

This article was downloaded by: [University of Rochester]

On: 03 February 2012, At: 06:20

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



IIE Transactions

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/uiie20>

Can flexibility be constraining?

Edieal Pinker^a, Hsiao-Hui Lee^a & Oded Berman^b

^a Simon School of Business, University of Rochester, Rochester, NY, 14627, USA

^b Rotman School of Management, University of Toronto, Toronto, ON, Canada, M5S 3E6

Available online: 20 Nov 2009

To cite this article: Edieal Pinker, Hsiao-Hui Lee & Oded Berman (2009): Can flexibility be constraining?, IIE Transactions, 42:1, 45-59

To link to this article: <http://dx.doi.org/10.1080/07408170903113789>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Can flexibility be constraining?

EDIEAL PINKER^{1,*}, HSIAO-HUI LEE¹ and ODED BERMAN²

¹*Simon School of Business, University of Rochester, Rochester, NY 14627, USA*
E-mail: pinker@simon.rochester.edu

²*Rotman School of Management, University of Toronto, Toronto, ON, Canada, M5S 3E6*
E-mail: berman@rotman.utoronto.ca

Received August 2008 and accepted May 2009

Five common options for workforce flexibility and their robustness under uncertain demand are investigated. 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. Numerical results are presented that 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. The forms of flexibility that give robust benefits are identified and an analysis on how different forms of flexibility interact with each other is performed.

[Supplemental materials are available for this article. Go to the publisher’s online edition of *IIE Transactions* for the following supplemental resource: Appendix with additional tables of results.]

Keywords: Flexibility, workforce management, optimization, scheduling, cross-training

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. We consider a high-volume factory that could be a manufacturing facility or a “service factory” that

processes mail, checks, insurance claims, online product orders etc. The key operational feature of our environment is that finished products cannot be inventoried. While a common feature of services, it also applies to make-to-order production systems.

As described in Abernathy *et al.* (1973), workforce planning typically involves three hierarchical phases: planning, scheduling and allocating. Goodale and Thompson (2004) have a slightly different formulation: forecasting demand, determining requirements, scheduling employees and real-time adjustment. Given a demand forecast the approaches are essentially the same. 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

*Corresponding author

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, we assume the manager can reschedule and reallocate them to adapt to changes in demand patterns. We model the following forms of flexibility: cross-training, intra-shift job switching, part-time workers, start time flexibility and work-in-process buffers. The ability to reschedule and reallocate the workers within the constraints of the above forms of flexibility is in and of itself a form of flexibility. However, the setting we have in mind is one in which the demand pattern changes are not occurring on a daily basis but rather last for at least weeks so that schedules are predictable to workers. Given this setting we think it is appropriate to ignore overtime which typically is not repeated day after day (see Easton and Rossin (1997) and references therein for examples of research on overtime). One form of flexibility we do not consider has to do with day of the week scheduling. When there is variability in work load across days of the week it can be challenging to create work schedules across days. Worker flexibility regarding days off is of relevance here. While well studied in the literature (see for example Loucks and Jacobs (1991)) it is beyond the scope of this paper.

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 def-

inition 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 cross-training 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 and large workforce. 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 cross-trained 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 is summarized in detail in Section 3. Our goal in this paper is more comprehensive in that we embed their model within the framework of a two-stage staffing problem with uncertain demand in the second stage. In Berman *et al.* (1997) flexibility is utilized to best match labor to a known work arrival pattern. In the current paper we attempt to identify the forms of flexibility that will enable a manager to best adapt to unexpected changes in the work arrival pattern.

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 cross-training called downgrading, in which higher-skilled workers can do lower-skilled jobs as needed. Campbell (1999) and Brusco (2008) consider cross-training 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 cross-training. Brusco and Johns (1998) consider the planning phase when analyzing alternative cross-training policies. They similarly omit workflow dependencies and other forms of flexibility.

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 is 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 cross-training configurations for a set of dynamic allocation policies for cross-trained workers. They ask the following questions: “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 cross-training 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 cross-training 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. The first stage is solved independently of the second stage as opposed to solving the first stage to minimize costs over all expected second-stage realizations. Our motivation for that is that we are trying to evaluate flexibility as a response to shocks that are by definition unpredictable. Managers often operate in this way. They lack distributional data on potential demand shocks and therefore plan based on a forecast and then rely on flexibility to enable them to adjust.

