

Can Flexibility be Constraining?

August 2008

Keywords: Workforce Flexibility, Planning, Crosstraining, Linear Programming

Edieal Pinker (Corresponding Author)
Simon Graduate School of Business
University of Rochester
Rochester, NY 14627
585-275-2131
Ed.pinker@simon.rochester.edu

Hsiao-Hui Lee
Simon Graduate School of Business
University of Rochester
Rochester, NY 14627

Oded Berman
Rotman School of Management
University of Toronto
105 St. George Street
Toronto, ON M5S 3E6

Can Flexibility be Constraining?

Abstract: We investigate five common options for workforce flexibility and their robustness under uncertain demand. In the first stage, a firm makes optimal staffing decisions according to estimated demand and a given workforce flexibility policy. In the second stage, it reallocates its workforce to react to demand shocks. Our numerical results show that flexibility can lead a firm to staff with too little slack to be flexible to demand shocks, thus leading to higher total costs i.e., staffing and inventory costs. We identify the forms of flexibility that give robust benefits and also analyze how different forms of flexibility interact with each other.

Can Flexibility be Constraining?

1 Introduction

There is a large literature on flexibility in production processes and in particular workforce flexibility. Upton (1995) defines flexibility as "the ability to change or react with little penalty in, time, effort, cost or performance." Another way to define flexibility is as the absence of constraints. As such flexibility allows managers to better match the supply of resources to the demand when there is variability in the timing, content and quantity of work. For this reason flexibility is generally viewed as a positive characteristic of a process or workforce. It is also generally accepted that flexible resources come at a cost. For example, cross-trained workers require more training, workers who have flexible hours require higher pay and often labor flexibility needs to be negotiated with labor unions. Given the potential costs of greater flexibility, researchers have been interested in quantifying the benefits of flexibility and determining the forms of flexibility that give the greatest benefits. In this paper we will address these issues in the context of labor flexibility. Unfortunately much of the literature on workforce flexibility has taken a narrow perspective on the subject thus leading to an incomplete and sometimes distorted view of flexibility.

As described in Abernathy et al. (1973) workforce planning typically involves three hierarchical phases, planning, scheduling and allocating. The planning phase determines the capacity limits of the production process. Scheduling determines when workers should be available and must account for working hour constraints. While the schedule determines the set of workers available at any point in time, the allocation phase determines what they do. The allocation can be reactive to changes in demand which is one way that cross-training or skill flexibility can be valuable. Much of the recent literature on cross-training has focussed on the allocation phase and ignored its connection to the scheduling and planning phases.

We find that if one takes the full spectrum of workforce flexibility into account, meaning working time flexibility in addition to cross-training, and consider all the planning phases, some dangers of flexibility are revealed. Scheduling and allocation are typically constrained because workers expect regular hours with shifts of fixed lengths, or at least minimum lengths,

and have specific skills and capabilities. As a result their schedules and allocations do not always exactly match up to the workflow leading to slack labor capacity. Some kinds of flexibility eliminate constraints and allow a better match. Therefore, some types of flexibility can allow a firm to achieve the same productivity with fewer labor resources, thus reducing costs. This is an example of taking advantage of scheduling flexibility to reduce capacity in the planning phase. However, a byproduct is that the process has less slack capacity and thus less flexibility to deal with demand fluctuations. In this paper we show that some forms of flexibility can actually make the system less flexible in other dimensions.

We consider a high volume production system defined by a network of workstations. Work arrives throughout the day according to some hourly pattern. Different types of work take different paths through the network. The available labor pool is made up of workers with differing skills at each workstation. Given a forecast of demand, a manager must decide how many of each worker type to hire and how to schedule them. Given the pool of workers hired, the manager has the flexibility to reschedule and reallocate them to adapt to changes in demand patterns. We model the following forms of flexibility: crosstraining, intra-shift job switching, part-time workers, start time flexibility, and work-in-process buffers. In Section 2 we review the related literature. In Section 3, we summarize the model of Berman et al. (1997) which we use as a basis for our analysis within a two-stage optimization framework. In Section 4, we report on extensive numerical experiments we conducted with the model and our findings regarding the importance and impact of the various forms of flexibility. We conclude in Section 5.

2 Literature Review

Production system flexibility has been widely studied, see reviews in De Toni and Tonchia (1998), Beach et al. (2000), and Vokurka and O’Leary-Kelly (2000). However the definition and classification schemes of flexibility are quite varied. Browne et al. (1984) categorize eight dimensions of flexibility: machine, product, process, operation, routing, volume, expansion, and production. Sethi and Sethi (1990) add three more dimensions: material handling, program, and market. Vokurka and O’Leary-Kelly (2000) expand the area of flexibility even

further by including four more dimensions: automation, new design, delivery, and labor. Many of these forms of flexibility involve technology or process design. In this paper we focus on workforce flexibility.

Workforce flexibility has been studied from two perspectives. One perspective is staff-scheduling while the other is as a response to stochasticity in demand. The scheduling literature has primarily focused on the computational challenges of determining optimal schedules in a deterministic setting. Ernst et al. (2004) give a comprehensive review of more than 700 papers in the area of personnel scheduling and rostering. Increased flexibility typically increases the complexity of the deterministic optimization because the additional options lead to greater numbers of variables. Jordan and Graves (1995) spurred much interest in evaluating how much flexibility is needed in stochastic environments to get the majority of the benefits. Their finding that relatively sparse and simple "skill-chaining" can achieve most of the benefits of crosstraining in some environments has led to a stream of related work. This paper straddles these two streams of research.

Berman et al. (1997) formulate a linear programming model for jointly optimizing the workforce shift schedule and workflow in a high volume factory with an exogenous deterministic demand. For a large workforce and Worker flexibility is displayed in terms of the number of available shift starting times, the option of part-time shifts, mid-shift job switching by crosstrained workers and the degree to which work in progress (WIP) can be buffered. Computational experiments show the way buffers and labor flexibility act as substitutes. It also shows how the benefits of different forms of flexibility depend upon the work arrival pattern. The model in that paper is also the core model in our paper, and the model is summarized in detail in Section 3. Bard (2004) looks at a similar setting with a deterministic integer programming model and solution methodology that does days off and break scheduling. He takes a similar approach focusing on a particular form of hierarchical crosstraining called downgrading, in which higher skilled workers can do lower skilled jobs as needed. Campbell (1999) and Brusco (2008) consider crosstraining in the allocation problem after demand has been realized for a work shift. As such they assume a fixed number of workers, and do not model shift schedules. They also restrict themselves to a work environment with independent departments and not the workflow between them. Both papers ignore other forms of

flexibility than crosstraining.

Graves and Tomlin (2003) analyze the benefits from process flexibility in multistage supply chain setting, where "multistage" means that more than one station in the process are considered. A two-stage sequential decision process is considered: decide flexibility configurations of the process first and then allocate production capacity to meet demand. In the first stage, demand is a random vector with a known distribution, and the flexibility configuration is determined by minimizing the expected total shortfall, which is defined as the amount of demands that cannot be met. In the second stage, the system allocates its capacity given the flexibility configuration solved in the first stage and the realized demand. Hopp et al. (2004) model a serial production line and test various crosstraining configurations for a set of dynamic allocation policies for crosstrained workers. They ask the following question: "how to decide which skill(s) are strategically more desirable for workers to gain", and "how to coordinate these workers to respond dynamically to congestion?" Their goal is to minimize the WIP-to-throughput ratio for serial production lines. Iravani et al. (2005), define measures called the structural flexibility indices to characterize in a very general way the flexibility provided by a particular crosstraining arrangement. In their setting capacity is assumed to have been set to be sufficient "on average". Flexibility enables the system to respond to various stochastic demand shocks. Through numerical experiments, on both parallel and serial production lines, they are able to show that characteristics of the worker skill matrix can give good indications of "a system's ability, provided by its structure of multi-capability sources to reallocate production to respond to change in demand". All of these papers have focussed exclusively on crosstraining as the source of flexibility and on the worker allocation phase. While they have attempted to draw conclusions about the best skill mix they have assumed capacity is given and have ignored scheduling and other forms of worker flexibility.

