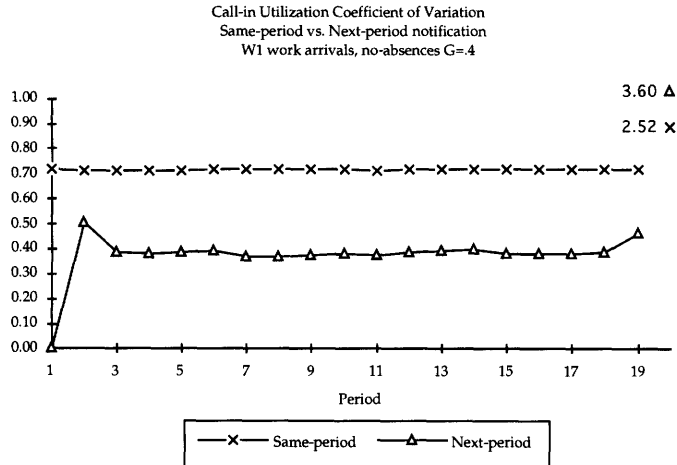


Figure 17:

The main observations are:

Obs1b: Expected call-in usage is higher in same-period case than next-period.

¹³Optimal staffing for same-period notification.

Obs2b: Same-period variation in call-in usage is higher than in next-period case.

Obs3b: In same-period case there is little change in system behavior from period to period.

We explain these observations by noting that because call-in decisions are made in the same-period case, when there is complete information about the workload in the system, the realizations of the work arrival process will directly determine call-in usage. Uncertain impact of the call-in workers in the next-period case keeps their usage lower, and the anticipatory nature of the decision keeps its variability lower as well. Since the work arrivals are iid. each period we can also explain Obs3b by the direct relation between same-period call-in usage and work arrivals.

To summarize, the behavior of the system with respect to call-in worker utilization over the course of the planning horizon can be very different for the same-period notification case than for the next-period and can be strongly effected by the call-in guarantee. We also see that the effects of the intitial conditions and the finite horizon are not important over the majority of the planning horizon, but that the finite call-in guarantee can effect the system throughout the planning horizon (see Obs1, Obs2, and Obs4a).

4.3 Finite horizon issues

We will now consider the question of what is the appropriate way to take into account the effect of final period backlog on the future performance of the system. Since we have constructed finite horizon models our primary concern is that late in the planning horizon we will start to accept higher levels of backlog than would be realistic. Setting the final period backlog cost at a very high level, to force the system to empty by the end of the planning horizon artifically creates a significant difference between our tolerance for backlog in the final period versus others.

Our approach to this problem is to optimize the staffing level over multiple planning horizons, rather than just one. That is, if the planning horizon has length V , we would solve the staffing problem over kV periods for some integer k . Every V periods we reset the call-in guarantee remaining to MGV . In this scheme we do not assign any special penalty

to backlog remaining at the end of the kV th period. The larger the value of k the less the edge effects.

In this section we have presented numerical results for specific problem instances that are relatively insensitive to the finite horizon. We found that after at most 3 planning horizons the end of planning horizon backlog stabilized and that the optimal staffing levels did not change from those determined for the single planning horizon. This need not always be the case and can be explained by the backlog penalty. The one period backlog penalty is high relative to the labor costs, therefore backlog is kept to a very low level in all periods. This means that successive planning horizons are almost independent of one another.

5 Extensions and Conclusions

In most practical applications of these staffing models there are different exogenous work arrival processes for different time periods. Typically there is a day of the week seasonality exhibited in the work arrival profile. We can accommodate this phenomenon within the framework of the models formulated in this chapter. To do this we must make some assumptions about how regular workers are utilized.

Implicit in the models is the assumption that all the regular workers are scheduled to work in each time period. This can be interpreted as all regular workers being full-time workers. When workloads vary from time period to time period because of day-of-the-week type seasonality it is common to use part-time workers, i.e. workers who are not scheduled to work each time period, to match workforce to workload. The models we have developed here are geared toward the stochastic variability as opposed to the deterministic/seasonal variability. Therefore, we do not explicitly model the scheduling of part-time workers. However, it is possible to consider call-in workers to be flexible part-time workers.

In this more general workload arrival case we will model the new work arriving in period t as $d_t = s_t + r_t$ where s_t is a deterministic component of workload and r_t is the stochastic component that is distributed according to some probability mass function $F_t(r)$. Note that this PMF is dependent upon the time period.