3.1. Summary of Berman *et al.* (1997)

Here we summarize the model formulated in Berman *et al.* (1997). 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 \mathcal{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 left-over in the system at the end of period T is carried over into period 1. This means that we are conducting an equilibrium

Table 1. Table of notations

Symbol	Variable
$X_{k(j_1, j_2)h\tau}$	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 $j, j = 1, \dots, n$
\mathcal{H}	the set of all allowed shift lengths
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 his/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	$1, \dots, K, (j_1, j_2) \in A_k, h \in \mathcal{H}, \tau \in ST$
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

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} is the number of jobs in the buffer at station j at the start of period t . Finally, the conservation of buffers results in

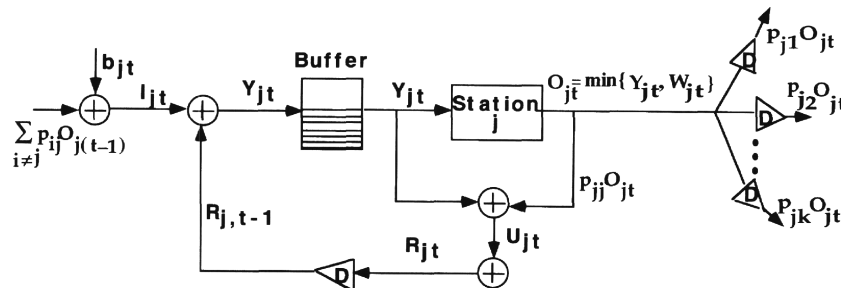
$$Y_{jt} = I_{jt} + R_{j(t-1)}, \quad (3)$$

where the right-hand side is the sum of the new jobs arriving at period t and residual jobs from period $t-1$. The workflow at a single workstation is depicted in Fig. 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

$$W_{jt} = \sum_{k \in M_j} \sum_{h \in \mathcal{H}} \left\{ \sum_{\tau \in F_{jh}} \beta_{kj} X_{k(j, j)h\tau} + \sum_{n_2 \in W_k} \sum_{\tau \in G_{jh}} \beta_{kj} X_{k(j, n_2)h\tau} + \sum_{n_1 \in W_k} \sum_{\tau \in Q_{jh}} \beta_{kj} X_{k(n_1, j)h\tau} \right\}, \quad (4)$$

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,

**Fig. 1.** Schematic of workflow at a single workstation.

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 . Lunch breaks are modeled as non-work periods that occur at the approximate mid-point of a shift. They are implemented via the sets F_{th} , G_{th} and Q_{th} . The actual productivity of station j at time t is not W_{jt} but O_{jt} , which satisfies:

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

and

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

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 multiplied by the worker capacity over all starting times. Having a higher number of starting times allows the system to take advantage of the overlapped workforce, e.g., a worker starting at 8 am overlaps a worker starting at 12 pm given that both work an 8-hour shift. When the load peak in this station is between 12 pm and 4 pm it is more efficiently covered than if the shifts only started at 8 am and 4 pm. 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:

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

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

and

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

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 $\sum_j \sum_t c R_{jt}$, we obtain the objective in the first stage:

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

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 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 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 workflow 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 1 we introduce Equation (11) which sets the number of workers for each type equal to the same number as in the first stage:

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

Table 2. Objective and constraints in both stages

Objective	Staffing	Rescheduling
	Minimizing staffing cost and WIP cost	Minimizing WIP cost
Workflow conservation	Yes (Equations (1)–(5b))	Yes (Equations (1)–(5b))
Buffer size	Yes (Equation (6))	No
Part-time ratio	Yes (Equation (7))	Yes (Equation (7))
Job switch ratio	Yes (Equation (8))	Yes (Equation (8))
Number of workers	No	Yes (Equation (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 system's ability to react to such shocks. In the next section we conduct extensive numerical experiments, to evaluate different flexibility policies in this way. It is important to note that we have made assumptions about the type of demand variability and its timing relative to the managerial staffing responses. Namely we are assuming that the demand pattern is known enough in advance that workers can be rescheduled. One can envision a scenario in which the Stage 1 shift schedule was fixed and after the demand pattern was observed the only recourse for the manager was to reassign workers to tasks during their scheduled work time. In such a scenario cross-training and job switching would have greater significance. Later we discuss what our results suggest would happen in such a situation.

4. Computational examples and results

In practice the staffing decision variables X and \tilde{X} in both the first and second-stage problems must take integer values because workers must be scheduled in whole quantities. As such the problems should be solved as Mixed-Integer Programs (MIPs). Our goal here is to investigate the benefits of flexibility and not to develop a new scheduling algorithm. For large facilities, i.e., those with many workers, the difference between the integer and continuous solutions will be relatively small. That said, one of the points we are trying to make in this paper is that slack capacity can make it easier for firms to adjust to changes in the work arrival pattern. Integer staffing provides some slack. To confirm that this effect is not significant for large facilities we solve the Stage 1 and Stage 2 problems for integer values in our experiments. To reduce the computational burden of our experiments we use a heuristic solution methodology rather

than solving the MIP to optimality. We solve the line programming relaxation and then use rounding to convert it to a feasible integer solution in the staffing variables. We find that across all the experiments the average deviation of the heuristic integer objective value from the continuous solution is less than 1% while the maximal deviations is 4%. These statistics indicate that we can be confident that the behaviors we observe in the following results will also be observed if the problems were solved using more exact methods.