In this paper, we attempt to bring together some of the various threads in the recent literature on workforce flexibility in a more holistic manner. Our goal is to illustrate the linkages between the capacity setting phase and the allocation phase in terms of determining the benefits of flexibility. We use the model of Berman et al. (1997) to determine the number of workers of each type (labor capacity) and their schedules to minimize the cost of

processing an estimated workload and arrival pattern. We view this first stage optimization as "short-term". Given the pool of workers determined in the first step we then reoptimize the schedule and allocation of workers to best process a perturbed work arrival pattern that recurs over time. We define this second stage optimization as "long-term".

3 Model Description

As discussed above we analyze a two-stage staffing process. In the first stage the manager uses an estimate of demand and chooses the staff to minimize the sum of staffing and WIP costs. This is done using the model of Berman et al. (1997). In the recurring second stage the same model is used with slight modification to reallocate the existing workers to minimize the WIP costs.

3.1 Summary of Berman et al (1997)

Here we summarize the model formulated in Berman et al (1997). While an MIP might be more accurate for a large workforce an LP formulation is sufficient to evaluate the value of various forms of flexibility. Consider a high volume factory containing n workstations. The factory is characterized by the workstation routing rules, which is represented by the routing probability p_{ij} , the fraction of jobs routed from station i to j for $i, j = 1, \dots, n$, and the flexibility policy which includes the number of starting times (set \mathbf{H}), the part-time allowance ratio (α), the job switch ratio (σ), the skill levels (β_{kj}), and the buffer size (γ_{jt}). These parameters and others are defined in Table 1 and are discussed later. Decision variables are $X_{k(j_1, j_2)h\tau}$, the number of type k workers working the first half of the shift at station j_1 and the second half at station j_2 for a shift length h that starts at time period τ . Constraints include the workflow conservation equations (equations 1, 2, and 3), the productivity equations (equations 4, 5a and 5b), the buffer constraints (equation 6), the part-time ratio constraint (equation 7), and the job switch ratio constraint (equation 8).

Let the deterministic demand b_{jt} be the number of jobs that arrive exogenously to station j in the beginning of time t . Denote I_{jt} as the new jobs at station j at the start of period t , and $O_{i(t-1)}$ as the output at station i at the end of period $t - 1$. Then the workflow

conservation for arrivals states that the number of new jobs at station j is equal to the exogenous input and the jobs routed to station j from other stations.

$$I_{jt} = \begin{cases} b_{jt} + \sum_{i \neq j} p_{ij} O_{i(t-1)}, & t = 2, 3, \dots, T, \\ b_{j1} + \sum_{i \neq j} p_{ij} O_{iT}, & t = 1. \end{cases} \quad (1)$$

Note that in this model time is cyclical so that what is leftover in the system at the end of period T is carried over into period 1. This means that we are conducting an equilibrium analysis. The departure flow has to be conservative as well, i.e. the remaining number of jobs equals the sum of reworks and unprocessed jobs. Denote $R_{j(t-1)}$ as the residual works remaining at station j from period $t - 1$. The departure flow conservation becomes:

$$R_{j(t-1)} = p_{jj} O_{j(t-1)} + [Y_{j(t-1)} - O_{j(t-1)}] \quad (2)$$

where, Y_{jt} = the number of jobs in the buffer at station j at the start of period t . Finally, the conservation of buffers results in:

$$Y_{jt} = I_{jt} + R_{j(t-1)}, \quad (3)$$

where the right hand side is the sum of the new jobs arriving at period t and residual jobs from period $t - 1$.

The productivity constraints consider both the labor capacity and the loading. Denote the maximum number of jobs that can be processed by the personnel assigned to station j during time t as the W_{jt} , i.e. labor capacity, which is the product sum of all scheduled workers and the corresponding individual capacity. It can be expressed as:

$$\begin{aligned} W_{jt} = & \sum_{k \in M_j} \sum_{h \in \mathbf{H}} \left\{ \sum_{\tau \in F_{th}} \beta_{kj} X_{k(j,j)h\tau} + \sum_{n_2 \in W_k} \sum_{\tau \in G_{th}} \beta_{kj} X_{k(j,n_2)h\tau} \right. \\ & \left. + \sum_{n_1 \in W_k} \sum_{\tau \in Q_{th}} \beta_{kj} X_{k(n_1,j)h\tau} \right\}, \end{aligned} \quad (4)$$

where M_j is the set of qualified worker types for station j , W_k is the set of stations for

which worker type k is qualified, F_{th} is the set of starting times such that a shift of length h includes period t , G_{th} is the set of starting times such that the first half of shift length h includes period t , and Q_{th} is the set of starting times such that the second half of shift length h includes period t . However the actual productivity of station j at time t is not W_{jt} but O_{jt} , which satisfies:

$$O_{jt} \leq Y_{jt} \text{ and} \quad (5a)$$

$$O_{jt} \leq W_{jt}, \quad (5b)$$

where equation 5a states that the output cannot exceed the number of jobs that are required to be processed at station j at time period t , and equation 5b means that the output cannot exceed the capacity at station j at time period t .

Labor capacity is the sum of scheduled workers times worker capacity over all starting times. Having a higher number of starting time allows the system to take advantages of the overlapped workforce, e.g. a worker starting at 8 a.m. overlaps a worker starting at 12 p.m. given that both work an eight-hour shift. When the load peak in this station is between 12 p.m. to 4 p.m., it is more efficiently covered than if the shifts only started at 8 a.m. and 4 p.m. Having cross-trained workers expands the scheduling pool and so the system is less constrained and the optimization potentially returns a better result.

For the rest of flexibility options, three more constraints (for buffer size, part-time ability, and job switch ability) are considered:

$$Y_{jt} \leq \gamma_{jt}, \quad (6)$$

$$(1 - \alpha) \sum_{k=1}^K \sum_{(j_1, j_2) \in A_k} \sum_{\substack{h \in \mathbf{H} \\ h \geq 8}} \sum_{\tau \in \mathbf{ST}} X_{k(j_1, j_2)h\tau} - \alpha \sum_{k=1}^K \sum_{(j_1, j_2) \in A_k} \sum_{\substack{h \in \mathbf{H} \\ h < 8}} \sum_{\tau \in \mathbf{ST}} X_{k(j_1, j_2)h\tau} \geq 0, \text{ and} \quad (7)$$

$$(1 - \sigma) \sum_{k=1}^K \sum_{\substack{(j_1, j_2) \in A_k \\ j_1 \neq j_2}} \sum_{h \in \mathbf{H}} \sum_{\tau \in \mathbf{ST}} X_{k(j_1, j_2)h\tau} - \sigma \sum_{k=1}^K \sum_{(j, j) \in A_k} \sum_{h \in \mathbf{H}} \sum_{\tau \in \mathbf{ST}} X_{k(j, j)h\tau} \leq 0, \text{ where} \quad (8)$$

Where part-time workers work less than 8 hours and A_k is the set of all (j_1, j_2) that are feasible for worker k , γ_{jt} is the capacity of buffers at station j during period t , α is the minimal percentage of full time workers, and σ is the maximal fraction of job switched workers.