If the deterministic components s_t are significant we can assume that a separate schedule for regular workers including part-time regulars has been determined to satisfy the deterministic needs. We can then proceed with our ordinary models concerned with only the stochastic component r_t . If the deterministic component of the work arrival is not significant with respect to the stochastic component we can just view d_t as being a stochastic variable as we did in all our models.

In practice it is also common that the workload information that is available when the call-in decision is made is somewhere between the extremes of same-period and next-period notification models. We call this mixture of the two, the mixed-info model. In this model we split periods into odd and even sub-stages as in the next-period notification cases. We model new work arrival at the start of each of these sub-stages with different probability distributions F_{odd}, F_{even} . The call-in decision is made after the realization of F_{odd} and the overtime decision is made after the realization of F_{even} .

Although we can accommodate non-homogeneous work arrival processes and mixed-info within the formulations presented in this paper, we cannot, in general, extend the analytical results of the previous section to this case. To summarize, we do not really have to change our models to adapt to non-identical workload arrivals. All we do is use the time period dependent distributions in all calculations of expected values involving workload in the dynamic programs.

In this paper we have defined a new family of realistic workforce management models that represent the many of the important characteristics of modern service providing firms in stochastic environments. We have created a classification scheme that distinguishes different problems by the presence of crosstraining, the availability of workload information, and absenteeism. We have formulated the general one-job problem 2.2 as an optimization over two dimensions; regular workforce size N and call-in workforce size M . The value of the objective function at a particular solution $S = (N, M)$ is the expected cost of the optimal solution to a finite horizon dynamic program. We have formulated all the one-job problems depicted in the problem tree in figure 1. That is, we have formulated the various problems as distinguished by notification and absenteeism. All the dynamic programming formulations are based upon a Markovian structure that governs the work in the system

and the cumulative utilization of call-in workers.

In the formulations involving absenteeism we characterized different cases distinguished by the relative timing of the two staffing decisions made each period and information about workload and absenteeism. Each case required a distinct formulation. Finally we showed how the formulations could be extended to situations in which the exogenous new work arrival each time period were not identically distributed. In all cases we have assumed that they are independent.

In the analysis section we have shown how this family of one-job models represents a wide range of realistic practical situations. We also derived characterizations of the optimal call-in and overtime decision policies for the inner dynamic optimization problem. In the numerical results section we demonstrated how the model can be used to determine the benefits of different forms of labor flexibility. We have also gained insight into the relationship between the benefits of flexibility, the degree of stochasticity of the work environment and the availability of system information. The implication of these results is that workforce flexibility, in and of itself, is not a panacea for a firm's operating costs. The form of flexibility that best takes advantage of system information will be the most effective. Furthermore, if the stochasticity of the system is not great, the benefits of flexibility will probably be marginal.

A Same-period notification analysis

In this appendix we demonstrate in detail that $f_t(x_t, \kappa_t)$ is convex when $C_{cw} \leq C_b \leq C_{ot}$. This result carries over to the other cost structures of interest, namely $C_{cw} \leq C_{ot} \leq C_b$ and $C_b \leq C_{cw} \leq C_{ot}$.

Lemma A.1 *If $C_{cw} \leq C_b \leq C_{ot}$ then $f_V(x_V, \kappa_V)$ is convex.*

Proof We know that we do not use any overtime hours before we have utilized all the call-in workers available. This means that overtime is not used unless x_t exceeds $N + M$.

In the final period we know we do not use overtime at all, since the cost of backlog for that period is less than the overtime cost. Therefore we write:

$$f_V(x_V, \kappa_V) = C_{cw}[u_V^* - \kappa_V]^+ + C_b[x_V - N - u_V^*]$$

$$\text{Where, } u_V^*(x_V) = \text{Min}(M, [x_V - N]^+)$$

We drop the subscript V and consider the following 3 regions of values for x :

R1: Region 1 is $x < N \Rightarrow u^*(x) = 0$

R2: Region 2 is $N \leq x < N + M \Rightarrow u^*(x) = x - N$

R3: Region 3 is $N + M \leq x \Rightarrow u^*(x) = M$

Therefore:

$$\text{For } x \in R1 : f(x, \kappa) = 0 \quad \forall \kappa$$

$$\text{For } x \in R2 : f(x, \kappa) = C_{cw}[x - N - \kappa]^+ \quad \forall \kappa$$

$$\text{For } x \in R3 : f(x, \kappa) = C_{cw}[M - \kappa]^+ + C_b(x - N - M) \quad \forall \kappa$$