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 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 \$15/hour and if they are qualified to work at a work station their productivity is 10 units per hour.

We study the following set of flexibility types:

- number of starting times (three, four, six or eight);
- part-time allowed (Yes or No);
- mid-shift jobs switching allowed (Yes or No);
- cross-training 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 one. For example, the starting times for eight 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 job switching 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 part-time workers and/or mid-shift job switching. 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: \$0.5, \$1.0 or \$2.0 per unit per time period.

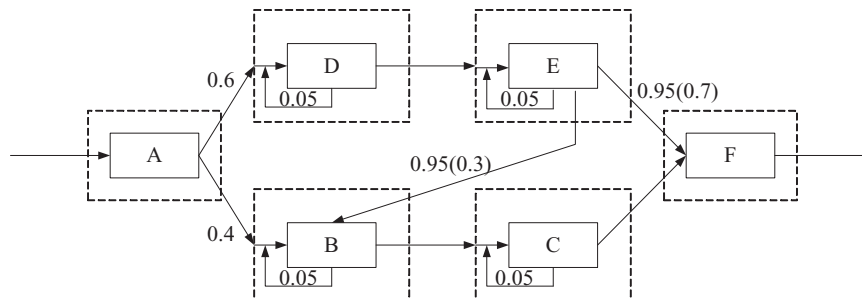


Fig. 2. Six-station process with network configuration.

The network configuration has six stations and is shown in Fig. 2. 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 four stations, where all jobs have to be handled by all four stations in the same order (Fig. 3). 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 Fig. 3) is similar to the definition in Jordan and Graves (1995, p. 580), 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.” Finally, “All” means that all workers are capable of performing the tasks at all four stations.

The third, parallel configuration is similar to the one in Jordan and Graves (1995), in which six parallel factories are considered (Fig. 4). We apply the same exogenous inputs

to all six factories. Three training levels are considered: “None,” “Chained pairs” (Fig. 4) and “All,” where workers are capable of performing tasks at 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 job switching policies into one index, and assign another index to the number of starting times (Table 4). Next, by summing these two numbers, we 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 considered (40), and part-time is allowed but mid-shift job switching 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)

Table 3. Skill mapping for different cross-training levels in the network configuration

Worker type	Cross-training 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

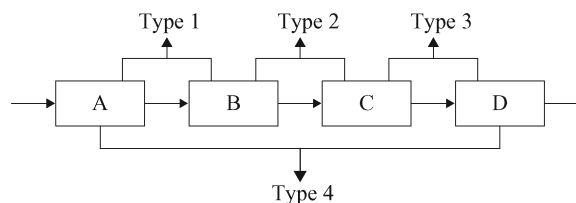


Fig. 3. Serial configuration showing paired chain cross-training.

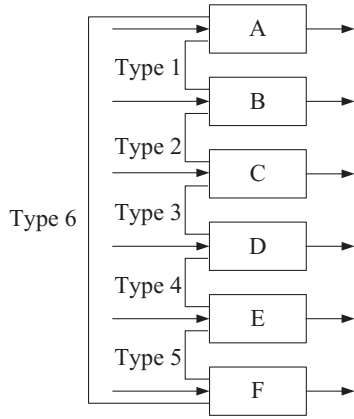


Fig. 4. Parallel configuration showing paired chain cross-training.

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 Fig. 5(a). In other words, the total number of arrivals and the maximum number of arrivals in any one period are the same across the 20 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 9600 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, 20 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 zero 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 $i, j \in$

Table 4. Indices for flexibility scenarios

Part-time	Mid-shift job switching	Index	Number of starting times	Index
No	No	2	3	30
No	Yes	4	4	40
Yes	No	6	6	50
Yes	Yes	8	8	60

{1, 2, ..., 48}. By adding a noise value to the demand at t_i and subtracting 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 Fig. 5. Figure 5(a) represents the demand with the peak at $t_p = 25$, Fig. 5(b) shows the demand where the maximum number of arrivals changes from 200 to 305 and Fig. 5(c) represents the first stage demand with noise. For each type of demand shocks we calculate the mean and 90th-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 configuration under demand shock t_p . We report more results in the online Appendix.

4.4.1. First stage

In Fig. 6 we plot the total Stage 1 cost for the network configuration. In Figs. 7(a) and 7(b) 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 cross-training 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,