3.2 First Stage: Staffing

Denote the estimate of the demand arrival pattern as b_{jt}^1 , the number of jobs that arrive exogenously to station j in the beginning of period t . The decision variables are $X_{k(j_1, j_2)h\tau}$, with corresponding unit staffing cost $C_{k(j_1, j_2)h\tau}$. The WIP at each station j at the end of period t is given by R_{jt} . We define c to be the unit cost per period of WIP. Therefore, by combining the staffing cost $\sum_{\text{All}} X_{k(j_1, j_2)h\tau} C_{k(j_1, j_2)h\tau}$ and the WIP cost $\sum_j \sum_t cR_{jt}$, we obtain the objective in the first stage

$$\min \sum_{\text{All}} X_{k(j_1, j_2)h\tau} C_{k(j_1, j_2)h\tau} + \sum_j \sum_t cR_{jt}. \quad (9)$$

Therefore, for a given flexibility policy we apply the constraints in Section 3.1 to this objective function and obtain the worker requirement in the first stage, where the work requirement includes not only a schedule for the first stage demand but also the total amount of workers that remains unchanged in the second stage.

3.3 Second Stage: Rescheduling

In the second stage, the firm observes a demand shock that perturbs the demand pattern. Let b_{jt}^2 be the new demand pattern. Because the recruiting and training new workers is time consuming, the firm must attempt to satisfy demand with the existing workforce and can only rearrange the schedule and allocation of workers to tasks. As a result in the second

Table 1: Table of notations

$X_{k(j_1, j_2)h\tau}$	= number of type k workers working the first half of the shift at j_1 and the second half at j_2 for shift length h that starts at time period τ .
$C_{k(j_1, j_2)h\tau}$	= cost of type k workers working the first half of the shift at j_1 and the second half at j_2 for shift length h that starts at time period τ .
n	= total number of workstations in the factory
I_{jt}	= total quantity of new work presented to station j at the start of period t
R_{jt}	= total work remaining at station j at the end of period t
Y_{jt}	= units of work in the buffer at station j at the start of period t
O_{jt}	= output of station j during period t
W_{jt}	= maximum number of jobs that can be processed by personnel assigned to station j during period t
T	= total number of equal length time periods during a working day
b_{jt}	= the number of units of work that arrive exogenously to station j and is presented there at the beginning of time period t , $t = 1, 2, \dots, T$
p_{ij}	= fraction of jobs processed at station i that are routed next to station j , $j = 1, \dots, n$
\mathbf{H}	= the set of all allowed shift lengths
\mathbf{ST}	= the set of all allowed starting times for shifts
β_{kj}	= units of work that worker type k can process per time period at station j
(j_1, j_2)	= a pair of station j_1 and j_2 , representing a worker's workstation assignments for the first and second halves of her shift, respectively
A_k	= the set of all (j_1, j_2) that are feasible for worker type k at station j_1 and the second half at j_2 for a shift length h that starts at time period τ ;
$k = 1, \dots, K$, $(j_1, j_2) \in A_k$, $h \in \mathbf{H}$, $\tau \in \mathbf{ST}$	
γ_{jt}	= capacity of buffer j during period t , measured in units of work
α	= the maximum fraction of workers that can be part-time allowance ratio.
σ	= the maximum fraction of workers that can switch tasks mid-shift.

stage we view labor costs as fixed and the firm minimizes WIP. To maintain feasibility total demand is kept the same as in the Stage 1 problem and the buffer constraints are relaxed. We define similar decision variables as in stage 1, $\tilde{X}_{k(j_1, j_2)h\tau}$, the number of type k workers working the first half of the shift at station j_1 and the second half at j_2 for a shift length h that starts at time period τ , and the corresponding remaining jobs in station j in the end of period t is \tilde{R}_{jt} . The objective is:

$$\min \sum_j \sum_t c\tilde{R}_{jt} \quad (10)$$

The constraints in this stage include the work flow conservation equations, the productivity equations, the work balance equations including reworks, the part-time ratio constraint (which is automatically satisfied), the job switch ratio constraints, and the workforce balance equations. In order to capture the fact that the workforce is unchanged from Stage one we introduce Equation (11) which sets the number of workers for each type equal to the same number as in the first stage.

$$\sum_{(j_1, j_2) \in A_k} \sum_{h \in \mathbf{H}} \sum_{\tau \in \mathbf{ST}} \tilde{X}_{k(j_1, j_2)h\tau} = \sum_{(j_1, j_2) \in A_k} \sum_{h \in \mathbf{H}} \sum_{\tau \in \mathbf{ST}} X_{k(j_1, j_2)h\tau} \text{ for all } k = 1, \dots, K. \quad (11)$$

The objectives and constraints in both stages are summarized in Table 2.

The optimal solution in the second stage shows the best the system can react to the demand pattern change (or demand shock). By comparing the optimal results across all the flexibility policies, we are able to see which flexibility combinations hurt the systems ability to react to such shocks. In the next section we conduct extensive numerical experiments, to evaluate different flexibility policies in this way.

4 Computational Examples and Results

4.1 Numerical Scenarios

For the purposes of our numerical experiments we analyze three different work configurations: a network similar to the example in Berman et al. (1997), a parallel system similar to that

Table 2: Objective and constraints in both stages

	Staffing	Rescheduling
Objective	minimizing staffing cost and WIP cost	minimizing WIP cost
Workflow conservation	Yes (Eq. 1-5b)	Yes (Eq. 1-5b)
Buffer size	Yes (Eq. 6)	No
Part-time ratio	Yes (Eq. 7)	Yes (Eq. 7)
Job switch ratio	Yes (Eq. 8)	Yes (Eq. 8)
Number of workers	No	Yes (Eq. 11)

analyzed in Jordan and Graves (1995), and a serial system similar to that studied in Iravani et al. (2005). We assume that we are modeling a day of operations broken into 48 half hour periods. The demand estimate for the Stage 1 problem is a uniform arrival of 200 units of work per period. To simplify the analysis we assume that all workers regardless of their skill mix or work hours are paid at the same rate of \$30/hour and if they are qualified to work at a work station their productivity is 60 units per hour.

We study the following set of flexibility types:

- Number of starting times (3, 4, 6, or 8)
- Part-time allowed (yes or no)
- Mid-shift jobswitching allowed (yes or no)
- Crosstraining level (“None”, “Pair”, or “All”)
- Buffer size at each station (200 or 400)

Overall, we have $4 \times 2 \times 2 \times 3 \times 2 = 96$ flexibility policies. For the number of starting times, we evenly distribute the starting times over the 48 periods with the first one being 1. For example, the starting times for 8 starting times are $ST = 1, 7, 13, 19, 25, 31, 37,$ and 43. A full time shift is 8.5 hours, which includes a half hour break, and two types of part-time shifts are considered: 6.5 and 4.5 hours. When part-time workers are allowed, we limit the percentage of part-time workers to be no higher than 20% of overall workers. Similarly, if jobswitching is allowed, we apply a limit of 30% on the percentage of workers who switch their jobs in the middle of the shifts. One may question the limits on the percentages of