All three of the above functions are clearly convex in their respective regions. To demonstrate convexity we need to show that convexity holds between points from different regions. These demonstrations are not very informative so we will only present the demonstration for points in $R2$ and $R3$. Define $x_o = \lambda_2 x_2 + \lambda_3 x_3$ for $\lambda_2, \lambda_3 \geq 0$ and $\lambda_2 + \lambda_3 = 1$ and similarly $\kappa_o = \lambda_2 \kappa_2 + \lambda_3 \kappa_3$. Where $x_i \in R_i$.

Step 1: Say we have chosen λ_2, λ_3 so that $x_o \in R3$.

$$\begin{aligned} \Rightarrow f(x_o, \kappa_o) &= C_{cw}[M - \kappa_o]^+ + C_b(x_o - N - M) \\ &\Rightarrow \lambda_2 f(x_2, \kappa_2) + \lambda_3 f(x_3, \kappa_3) - f(x_o, \kappa_o) \\ &= \lambda_2 C_{cw}[x_2 - N - \kappa_2]^+ + \lambda_3 C_{cw}[M - \kappa_3]^+ + \lambda_3 C_b(x_3 - N - M) - C_{cw}[M - \kappa_o]^+ - C_b(x_o - N - M) \\ &= -\lambda_2 C_b(x_2 - N - M) + Q_1 \end{aligned}$$

Where

$$Q_1 = C_{cw}\{\lambda_2[x_2 - N - \kappa_2]^+ + \lambda_3[M - \kappa_3]^+ - [\lambda_2M + \lambda_3M - \lambda_2\kappa_2 - \lambda_3\kappa_3]^+\}$$

Note: $Q_1 \geq C_{cw}\lambda_2(M - x_2 - N)$ and $x_2 \in R2 \Rightarrow M - x_2 - N \leq 0$.

$$C_{cw} \leq C_b \Rightarrow -\lambda_2C_b(x_2 - N - M) + Q_1 \geq 0$$

Therefore, convexity holds for $x_o \in R3$.

Step 2: Say we have chosen λ_2, λ_3 so that $x_o \in R2$.

$$\begin{aligned} &\Rightarrow \lambda_2f(x_2, \kappa_2) + \lambda_3f(x_3, \kappa_3) - f(x_o, \kappa_o) \\ &= \lambda_2C_{cw}[x_2 - N - \kappa_2]^+ + \lambda_3C_{cw}[M - \kappa_3]^+ + \lambda_3C_b(x_3 - N - M) - C_{cw}[x_o - N - \kappa_o]^+ \\ &= \lambda_3C_b(x_3 - N - M) + C_{cw}Q_2 \end{aligned}$$

Where

$$Q_2 = \lambda_2[x_2 - N - \kappa_2]^+ + \lambda_3[M - \kappa_3]^+ - [\lambda_2(x_2 - N - \kappa_2) + \lambda_3(x_3 - N - \kappa_3)]^+$$

Note: $Q_2 \geq \lambda_3(M - x_3 - N)$ and $x_3 \in R3 \Rightarrow M - x_3 + N \leq 0$.

$$C_{cw} \leq C_b \Rightarrow \lambda_3C_b(x_3 - N - M) + C_{cw}Q_2 \geq 0$$

Therefore, convexity holds for $x_o \in R2$.

Results of steps 1 and 2 complete proof. ■ .

Lemma A.2 *If $C_{cw} \leq C_b \leq C_{ot}$ then $f_{V-1}(x_{V-1}, \kappa_{V-1})$ is convex.*

Proof If $x_{V-1} \leq (N + M)$ then we know that:

$$f_V(x_{V-1}, \kappa_{V-1}) = C_{cw}[x_{V-1} - N - \kappa_{V-1}]^+ + E_{d_V}[f_V(d_V, \kappa_V)]$$

which is convex.