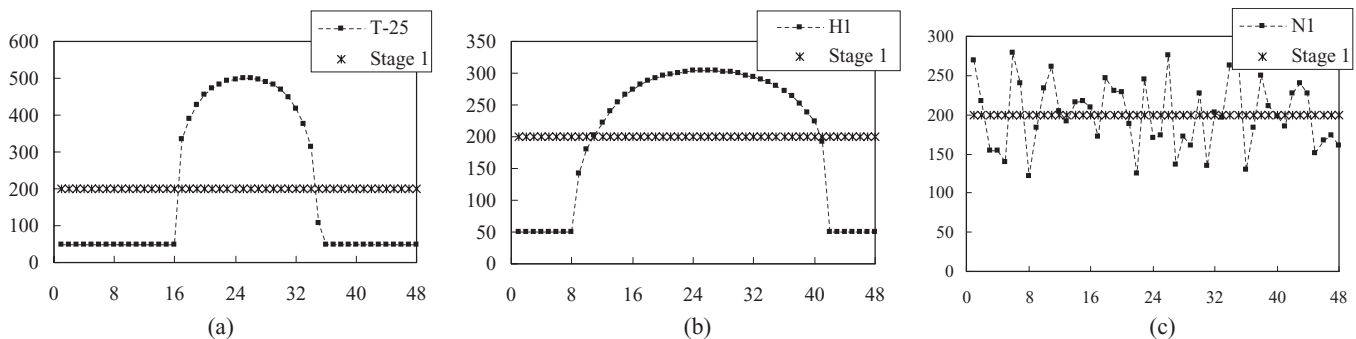


Fig. 5. Second-stage demand arrival pattern for: (a) shock t_p ; (b) shock H; and (c) shock N.

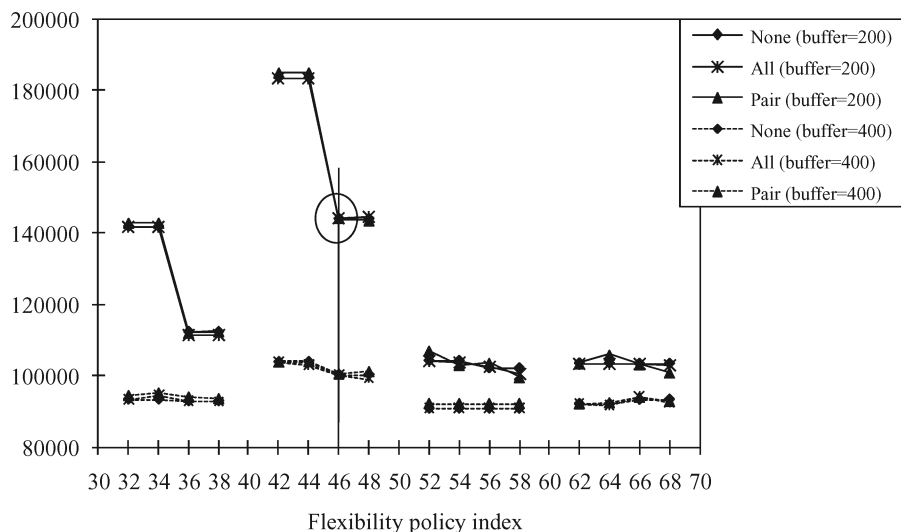


Fig. 6. Total Stage 1 cost for network configuration (WIP cost of \$0.5/unit/time).

the part-time ability and the mid-shift job switching ability (Table 4). For example, the three coinciding circled points, in Fig. 6, 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 cross-training levels of “None” (diamond marker), “All” (star marker) and “Pair” (triangle marker).

First, we observe that there is essentially no benefit from cross-training evidenced in the results because the work arrival profile is deterministic and uniform. Cross-training and job switching are advantageous when the workload at different stations is out of phase, meaning that one station has a heavy workload when the other has a light load. In such a case, a job switching cross-trained worker can spend part of his/her shift at one station and then move where the work is at the other station. In our Stage 1 problem demand arrivals are uniform so the stations are not out of phase and

there is little benefit from job switching. Also, note that in the Stage 1 problem, cross-training in and of itself, with no job switching ability, will not have any value. In the Stage 2 problem the various demand shocks create opportunities to take advantage of both cross-training and job switching. In fact, when we study the Stage 2 problem we find that cross-training on its own can have benefits without any job switching.

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 (Fig. 6). Figure 7(a) shows that this reduction is coming from the savings on staffing, especially when the

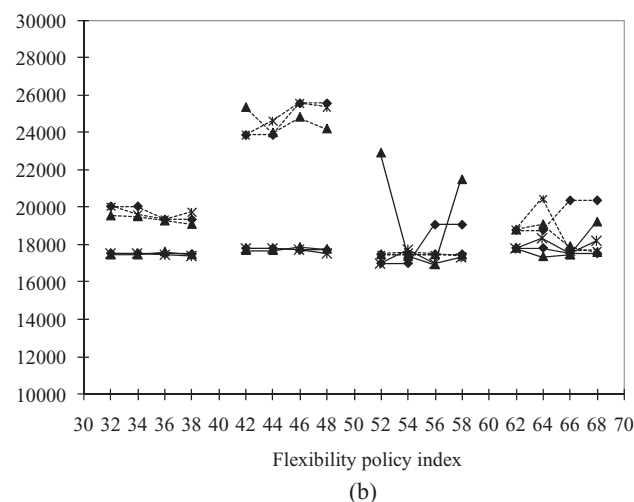
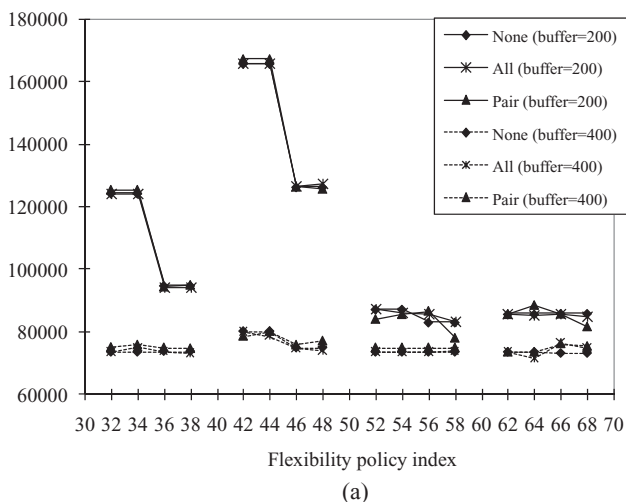