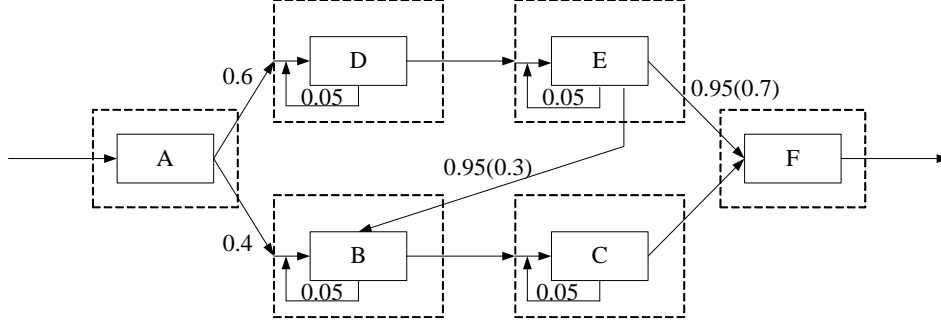


Figure 1: Six station process with network configuration

Worker Type	Crosstraining Levels		
	None	Pair	All
1	A	A, D	A, B, C, D, E, F
2	B	D, E	N/A
3	C	E, F	N/A
4	D	A, B	N/A
5	E	B, C	N/A
6	F	C, F	N/A

Table 3: Skill mapping for different crosstraining levels in the network configuration

part-time workers and/or mid-shift jobswitching. However, in practice we often see the limits, and hence it is more realistic to apply these limits. We note again that we assume the productivity and the labor cost to be the same for each type of worker. This is not a restriction of the model but rather a way to reduce the number of cases considered in this analysis. Including a pay difference between more and less cross-trained workers does not qualitatively alter the results. We also analyze three different cases for the WIP cost, \$0.1, \$0.25, or \$0.5 per unit per time period.

The network configuration has six stations and is shown in Figure 1. Icons A to F represent the six stations, the directed arcs represent the workflow, and the values above the arcs are the percentages of jobs directing to the next stations. Note that there is a 5% rework rate in stations B, C, D, and E. We assume that jobs only enter the system from station A and leave the system from station F. Workers can be classified in the three training levels: “None”, “Pair”, or “All”, which are summarized in Table 3.

The serial configuration is similar to the example in Iravani et al. (2005). We consider

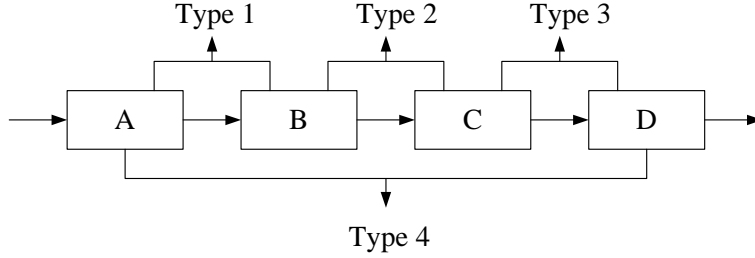


Figure 2: Serial configuration showing paired chain crosstraining

four stations, where all jobs have to be handled by all four stations in the same order (Figure 2). The jobs only enter the system from station A and exit the system from station D. Consider three training levels: "None", "Chained pairs", and "All". The first one means that there are four worker types and each type can only work at one station. Chained pairs (shown in Figure 2) is similar to the definition in Jordan and Graves (1995), in which "*A chain is a group of products and plants which are all connected, directly or indirectly, by product assignment decisions. In terms of graph theory, a chain is a connected graph. Within a chain, a path can be traced from any product or plant to any other product or plant via the product assignment links.*" (P580). Finally, "All" means that all workers are capable of performing the tasks in all four stations.

The third, parallel configuration is similar to the one in Jordan and Graves (1995), in which six parallel factories are considered (Figure 3). The exogenous inputs are evenly¹ distributed to all six factories. Three training levels are considered: "None", "Chained pairs" (Figure 3), and "All", where workers are capable of performing tasks in all six stations.

4.2 First Stage: Staffing

Given a flexibility policy out of the 96 policies defined in Section 4.1, a staffing schedule for each station can be calculated based on the model in Section 3.2, and the objective gives the total operating cost for this flexibility policy. In order to simplify the presentation, we combine the part-time and the jobswitching policies into one index, and assign another index to the number of starting times (Table 4). Next, by summing these two numbers, we

¹The demand can also be unevenly distributed to the six factories, but the results are similar.

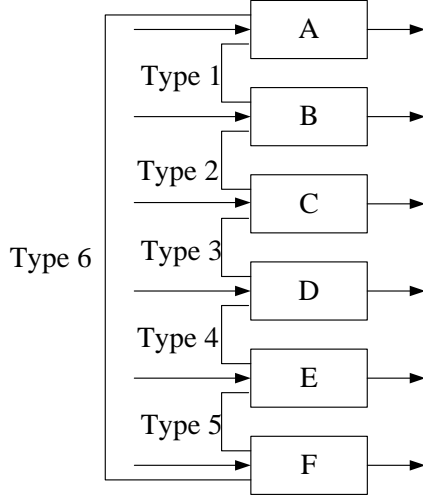


Figure 3: Parallel configuration showing paired chain crosstraining

Part-time	Mid-shift jobswitching	index	No. of starting times	index
No	No	2	3	30
No	Yes	4	4	40
Yes	No	6	6	50
Yes	Yes	8	8	60

Table 4: Indices for flexibility scenarios

obtained an overall index for a flexibility policy that consists of the part-time ability, the job switch ability, and the number of starting times. For example, for a flexibility policy in which four starting times are consider (40), and part-time is allowed but mid-shift jobswitching is not allowed (6), the corresponding index is $40 + 6 = 46$.

4.3 Second Stage: Rescheduling

After a pool of workers is hired, trained, and scheduled based on the first stage demand, we proceed to the second stage rescheduling. In the second stage, the original uniform demand arrival pattern is perturbed in three different ways. For two types of perturbation we change the arrival pattern into a peak shape which varies in time, t_p , or in peak magnitude, H . The third type of demand shock involves adding random noise to the original uniform pattern.

The first type of demand shock captures the variation when the true demand is unimodal (as opposed to uniform) with a peak of 500 units of work, while the total number of jobs

remains the same. Twenty scenarios are generated with the peak position ranging from $t_p = 14$ to $t_p = 33$, where each scenario has the same shape as the one shown in Figure 4a. In other words, the total number of arrivals and the maximum number of arrivals in any one period are the same across the twenty scenarios. Given the original staffing level for each flexibility policy, we can reschedule the staff in order to minimize the WIP.

The second demand shock type represents the situation where the peak is located at $t_p = 25$ but varies in magnitude. We randomly generate scenarios such that in each scenario, the peak locates at $t_p = 25$ with the total number of jobs being 9,600 jobs, but the maximum number of arrivals is drawn from a normal distribution with a mean of 450 jobs and a standard deviation of 100 jobs.

The third demand shock type assumes that the first-stage demand estimate is relatively accurate but is not completely deterministic. For this type of demand shock, twenty scenarios are also considered where noise is randomly added to the first-stage (base) demand but the total amount of jobs is kept the same. In each scenario, we first generate 24 noise values from a normal distribution with a mean of 0 and a standard deviation of 50. Second, we randomly pair the 48 time periods into 24 pairs, e.g., t_i and t_j , for $i, j \in \{1, 2, \dots, 48\}$. By adding a noise value to the demand at t_i and subtract the same noise to the demand at t_j for all 24 noises, we complete the process of generating the demand with noise and maintain the total number of jobs the same.

Examples of the three demand shocks are given in Figures 4. Figure 4a represents the demand with the peak at $t_p = 25$, Figure 4b shows the demand where the maximum number of arrivals changes from 200 to 305, and Figure 4c represents the first stage demand with noise. For each type of demand shocks we calculate the mean and ninetieth percentile of the sum of the staffing and the WIP costs (denoted as the total Stage 2 cost) across 20 scenarios for each flexibility policy.