If $x_{V-1} > N + M$ then we define $y_{V-1} = x_{V-1} - N - M$ and write:

$$f_{V-1}(x_{V-1}, \kappa_{V-1}) = \text{Min}_{0 \leq \omega \leq y} \{C_{cw}[M - \kappa_{V-1}]^+ + C_b[y_{V-1} - \omega_{V-1}] \\ + C_{ot}\omega_{V-1} + E_{d_V}[f_V(d_V + y_{V-1} - \omega_{V-1}, \kappa_{V-1} - M)]\}$$

We define $b_{V-1} = y_{V-1} - \omega_{V-1}$ to be the backlog at the end of period $V - 1$ and rewrite the cost-to-go function as:

$$f_{V-1}(x_{V-1}, \kappa_{V-1}) = \min_{[y - OT_{max}(N, M)]^+ \leq b \leq y} \{(C_b - C_{ot})b_{V-1} + E_{d_V}[f_V(d_V + b_{V-1}, \kappa_{V-1} - M)]\} \\ + C_{cw}[M - \kappa_{V-1}]^+ + C_{ot}y_{V-1}$$

Using the convexity of $f_V(x_V, \kappa_V)$ we see that the quantity within the minimization brackets is a convex function of the variables b_{V-1} and κ_{V-1} . This means that there is some minimizer $\beta_{V-1}(\kappa_{V-1})$ of this function for every possible value of κ_{V-1} . This implies the following:

$$b_{V-1}^* = \begin{cases} y_{V-1} & \text{if } \beta_{V-1}(\kappa_{V-1}) \geq y_{V-1} \\ \beta_{V-1}(\kappa_{V-1}) & \text{if } [y - OT_{max}(N, M)]^+ \leq \beta_{V-1}(\kappa_{V-1}) < y_{V-1} \\ [y - OT_{max}(N, M)]^+ & \text{if } \beta_{V-1}(\kappa_{V-1}) < [y - OT_{max}(N, M)]^+ \end{cases}$$

Therefore:

$$\omega_{V-1}^* = y_{V-1} - b_{V-1}^*$$

The above argument was based upon the convexity of $f_V(x_V, \kappa_V)$. This also allows us to demonstrate the convexity of f_{V-1} . Let's define:

$$G(b, \kappa) = (C_b - C_{ot})b + E_{d_V}[f_V(d_V + b, [\kappa - M]^+)]$$

We know that $G(b, \kappa)$ is a convex function of (b, κ) . We can now write:

$$f_{V-1}(x_{V-1}, \kappa_{V-1}) = L(x, \kappa) + C_{cw}[M - \kappa_{V-1}]^+ + C_{ot}y_{V-1}$$

$$\text{Where } L(x, \kappa) = \min_{[y-OT_{max}(N,M)]^+ \leq b \leq y} \{G(b, \kappa)\}.$$

We focus our attention on $L(x, \kappa)$. If $x_o = \lambda_1 x_1 + \lambda_2 x_2$ for $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$ and similarly $\kappa_o = \lambda_1 \kappa_1 + \lambda_2 \kappa_2$, we define the following sets:

$$S_o \equiv [[x_o - N - M - OT_{max}(N, M)]^+, x_o - N - M]$$

$$S_1 \equiv [[x_1 - N - M - OT_{max}(N, M)]^+, x_1 - N - M]$$

$$S_2 \equiv [[x_2 - N - M - OT_{max}(N, M)]^+, x_2 - N - M]$$

By the definition of $L()$ we have that:

$$L(x_o, \kappa_o) \leq G(b, \kappa_o) \text{ for all } b \in S_o$$

or,

$$L(x_o, \kappa_o) \leq G(\lambda_1 b_1 + \lambda_2 b_2, \lambda_1 \kappa_1 + \lambda_2 \kappa_2) \text{ for all } b_i \in S_i$$

by the convexity of $G()$ we have:

$$L(x_o, \kappa_o) \leq \lambda_1 G(b_1, \kappa_1) + \lambda_2 G(b_2, \kappa_2) \text{ for all } b_i \in S_i$$

$$\Rightarrow L(x_o, \kappa_o) \leq \lambda_1 \min_{b_1 \in S_1} G(b_1, \kappa_1) + \lambda_2 \min_{b_2 \in S_2} G(b_2, \kappa_2)$$

Therefore, $L(x, \kappa)$ is convex. We see that $f_{V-1}(x_{V-1}, \kappa_{V-1})$ is the sum of convex functions and therefore is convex as well when $x_{V-1} > N + M$. To complete the proof of the convexity of $f_{V-1}(x_{V-1}, \kappa_{V-1})$ we need to connect the two cases for $x_{V-1} > N + M$ and $x_{V-1} \leq N + M$. This can be done in a similar fashion to that in the proof of the convexity of $f_V(x_V, \kappa_V)$. ■ .