Fig. 7. (a) Staffing cost and (b) WIP cost for the network configuration in the first stage (WIP cost of \$0.5/unit/time).

starting times are restricted (30s and 40s). WIP costs may actually increase somewhat due to buffering (Fig. 7(b)).

Third, flexibility in the number of shift start times can also reduce cost significantly. By comparing the 30s and 50s or 40s and 60s (Fig. 6 and Fig. 7(a)), 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 30s). 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 Fig. 6 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 times 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 times 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.

4.4.2. Second stage

In Fig. 8, 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 cross-training can extract almost all the benefits of cross-training if the skills are “chained.” Jordan and Graves (1995) introduced the concept of pairwise skill chaining for a parallel system. In the context of a network, a skill chain is less well defined. Nonetheless, the pairwise cross-training configuration we implement is very effective. For example, Figure 8 shows 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 cross-training or no cross-training 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 Fig. 8(b), with eight starting times, the system with both job switching and part-time (68) has a higher total cost than the system without job switching (66). Figure 9 separates the total cost into the staffing (Fig. 9(a)) and WIP costs (Fig. 9(b)). In Fig. 9(a), the staffing cost for 68 is the lower than the staffing cost for 66, but it has a higher 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

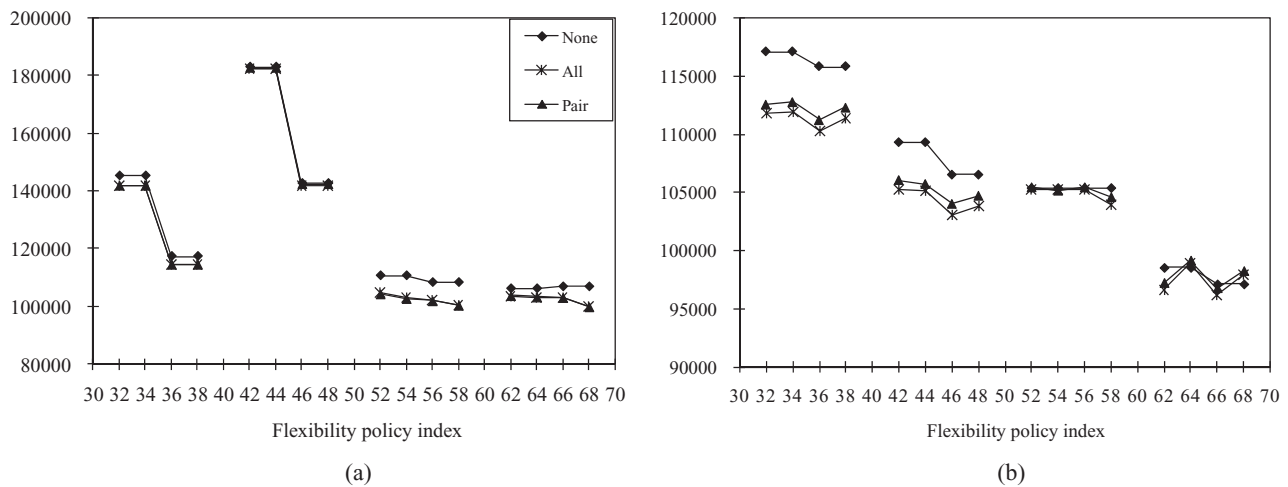


Fig. 8. Total Stage 2 cost for the network configuration under shock t_p with unit WIP cost of \$0.5/unit/time: (a) buffer size = 200; and (b) buffer size = 400.