4.4 Flexibility Effects

In this subsection we explore flexibility effects for both the first stage and second stage problems. We find that the results are similar across the configurations, demand shocks and unit WIP costs. Therefore, we only show a representative set of results for the network

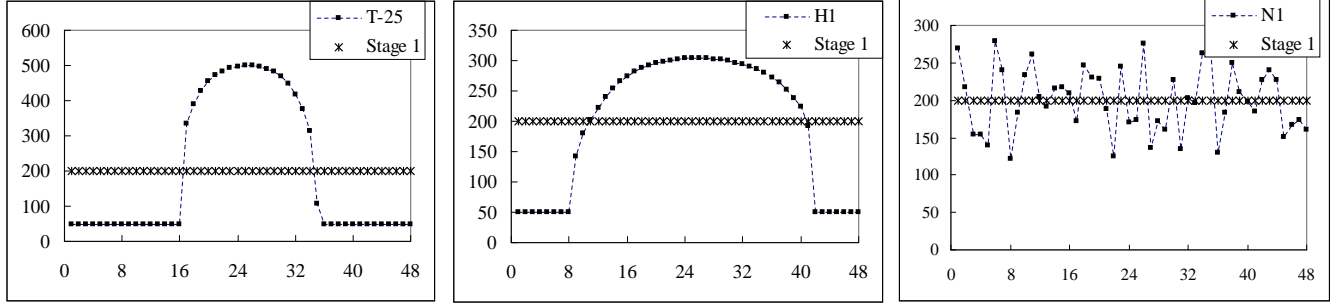


Figure 4: Second-stage demand arrival pattern for (a) shock t_p (left); (b) shock H (middle); (c) shock N (right)

configuration under demand shock t_p . We report more results in the online appendix.

4.4.1 First Stage

In Figure 5 we plot the total Stage 1 cost for the network configuration. In Figures 6a and 6b we respectively separate out the staffing and WIP costs. The solid (dashed) lines are the costs when the buffer size is 200 (400). The three crosstraining levels are represented by three markers: diamond for “None”, star for “All”, and triangle for “Pair”. The index on the horizontal axis can be mapped into the flexibility policy that includes the number of starting times, the part-time ability, and the mid-shift jobswitching ability (Table 4). For example, the three coinciding circled points, in Figure 5, are the total Stage 1 costs for the flexibility policy with buffer size of 200 (solid line), four starting times and part-time workers (index 46), and crosstraining levels of “None” (diamond marker), “All” (star marker), and “Pair” (triangle marker). First we observe that there is essentially no benefit from crosstraining evidenced in the results. The reason for this is that the work arrival profile is deterministic and uniform. When we study the Stage 2 problem we do find benefits from this form of flexibility.

Second, we also find that buffers are a powerful tool in lowering total Stage 1 costs. The costs are reduced because buffering allows the manager to accumulate work in time so that worker shifts can be closely matched to the time at which work is done. By comparing the results of a buffer size of 200 (solid lines) and the one of 400 (dashed lines), we see that doubling the buffer size lowers the total Stage 1 cost significantly (Figure 5). Figure 6a shows

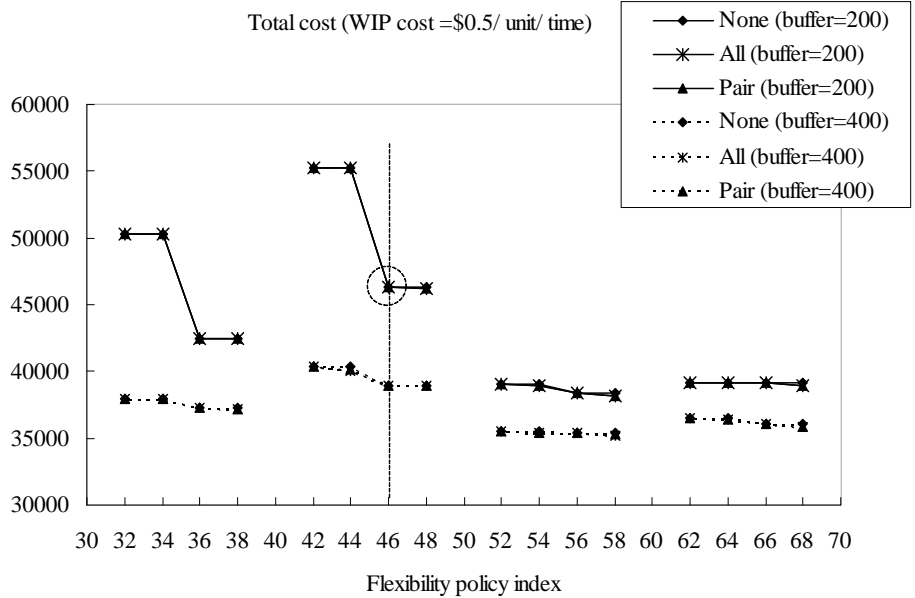


Figure 5: Total Stage 1 cost for network configuration (WIP cost of \$0.5/unit/time)

that this reduction is coming from the savings on staffing, especially when the starting times are restricted (30's and 40's). WIP costs may actually increase somewhat due to buffering (Figure 6b).

Third, flexibility in the number of shift start times can also reduce cost significantly. By comparing the 30's and 50's or 40's and 60's (Figure 5 and Figure 6a), whether the factory has a tight buffer size or not, doubling the number of starting times significantly decreases the staffing cost and the total Stage 1 cost. The change in the total staffing cost from doubling the number of starting times is almost the same as the change of doubling the buffer size (the difference between the solid and dashed lines in 30's). More start times is another mechanism for matching labor to work in time.

Finally, allowing part-time workers lowers the staffing cost significantly, when a tight buffer size and a small number of starting times are considered. For example, in Figure 5 when the buffer size is 200 (solid lines) the total Stage 1 cost of 36 (part-time allowed) is significantly lower than the one of 32 (part-time is not allowed). However, if the buffer size is doubled (see the pair (32,36) on the dashed line), including this flexibility type does not lower the staffing cost significantly. Similarly, if the number of starting time is relaxed

(compare the pairs (32, 36) and (52, 56) or (42, 46) and (62, 66)), the part-time ability also does not affect the staffing cost significantly. Understandably, hiring a part-time worker can avoid the excess capacity or cost incurred by a full-time worker when the number of starting time is small and buffer size is tight. For this type of process, in which the demand has to be fulfilled quickly but the starting time set is small, part-time workers allow the firm to handle the demand peak without paying full-time hours. However, when buffer constraints are relaxed, jobs can wait until the next available worker and hence the benefit of having part-time workers is greatly reduced.

Thus far, we discussed the benefits from all five flexibility types when there is no demand fluctuation. In the next section, we show that flexibility can lead a firm to staff with too little slack to be flexible to demand shocks, thus leading to a higher total cost, i.e., the sum of the staffing and WIP costs.