Proposition A.1 *If $C_{cw} \leq C_b \leq C_{ot}$ then $f_t(x_t, \kappa_t)$ is convex.*

Proof We can apply the arguments of Lemmas A.1 and A.2 inductively and show that: $f_t(x_t, \kappa_t)$ is convex for all t . ■ .

Proposition A.2 *The objective function of problem P1:*

$$C(S) = C_{rw}NV + C_{cw}MGV + C_f(N + M) + f_1^S(x_1, \kappa_1)$$

is convex.

Proof From proposition A.1 we know that $f_1(x_1, \kappa_1)$ is convex in (x_1, κ_1) . Since $\kappa_1 = MGV$ we have that $f_1(x_1, \kappa_1)$ is convex in M as well.

If we redefine x_t to be the work in the system in excess of the regular staff N and d_t to be the new work arriving in excess of the regular staff processing capacity, we would have cost-to-go functions of the form: $f_1(x_1^{old} - N, \kappa_1)$. Standard convexity results show that $f_1(x_1^{old} - N, \kappa_1)$ is convex in (N, M) [Baz93].

$\Rightarrow f_t^S(x_1, \kappa_1)$ is convex in S . We have noted before that the deterministic part of $C(S)$ is a combination of linear functions. $\Rightarrow C(S)$ is the sum of linear terms and a convex term.

$\Rightarrow C(S)$ is convex in S . ■ .

B Next-period notification with absenteeism analysis

In this appendix we focus on the cost structure with $C_b \leq C_{cw} \leq C_{ot}$ but it should be clear that they carry over to the other cases of interest.

Lemma B.1 *If $C_b \leq C_{cw} \leq C_{ot}$, then $f_{2V}(x_{2V}, \kappa_{2V})$ is convex.*

Proof The final stage, $2V$ is an even stage in which we make an overtime decision. Since $C_{ot} > C_b$ we will never utilize overtime:

$$\Rightarrow f_{2V}(x_{2V}, \kappa_{2V}) = C_b x_{2V} \text{ a convex function}$$

■ .

Lemma B.2 *If $C_b \leq C_{cw} \leq C_{ot}$, then $f_{2V-1}(x_{2V-1}, \kappa_{2V-1})$ is convex.*

Proof In stage $t = 2V - 1$ we will not use any more call-in workers than we have guarantee remaining since $C_{cw} > C_b$. This means we try to utilize $u^* = \min(\kappa_t, M)$ call-in workers,

and will actually use $a = \min(u^*, m)$. Where m is the number of call-in workers who are available.

$$\Rightarrow f_t(x_t, \kappa_t) = C_b E_{d,a}[[x_t + d - a]^+]$$

We see that $f_{2V-1}(x_{2V-1}, \kappa_{2V-1})$ is convex in $(x_{2V-1}, \kappa_{2V-1})$. ■ .

Lemma B.3 *If $C_b \leq C_{cw} \leq C_{ot}$, then $f_{2V-2}(x_{2V-2}, \kappa_{2V-2})$ is convex.*

Proof When $t = 2V - 2$,

$$f_t(x_t, \kappa_t) = \min_{\omega_t \leq x_t} \{C_{ot}\omega_t + C_b(x_t - \omega_t) + E_n[f_{t+1}([x_t - \omega_t]^+ - n, \kappa_t)]\}$$

Dropping the stage subscript we can write this cost-to-go function as:

$$f(x, \kappa) = \min_{0 \leq b \leq x} \{C_b - C_{ot}b + E_n[G(b - n, \kappa)]\} + C_{ot}x$$

Where $G(b - n, \kappa)$ is a convex function as is everything within the minimization. As we showed in proposition A.1 this implies that $f_t(x_t, \kappa_t)$ is convex. ■ .