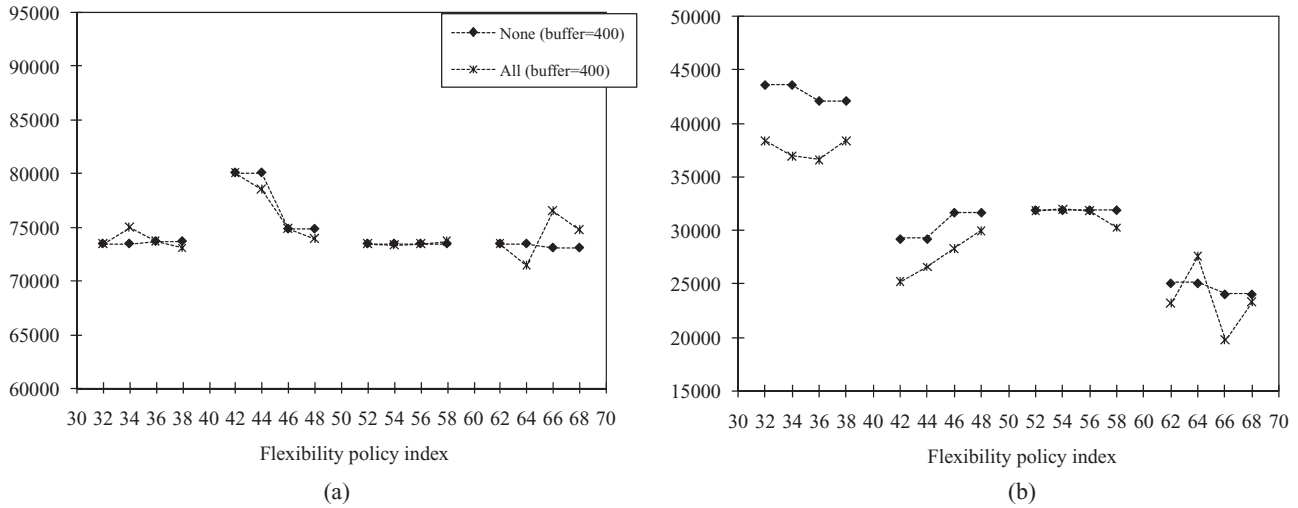


Fig. 9. (a) Staffing cost for the network configuration under shock t_p with a buffer size of 400 units and a unit WIP cost of \$0.5/unit/time; and (b) WIP cost for the network configuration under shock t_p with a buffer size of 400 units and a unit WIP cost of \$0.5/unit/time.

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.

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 versus six and four versus eight start times¹.

For each configuration and demand perturbation type we calculate $(E^+ - E^-)/2$ for each factor and 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 flexibility is included.

As an example suppose that our goal is to observe the effect of doubling the number of starting times from three to six 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 three (E^-) and the average when the number of starting times is six (E^+). Note that when the number of starting times is three (or six), we have $2 \times 2 \times 2 = 8$ policies, in which the 2s stand for the cross-training level ("None" or "All"), job switching ability (allowed or not) and part-time ability (allowed or not). Then, the effect of doubling the number of starting times from three to six is $(E^+ - E^-)/2$.

First we conduct four distinct DOE analyses identified by whether the start time levels are three versus six or four versus 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 job switching influences the results only when the cross-training level is "All," we consider only the cross-effect of training level and job switching in the DOE analyses instead of the main effect of job switching.² Therefore, the job switching

¹Depending on the demand pattern, having four starting times is not necessarily more effective than having three starting times whereas doubling the number of start times with a similar distribution within the day is increasing flexibility.

²Although other cross-effects, such as the number of starting times 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.

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

		<i>ST</i> (3, 6)					<i>ST</i> (4, 8)			
		<i>Unit WIP cost</i>	0.5	1.0	2.0	<i>Unit WIP cost</i>	0.5	1.0	2.0	
<i>B</i> = 200	<i>TL</i>		-2619	-4478	-8500	<i>B</i> = 200	<i>TL</i>	-1250	-1575	-2644
	<i>ST</i>		-11 818	-11 628	-10 913		<i>ST</i>	-28 952	-28 594	-29 056
	<i>PT</i>		-7496	-6521	-6725		<i>PT</i>	-10 233	-10 000	-9888
	<i>JS</i>		-2039	-3118	-5692		<i>JS</i>	-1143	-1175	-1846
<i>B</i> = 400	<i>TL</i>		-1388	-3081	-6127	<i>B</i> = 400	<i>TL</i>	-985	-2081	-1498
	<i>ST</i>		-4371	-7269	-12579		<i>ST</i>	-4238	-7046	-9001
	<i>PT</i>		-372	-2006	-2654		<i>PT</i>	-822	-669	-929
	<i>JS</i>		-935	-2413	-3789		<i>JS</i>	-264	-859	-416
Overall <i>B</i> effect			-4170	2283	12 217	Overall <i>B</i> effect		-15 771	-13 051	-11 285

factor is low if job switching is not allowed or if the training level is “None,” and high if both job switching 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., $(E^+ - E^-)/2$.

4.4.3.1. 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 cases 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., the effect of moving from three to six start times, increases in magnitude from -4371 to -7269 to -12579 in Table 5. When the unit WIP cost is high, the factory staffs more than when the unit WIP cost is low. 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. However, when buffers are small,

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 -28 952 to -4238 in (4, 8) 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 -4170 to +12 217 in (3, 6) in Table 5).