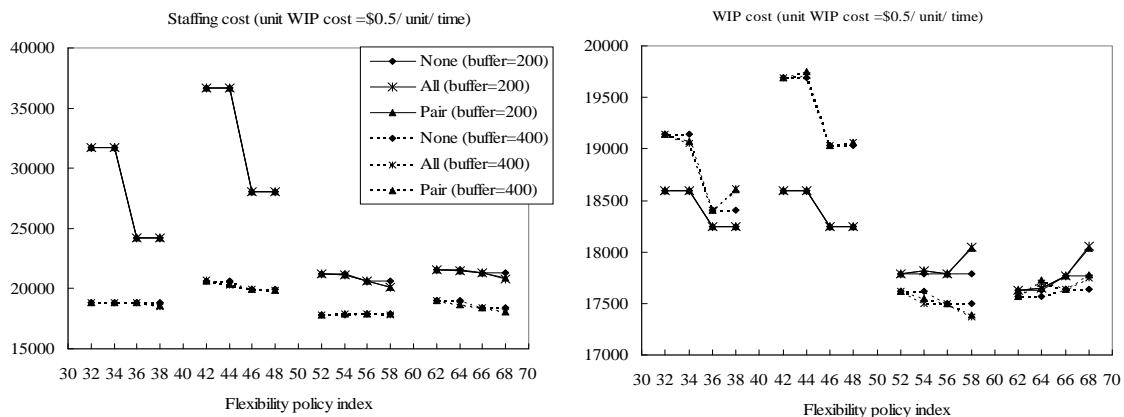


Figure 6: (a) Staffing cost (left) and (b) WIP cost (right) for the network configuration in the first stage (WIP cost of \$0.5/unit/time)

4.4.2 Second Stage

In Figure 7a(b), we plot the mean of the total Stage 2 cost across 20 demand scenarios for the network configuration under shock t_p , when the unit WIP cost is \$0.5/unit/time and buffer size is 200 units (400 units). We confirm the finding in previous papers that a small degree of crosstraining can extract almost all the benefits of crosstraining if the skills are “chained”. Jordan and Graves (1995) introduced the concept of pairwise skill-chaining for a

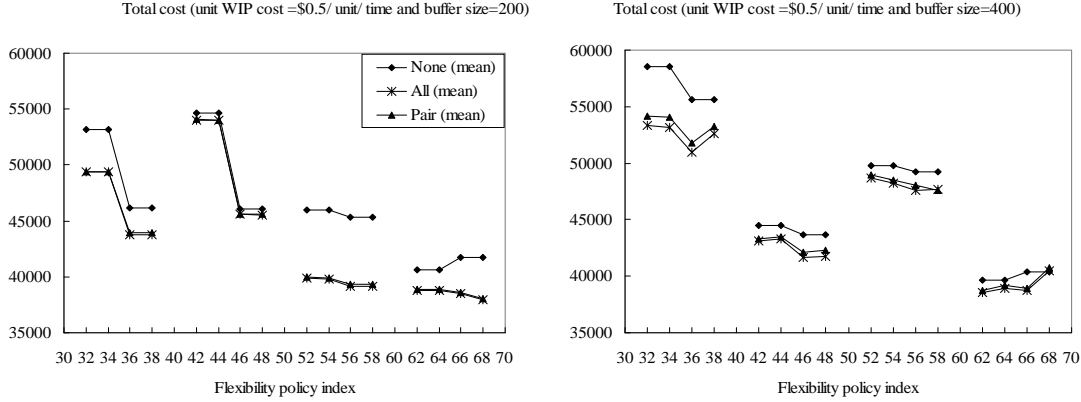


Figure 7: Total Stage 2 cost for the network configuration under shock t_p with unit WIP cost of \$0.5/unit/time and (a) buffer size=200 (left); (b) buffer size=400 (right)

parallel system. In the context of a network, a skill chain is less well defined. Nonetheless the pairwise crosstraining configuration we implement is very effective. For example, Figures 7a(b) show results for the network configuration. From the figure, we see that the training level “Pair” (triangle marker) performs almost the same as the training level “All” (star marker). Therefore in the following we only look at full crosstraining or no crosstraining cases in our comparisons.

A more flexible system does not necessarily outperform a less flexible one in terms of total Stage 2 costs in the long run. For example, in Figure 7b, with eight starting times, the system with both jobswitching and part-time (68) has a higher total cost than those systems without jobswitching and/or part-time (62, 64, and 66). Figure 8 separates the total cost into the staffing (Figure 8a) and WIP costs (Figure 8b). In Figure 8a, the staffing cost for 68 is the lowest among the four, but it has the highest mean WIP cost when demand fluctuation is considered. As a result, when the factory is more flexible, it tends to staff with too little slack and hence it cannot effectively respond to demand shocks. On the other hand, for those systems that are less flexible, the staffing level is higher, and thus they have potential to perform better in terms of the total Stage 2 cost. In other words, the apparent immediate benefits from flexibility (a reduction in staffing cost) may lead to a higher total cost as the work demand profile evolves over time.

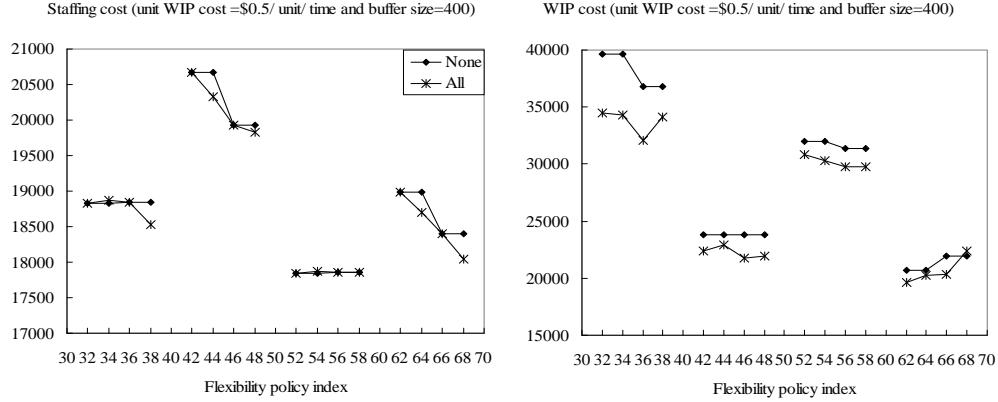


Figure 8: (a) staffing cost for the network configuration under shock t_p with a buffer size of 400 units and a unit WIP cost of \$0.5/unit/time (right); (b) staffing cost for the network configuration under shock t_p with a buffer size of 400 units and a unit WIP cost of \$0.5/unit/time (left).

4.4.3 Flexibility Effectiveness

In the previous section we have shown differences between the short-term and long-term effects of various types of flexibility. In this section we try to quantify the long-term effectiveness of each type of flexibility. To do this we borrow from the methodologies of design of experiments (DOE) analysis (Montgomery 2004). The three main components in a DOE analysis are the response (output), the factors (inputs), and the levels (values of factors). We change the value of a factor in order to observe the effects (responses) due to the factor's change. For our purposes the response is the mean total Stage 2 costs across the demand realizations. The factors are the flexibility types. The levels are the amount of each type of flexibility employed. In order to have binary levels for each flexibility type we bifurcate the start times into three vs. six and four vs. eight start times².

For each configuration and demand perturbation type we calculate $(E^+ - E^-)/2$ for each factor. Where E^+ is the average of the responses across scenarios when the factor level is high and E^- is the average of the responses when the factor level is low. Thus, the effectiveness of a flexibility type is defined as the mean improvement in the total Stage 2 cost when the

²Depending on the demand pattern, having 4 starting times is not necessarily more effective than having 3 starting times whereas doubling the number of start time with a similar distribution within the day is increasing flexibility

flexibility is included.

As an example suppose that our goal is to observe the effect of doubling the number of starting times from 3 to 6 for the network configuration under shock t_p , when the buffer size is tight. First, we calculate the average of the total Stage 2 costs when the number of starting times is 3 (E^-) and the average when the number of starting times is 6 (E^+). Note that when the number of starting times is 3 (or 6), we have $2 \times 2 \times 2 = 8$ policies, in which the 2's stand for the crosstraining level ("None" or "All"), jobswitching ability (allowed or not), and part-time ability (allowed or not). Then the effect of doubling the number of starting times from 3 to 6 is $(E^+ - E^-)/2$.