Lemma B.4 *If $C_b \leq C_{cw} \leq C_{ot}$, then $f_{2V-3}(x_{2V-3}, \kappa_{2V-3})$ is convex.*

Proof When $t = 2V - 3$,

$$f_t(x_t, \kappa_t) = \min_{u_t} \{C_{cw}[u_t - \kappa_t]^+ + E_{d_{t+1}}[f_{t+1}([x_t - u_t + d_{t+1}]^+, [\kappa_t - u_t]^+)]\}$$

Based upon our previous results we know that the quantity within the minimization is a convex function of (x_t, κ_t, u_t) . We can then write:

$$f_t(x_t, \kappa_t) = \min_{u_t} H(x_t, \kappa_t, u_t)$$

In the continuous approximation, the convexity of $H()$ implies the convexity of $f_t()$. ■ .

We can apply these arguments to inductively to show that:

Proposition B.1 *$f_t(x_t, \kappa_t)$ is convex, for all t .*

■ .

As in the same-period notification analysis, the above convexity results can be used to demonstrate that the objective function of problem $P1$ is convex even when we have next-period notification and absenteeism in the regular and call-in workforces.

References

- [Bak73] Kenneth Baker. An optimal procedure for allocating manpower with cyclic requirements. *AIIE Trans*, 5(2), 1973.
- [Baz93] Mokhtar S. etal Bazaara. *Nonlinear Programming Theory and Algorithms*. John Wiley and Sons, Inc., New York, 1993.
- [Ber87] Dimitri Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Inc, Englewood Cliffs, N.J., 1987.
- [BL93a] Oded Berman and Richard C. Larson. Optimal workforce configuration incorporating absenteeism and daily workload variability. *Socio-Econ Planning Sci.*, 27(2):91–96, 1993.
- [BL93b] Gabriel R. Bitran and Maureen Lojo. A framework for analyzing service operations. *European Management Journal*, 11(3):271–282, 1993.
- [BL94] Oded Berman and Richard C. Larson. Determining optimal pool size of a temporary call-in work force. *European Journal of Operations Research*, 1994.
- [BLP94] O. Berman, R. Larson, and E. Pinker. Scheduling workforce and workflow in a service factory. *MIT Operations Research Center Working Paper*, OR 287-94, 1994.
- [Fie94] Jaclyn Fierman. The contingency workforce. *Fortune*, pages 30–35, January 24, 1994.
- [Fod95] Barnaby J. Foder. Bigger roles for suppliers of temporary workers. *New York Times*, page 37, April 1, 1995.

- [HB90] Bennet Harrison and Barry Bluestone. Wage polarisation in the U.S. and the 'flexibility' debate. *Cambridge Journal of Economics*, 14(2):351–373, 1990.
- [HM60] Charles C. Holt and Franco Modigliani. *Planning Production, Inventories, and Work Force*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1960.
- [Kle94] Janice Klein. Maintaining expertise in multi-skilled teams. *Advances in Interdisciplinary Work Teams*, 1:145–165, 1994.
- [Mic87] Francois Michon. Time and flexibility: Working time in the debate on flexibility. *Labour and Society*, 12(1):153–176, 1987.
- [MR73] C. Maier-Roth. Cyclic scheduling and allocation of nursing staff. *Socio-Economic Planning Ser.*, 7:471–487, 1973.
- [Pol89] Anna E. Polivka. On the definition of 'contingent work'. *Monthly Labor Review*, pages 9–16, December 1989.
- [Reb95] James Rebitzer. Job safety and contract workers in the petrochemical industry. *Industrial Relations*, 34(1):40–57, January 1995.
- [Ser94] United States Postal Service. *Agreement between USPS and APWU and NALC*. United States Postal Service, Washington D.C., 1990-1994.
- [Sta95] Current Employment Statistics. *Current Population Survey*. Bureau of Labor Statistics, Washington D.C., March, 1995.
- [Tre89] M. Trevelen. A review of the dual resource constrained system research. *IIE Transactions*, 21(3):279–287, 1989.
- [Tre92] Tiziano Treu. Labor flexibility in Europe. *International Labour Review*, 131(4-5):497–512, 1992.
- [War72] D. Warner. A mathematical programming model for scheduling nursing personnel in hospitals. *Management Science*, 19:411–422, 1972.
- [WR93] Bernhard Wild and Christoph Schneeweiss R. Manpower capacity planning - a hierarchical approach. *International Journal of Production Economics*, 30-31:95–106, 1993.