4.4.4. Flexibility robustness

In the previous section, we investigated the aggregated effect of each factor and saw the effectiveness of each factor. In this section, we discuss the robustness of each factor by checking for consistent reduction in the total Stage 2 costs when a factor’s level is changed from low to high. For example, if the goal is to investigate the robustness of allowing job switching for the network configuration when $B = 200$ and unit WIP cost of \$0.5 we change the job switching ability from No to Yes (the other factors remain unchanged), and then compare the total Stage 2 costs (before and after change). If the expected cost decreases for a high proportion of cases we view it as a robust cost saving mechanism. We perform this same comparison for 2 (training level) \times 2 (part-time) \times 4 (number of starting times) = 16 policies and three 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 94%, which means that 94% of times including job switching 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) Robustness test of allowing jobswitching							
	B=200			B=400			overall %
unit WIP cost	0.5	1.0	2.0	0.5	1.0	2.0	
% robustness for Network configuration	94%	98%	98%	77%	71%	75%	85%
% robustness for Serial configuration	100%	96%	96%	75%	71%	69%	84%
% robustness for Parallel configuration	100%	94%	94%	88%	81%	88%	91%
(b) Robustness test of allowing part-time workers							
	B=200			B=400			overall %
unit WIP cost	0.5	1.0	2.0	0.5	1.0	2.0	
% robustness for Network configuration	88%	88%	88%	94%	90%	88%	89%
% robustness for Serial configuration	96%	96%	96%	94%	94%	81%	93%
% robustness for Parallel configuration	100%	100%	92%	96%	81%	83%	92%
(c) Robustness test of having cross-training (change from "None" to "All")							
	B=200			B=400			overall %
unit WIP cost	0.5	1.0	2.0	0.5	1.0	2.0	
% robustness for Network configuration	100%	100%	100%	100%	98%	100%	100%
% robustness for Serial configuration	100%	100%	100%	96%	98%	100%	99%
% robustness for Parallel configuration	71%	83%	83%	92%	90%	92%	85%
(d) Robustness test of doubling the number of starting times							
	B=200			B=400			overall %
unit WIP cost	0.5	1.0	2.0	0.5	1.0	2.0	
% 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%

Fig. 10. (a) Robustness test of allowing job switching; (b) robustness test of allowing part-time workers; (c) robustness test of having cross-training (change from "None" to "All"); and (d) robustness test of doubling the number of starting times (three to six or four to eight).

a form of flexibility as robust³. In Fig. 10, 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 job switching (Fig. 10(a)) and part-time (Fig. 10(b)) flexibility are not particularly robust. When the buffer is small ($B = 200$), job switching is robust (though less so for the parallel configuration), while if $B = 400$, job switching is not robust. Part-time work is not very robust regardless of the buffer size.

Cross-training 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 cross-training level from "None" to "All," almost all the comparisons of the total Stage 2 costs indicate that having a higher cross-training level reduces the total Stage 2 costs (Fig. 10(c)). When $B = 400$ cross-training

is not always beneficial. The cases in which cross-training makes cost increase have one point in common: job switching is allowed in the first stage. Since the system is flexible enough (due to the relaxed buffer size), applying both cross-training and job switching makes the system rely too much on flexibility. However, if we exclude the job switching ability in the first stage, cross-training benefits in the second are very robust. We note that the system can improve the performance even more by excluding the job switching ability in the first stage but implementing it in the second stage.

The number of starting times is the one that is both effective (Table 5) and robust (Fig. 10(d)). Figure 10(d) shows that the number of starting times has highest overall robustness scores for the three system configurations (Fig. 10(d): 99, 99 and 98%). Moreover, most cases in which increasing the number of starting times is not effective allow part-time workers. When part-time work 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 a larger number of starting times because less workers

³We acknowledge this threshold is somewhat arbitrary but it forms a natural break for our results.

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%

Fig. 11. Robustness test of doubling the buffer size (from $B = 200$ to $B = 400$).

are available to handle demand shocks. 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 the 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 Fig. 11). When the unit WIP cost is low, doubling the buffer size is robust, but robustness decreases significantly when the unit WIP cost increases. Overall, buffer size, job switching and part-time flexibility are not robust, but cross-training without job switching and high number of starting times without part-time are robust combinations.

5. Conclusions

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, job switching ability, cross-training 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 cross-training, 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 cross-training was found to be generally robust we also found that its robustness was diminished when combined with job switching. 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 general rules 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 workers 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 cross-training but set staffing levels assuming job switching is not allowed.