First we conduct four distinct DOE analyses identified by whether the start time levels are three vs. six or four vs. eight and whether the buffer size is B=200 or B=400. For these four DOE analyses, four factors are considered in each set of experiments: training level (TL), number of starting times (ST), part-time (PT), and job switching (JS). For the training level factor, we set the factor to high if the training level is "All", and low if it is "None". For the number of starting times, we set (3,6) or (4,8) as the low and the high levels. Similarly, we set the part-time factor as high if we allow part-time work and low otherwise. Because jobswitching influences the results only when the crosstraining level is "All", we consider only the cross effect of training level and jobswitching in the DOE analyses instead of the main effect of jobswitching.³ Therefore, jobswitching factor is low if jobswitching is not allowed or if training level is "None", and high if both jobswitching is allowed and the training level is "All". Given the above definitions each calculation of the effect $(E^+ - E^-)/2$ is performed over 16 flexibility policies.

Finally, we calculate the overall buffer size effect by performing an additional set of DOE analysis for (3,6) and (4,8) around the factor B, which is named as "overall B effect" in the results (see Table 5 for examples). For this DOE analysis, we include all 64 flexibility policies. When we calculate the negative effect(E^-), we average 32 policies with B = 200, and when we calculate the positive effect(E^+), we average another 32 policies with B = 400. The overall B effect is half of the difference between the positive and negative effects, i.e.,

³Although other cross effects, such as the number of starting time and the part-time ability, may affect the system, we only focus on which flexibility type is more effective instead of which "combination" of flexibility types is more effective. Thus we discard all other cross effects.

ST(3,6)					ST(4,8)				
Unit WIP cost		0.1	0.25	0.5	Unit WIP cost		0.1	0.25	0.5
B=200	TL	-599	-1200	-2301	B=200	TL	-314	-495	-802
	ST	-3062	-2966	-2765		ST	-4996	-4978	-5119
	PT	-1924	-1673	-1739		PT	-2264	-2111	-2061
	JS	-486	-845	-1536		JS	-310	-411	-586
B=400	TL	-283	-798	-1503	B=400	TL	-245	-421	-613
	ST	-902	-1777	-3011		ST	-663	-1277	-1840
	PT	-81	-475	-717		PT	-143	3	-84
	JS	-172	-603	-914		JS	-94	-119	-212
Overall B effect		-1391	689	3222	Overall B effect		-3077	-2351	-1768

Table 5: The main and cross effects of the network configuration under shock t_p

$$(E^+ - E^-)/2.$$

DOE Analysis Results Table 5 is a typical table from the analysis for the network configuration and t_p type demand shock. As discussed in the previous section, we only compare the effects of TL, ST, PT, and JS. Because we are minimizing cost a more negative number is considered more cost effective.

The number of starting times reduces the total Stage 2 cost the most regardless of the buffer size (B=200 or 400). This conclusion is valid across all the policies for the network (for example, see Table 5) and serial configurations. Some counter examples exist in the parallel configuration, in which jobs are processed in *one* station instead of sequential stations. This special structure greatly reduces the benefit of a large number of starting times.

Second, the ST effect increases with the unit WIP cost when B=400, e.g., in three vs. six start times, the effects increases from -902 to -1777 to -3011 in Table 5. When the WIP is costly (a high unit WIP cost), the factory staffs more than the ones with a low unit WIP cost. Combining this excess workforce and a large number of starting times, the factory can reallocate workforce more efficiently to the stations where needed. As a result, the ST effect is more significant (saves more on the total Stage 2 cost) for the factory with a high unit WIP cost than for the factory with a low one. But when a tight buffer size is considered, the staffing decision is more likely to be restricted by buffer constraints, and hence this increasing ST effect disappears. Furthermore, when the unit WIP cost is low, the ST effect is greatly reduced when B is relaxed (for example, from -3062 to -902 in (3,6) in Table 5). When jobs

can be carried over to the next period, the factory accumulates jobs and matches labor to work in time. Therefore, a high number of starting times does not lower total costs as much as the situation in which jobs cannot be carried over.

Finally, the overall buffer size effect decreases with respect to the unit WIP cost. This confirms that relaxing buffer size is appropriate when the unit WIP cost is low. However, it may lead to a high total Stage 2 cost when the unit WIP cost is high (from -1391 to +3222 in (3,6) in Table 5).

4.4.4 Flexibility Robustness

In previous section, we investigate the aggregated effect of each factor and see the effectiveness of each factor. In this section, we discuss the robustness of each factor by comparing the total Stage 2 costs when a factor's level is changed from low to high. If the result of a high level has a lower total Stage 2 cost than the one with a low level, we consider this as a robust result. For example, if the goal is to investigate the robustness of allowing jobswitching for the network configuration when $B=200$ and unit WIP cost = \$0.5. We first change the jobswitching ability from No to Yes (the other factors remain unchanged), and then compare the total Stage 2 costs (before and after change), if the cost decreases for a high proportion of cases we view it as a robust cost saving mechanism. We perform this same comparison for $2(\text{training level}) \times 2(\text{part-time}) \times 4(\text{number of starting times}) = 16$ policies and 3 shocks (48 cases in total), and calculate the percentage of cases with cost decreases (out of 48) generating a robustness score. For example, the robustness score in this case is 96%, which means that 96% of times including jobswitching improves the system performance (lower total Stage 2 costs). See the online appendix for more detailed results of the robustness tests. For the purposes of our discussion we use 90% as a threshold for categorizing a form of flexibility as robust⁴. In Figure 9, the shadowed cells are the cases where the designated form of flexibility percentage is not robust; namely, less than 90% of the 48 combinations of policies and shocks lead to cost savings.

We find that jobswitching (Figure 9a) and part-time (Figure 9b) flexibility are not particularly robust. When the buffer is small ($B=200$) jobswitching, is robust (though less so

⁴We acknowledge this threshold is somewhat arbitrary but it forms a natural break for our results.

for the parallel configuration), while if $B=400$, jobswitching is not robust. Part-time worker is not robust regardless of the buffer size.

Crosstraining is a robust flexibility type, except when the parallel configuration is considered and when $B=200$ (the staffing is more constrained by buffer size). When we change the crosstraining level from "None" to "All", all the comparisons of the total Stage 2 costs indicate that having a higher crosstraining level reduces the total Stage 2 costs (Figure 9c). When $B=400$ crosstraining is not always beneficial. The cases in which crosstraining makes cost increase have one point in common: jobswitching is allowed in the first stage. Since the system is flexible enough (due to the relaxed buffer size), applying both crosstraining and jobswitching makes the system rely too much on flexibility. However, if we exclude the jobswitching ability in the first stage, crosstraining benefits in the second are very robust. A quick note here, the system can improve the performance even more by excluding the jobswitching ability in the first stage but implementing it in the second stage.

The number of starting times is the one that is both effective (Figure 9) and robust (Figure 10d). The results (Figure 9d) show that the number of starting times has highest overall robustness scores among the four tests for the three configurations (Figure 9d: 99%, 99%, and 98%). Moreover, all the cases in which the number of starting is not robust allow part-time workers. When part-time is allowed, the system with a smaller number of starting times has to staff more than the one with a larger number of starting times. The difference in staffing level leads to a higher total cost for the system with larger number of starting times. This verifies again that a more flexible system does not necessarily have a lower total Stage 2 cost when demand fluctuation is considered.

Finally, doubling buffer size is not robust. When the unit WIP cost is high, having a loose buffer size leads to less staff in the first stage. When demand shocks occur, the factory fulfills demand by delaying jobs causing a high WIP cost and hence a high total cost (see Figure 10). When the unit WIP cost is low, doubling the buffer size is robust (96%, 100%, and 92%), but robustness decreases significantly when the unit WIP cost increases. Overall, buffer size, jobswitching, and part-time flexibility are not robust, but crosstraining without jobswitching and high number of starting times without part-time are robust combinations.

(a) Robustness test of allowing jobswitching							
	B=200			B=400			overall %
unit WIP cost	0.1	0.25	0.5	0.1	0.25	0.5	
% robustness for Network configuration	100%	100%	96%	79%	71%	69%	86%
% robustness for Serial configuration	100%	100%	100%	79%	71%	63%	85%
% robustness for Parallel configuration	100%	94%	94%	88%	81%	85%	90%
(b) Robustness test of allowing part-time workers							
	B=200			B=400			overall %
unit WIP cost	0.1	0.25	0.5	0.1	0.25	0.5	
% robustness for Network configuration	77%	77%	75%	81%	81%	79%	78%
% robustness for Serial configuration	83%	83%	83%	81%	81%	77%	82%
% robustness for Parallel configuration	83%	83%	83%	85%	69%	73%	80%
(c) Robustness test of having crosstraining (change from "None" to "All")							
	B=200			B=400			overall %
unit WIP cost	0.1	0.25	0.5	0.1	0.25	0.5	
% robustness for Network configuration	100%	100%	100%	96%	96%	98%	98%
% robustness for Serial configuration	100%	100%	100%	94%	96%	98%	98%
% robustness for Parallel configuration	100%	100%	100%	88%	88%	92%	94%
(d) Robustness test of doubling the number of starting times							
	B=200			B=400			overall %
unit WIP cost	0.1	0.25	0.5	0.1	0.25	0.5	
% robustness for Network configuration	100%	100%	96%	100%	100%	100%	99%
% robustness for Serial configuration	100%	100%	100%	100%	98%	98%	99%
% robustness for Parallel configuration	100%	100%	100%	98%	100%	92%	98%

Figure 9: (a) Robustness test of allowing jobswitching; (b) robustness test of allowing part-time workers; (c) robustness test of having crosstraining (change from "None" to "All"); (d) robustness test of doubling the number of starting times (3 to 6 or 4 to 8).

Robustness test of doubling the buffer size				
unit WIP=	0.1	0.25	0.5	overall %
% robustness for Network configuration	96%	73%	59%	76%
% robustness for Serial configuration	100%	77%	58%	78%
% robustness for Parallel configuration	92%	57%	51%	67%

Figure 10: Robustness test of doubling the buffer size (from B=200 to B=400)

5 Conclusion

By using a two-stage model to consider demand pattern changes, this paper investigates the distinction between the short-term and the long-term benefits from workforce flexibility, which includes number of starting times, part-time ability, jobswitching ability, crosstraining level, and buffer size. We also analyze the interaction between these different forms of flexibility showing their dependencies. We confirm what previous studies have shown about crosstraining, namely that a small amount of flexibility can achieve most if not all the benefits. However, more importantly we have shown that workforce flexibility is far richer than that when one considers the impact of flexibility on staff sizing decisions and multiple forms of flexibility simultaneously.

While crosstraining was found to be generally robust we also found that its robustness was diminished when combined with jobswitching. We also found that its benefits were not as strong as other forms of flexibility such as starting time flexibility. Increased buffer sizes, which give flexibility on the timing of processing, were effective when WIP penalties were low and could make other forms of flexibility insignificant for controlling staff costs. On the other hand for larger WIP penalties buffer flexibility was unreliable.

Our results suggest a set of rules of thumb for workforce planning. Work to create acceptance of starting time flexibility in the workforce and then set staff size in the planning phase assuming that buffers are small. Take advantage of buffers when actually scheduling worker after demand has been realized. If start time flexibility is not feasible then use part time work but not both. Finally, if part-time work is not possible use crosstraining but set staffing levels assuming jobswitching is not allowed.

It is important to note that in our study the uncertainty in demand was temporal. We did not consider uncertainty in the mix of work. As a result crosstraining came out as the weakest form of flexibility relative to the others that are themselves working time related. If the timing of work was less uncertain than the actual content, the relative effectiveness and robustness of each form of flexibility studied here would most likely be different. The modeling and analysis framework we have presented here could be easily applied to studying work mix uncertainty and that is our intention for future research.

References

- [1] Abernathy, W. J., N. Baloff, J. C. Hershey, S. Wandel (1973) A three-stage manpower planning and scheduling model—A service-sector example. *Operations Research* **21**(3) 693-711.
- [2] Bard, J.F. (2004) Staff Scheduling in High volume Service factories with Downgrading. *IIE Transactions*. **36**(10) 985-997.
- [3] Beach, R., A. P. Muhlemann, D. H. R. Price, A. Paterson, J. A. Sharp (2000) A review of manufacturing flexibility. *European Journal of Operational Research* **122** 41-57.
- [4] Berman, O., R. C. Larson, E. Pinker (1997) Scheduling workforce and workflow in a high volume factory. *Management Science* **43**(2) 158-178.
- [5] Browne, J., D. Dubois, K. Rathmill, S. P. Sethi, K. E. Stecke (1984) Classification of flexible manufacturing systems. *The FMS Magazine* **2**(2) 114-117.
- [6] Brusco, M. J. (2008) An Exact Algorithm for a Workforce Allocation Problem with Application to an Analysis of Crosstraining Policies. *IIE Transactions*. **40**(5) 495-508.
- [7] Campbell, G. M. (1999) Cross-utilization of workers whose capabilities differs. *Management Science* **45**(4) 722-732.
- [8] De Toni, A., S. Tonchia (1998) Manufacturing flexibility: a literature review. *International Journal of Production Research*, **36**(6), 1587-1627.
- [9] Ernst, A. T., H. Jiang, M. Krishnamoorthy, B. Owens, D. Sier (2004) An annotated bibliography of personnel scheduling and rostering. *Annals of Operations Research* **127**(1-4) 21-144.
- [10] Graves, S. C., B. T. Tomlin (2003) Process flexibility in supply chains. *Management Science* **49**(7) 907-919.
- [11] Hopp, W. J., E. Tekin, M. P. Van Oyen (2004) Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science* **50**(1) 83-98.

- [12] Iravani, S. M., M. P. Van Oyen, K. T. Sims (2005) Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science* **51**(2) 151-166.
- [13] Jordan, W. C., and S. C. Graves. (1995) Principles on the Benefits of Manufacturing Process Flexibility. *Management Science* **41**(4) 577-594.
- [14] Lenz, J. E. (1992) The need for both labor and machine flexibility in manufacturing. *Industrial Engineering* 122-23.
- [15] MacDuffie, J. P. (1995) Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry. *Industrial and Labor Relations Review* **48**(2) 197-221.
- [16] Merchant, M. E. (1983) Current status of and potential for automation in the metal working manufacturing industry. *Annals of the CIRP* **24**(2), 573-574.
- [17] Montgomery, D. (2004) *Design and Analysis of Experiments*. Wiley.
- [18] Newman, W. R., M. Hanna, M. J. Maffei (1993) Dealing with the uncertainties of manufacturing: flexibility, buffers and integration. *International Journal of Operations and Production Management* **13**(1) 19-34.
- [19] Sethi, A. K., P. S. Sethi (1990) Flexibility in manufacturing: A survey. *International Journal of Flexible Manufacturing Systems* **2**(4) 289-328.
- [20] Vokurka, R. J., S. W. O'Leary-Kelly (2000) A review of empirical research on manufacturing flexibility. *Journal of Operations Management* **18** 485-501.
- [21] Upton, D. M. (1995) Flexibility as process mobility: the management of plant capabilities for quick response manufacturing. *Journal of Operations Management* **12**(3-4) 205-224.