It could be argued that rescheduling workers to different start times is not always possible. Our results suggest what would occur in such a case. Start time flexibility would no longer be effective because it would only apply in the first stage. In fact it would worsen performance because using it in Stage 1 would reduce the workers available in Stage 2 without any benefit of increased Stage 2 flexibility. The manager would be left less able to adjust to demand shocks. This would be similar to staffing in Stage 1 with large buffers and then operating in Stage 2 with small buffers. For this reason we did not consider this to be an interesting case to analyze in this paper.

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 cross-training 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

- Abernathy, W.J., Baloff, N., Hershey, J.C. and Wandel, S. (1973) A three-stage manpower planning and scheduling model – a service-sector example. *Operations Research*, **21**(3), 693–711.
- Bard, J.F. (2004) Staff scheduling in high volume service factories with downgrading. *IIE Transactions*, **36**(10), 985–997.
- Beach, R., Muhlemann, A.P., Price, D.H.R., Paterson, A. and Sharp, J.A. (2000) A review of manufacturing flexibility. *European Journal of Operational Research*, **122**, 41–57.
- Berman, O., Larson, R.C. and Pinker, E. (1997) Scheduling workforce and workflow in a high volume factory. *Management Science*, **43**(2), 158–178.
- Browne, J., Dubois, D., Rathmill, K., Sethi, S.P. and Stecke, K.E. (1984) Classification of flexible manufacturing systems. *The FMS Magazine*, **2**(2), 114–117.
- 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.
- Brusco, M.J. and Johns, T.R. (1998) Staffing a multiskilled workforce with varying levels of productivity: an analysis of crosstraining policies. *Decision Sciences*, **29**, 499–515.
- Campbell, G.M. (1999) Cross-utilization of workers whose capabilities differ. *Management Science*, **45**(5), 722–732.
- De Toni, A. and Tonchia, S. (1998) Manufacturing flexibility: a literature review. *International Journal of Production Research*, **36**(6), 1587–1627.
- Easton, F.F. and Rossin, D.F. (1997) Overtime schedules for full-time service workers. *Omega*, **25**, 285–299.

- Ernst, A.T., Jiang, H., Krishnamoorthy, M., Owens, B. and Sier, D. (2004) An annotated bibliography of personnel scheduling and rostering. *Annals of Operations Research*, **127**(1–4), 21–144.
- Goodale, J.C. and Thompson, G.M. (2004) A comparison of heuristics for assigning individual employees to labor tour schedules. *Annals of Operations Research*, **128**(1–4), 47–63.
- Graves, S.C. and Tomlin, B.T. (2003) Process flexibility in supply chains. *Management Science*, **49**(7), 907–919.
- Hopp, W.J., Tekin, E. and Van Oyen, M.P. (2004) Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science*, **50**(1), 83–98.
- Iravani, S.M., Van Oyen, M.P. and Sims, K.T. (2005) Structural flexibility: a new perspective on the design of manufacturing and service operations. *Management Science*, **51**(2), 151–166.
- Jordan, W.C. and Graves, S.C. (1995) Principles on the benefits of manufacturing process flexibility. *Management Science*, **41**(4), 577–594.
- Loucks, J. and Jacobs, F. (1991) Tour scheduling and task assignment of a heterogeneous workforce: a heuristic approach. *Decision Sciences*, **22**, 719–738.
- Montgomery, D. (2004) *Design and Analysis of Experiments*, Wiley, New York, NY.
- Sethi, A.K. and Sethi, P.S. (1990) Flexibility in manufacturing: a survey. *International Journal of Flexible Manufacturing Systems*, **2**(4), 289–328.
- Vokurka, R.J. and O’Leary-Kelly, S.W. (2000) A review of empirical research on manufacturing flexibility. *Journal of Operations Management*, **18**, 485–501.
- 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.

Biographies

Edieal J. Pinker is an Associate Professor of Computers and Information Systems at the Simon School of Business, University of Rochester. He conducts research on the use of contingent workforces, cross-training and experience-based learning in service sector environments as it applies to work and workflow design. He also studies the use of online auctions in electronic commerce and the issues faced by legacy firms trying to transition into electronic commerce. He has consulted for the United States Postal Service, the financial services industry and the auto industry. His work has been widely published in leading operations journals. He serves on the Editorial Boards of *Management Science*, *Operations Research*, *POMS*, *Decision Sciences* and *IJOR*. He earned his M.S. and Ph.D. in Operations Research from the Massachusetts Institute of Technology.

Hsiao-Hui Lee is a doctoral candidate in Operations Management at the Simon School of Business, University of Rochester. She holds B.S. and M.S. degrees in Civil Engineering from the National Taiwan University.

Oded Berman is the Sydney C. Cooper Chair in Business and Technology, and Professor of Operations Management at the Rotman School of Management at the University of Toronto. He conducts research on logistics, operations management in the service industry, workforce management and software reliability. His work has been widely published in leading operations journals. He is on the Editorial Boards of *Computers & Operations Research*, *Management Science* and *Transportation Science*. He earned his M.S. and Ph.D. in Operations Research from the Massachusetts Institute of Technology.