International Journal of Production Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tprs20

Requirement estimation for indirect workforce allocation in semiconductor manufacturing

Chen-Fu Chien [a] , Wen-Chih Chen [b] & Shao-Chung Hsu [a]

[a] Department of Industrial Engineering and Engineering Management , National Tsing Hua University , Taiwan

[b] Department of Industrial Engineering and Management , National Chiao Tung University , Taiwan

Published online: 15 Jan 2010.

To cite this article: Chen-Fu Chien , Wen-Chih Chen & Shao-Chung Hsu (2010) Requirement estimation for indirect workforce allocation in semiconductor manufacturing, International Journal of Production Research, 48:23, 6959-6976, DOI: 10.1080/00207540903352678

To link to this article: http://dx.doi.org/10.1080/00207540903352678

International Journal of Production Research
Vol. 48, No. 23, 1 December 2010, 6959–6976

Requirement estimation for indirect workforce allocation in semiconductor manufacturing

Chen-Fu Chien[a], Wen-Chih Chen[b]* and Shao-Chung Hsu[a]

[a]*Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan;* [b]*Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan*

(*Received 21 January 2009; final version received 21 September 2009*)

This work, based on a real case, presents a model to estimate indirect workforce requirements of semiconductor fabrication facilities (fabs) so that the workforce can be fairly allocated. There is a concern in a real setting to fairly allocate the overall corporate workforce among the fabs, particularly when they compete in performance, and properly determining the actual requirement is the most critical in the decision process. The actual requirements of the workforce, especially the indirect workforce, for fabs may be indeterminate due to the lack of a well-defined workforce-output relationship. This paper presents a non-parametric frontier approach for estimating the indirect workforce, and the estimate is based on the best past performance adjusted to reflect the expected productivity growth. An empirical study was conducted in a leading foundry in Taiwan that has a number of 8-inch fabs. The proposed (re)allocation approach can provide an explicit decision support mechanism to balance the workloads in light of various production environments to enable an equitable basis for performance evaluation to foster constructive competition among the fabs.

Keywords: decision analysis; frontier models; semiconductor manufacturing; human capital; manufacturing strategy

1. Introduction

Following Moore's Law, the semiconductor industry has achieved unparalleled growth in the last few decades via maintaining rapid technology nodes advances and competitive advantages in productivity. The semiconductor industry is knowledge and capital intensive. In addition to physical capital investment, human capital enhancement and human resource management have recently attracted increasing attention (Chien and Chen 2007).

This work is motivated by the needs of the semiconductor manufacturer in real settings in Taiwan due to numerous reasons. First, the demand for knowledge workers is growing, even considering the rising costs of labour and automation. Indeed, knowledge workers including engineers and technical staff play increasingly important roles in modern semiconductor companies that are operated with highly automated and intelligent manufacturing facilities. Second, historically the semiconductor industry used to attract and retain talent by offering generous stock dividends in lieu of high salaries.

*Corresponding author. Email: wenchih@faculty.nctu.edu.tw

ISSN 0020–7543 print/ISSN 1366–588X online

DOI: 10.1080/00207540903352678
http://www.informaworld.com

Beginning in 2008, however, new Taiwanese accounting rules that require expensing employee bonuses are affecting the bottom line at most high-tech companies. Third, where annual turnover typically ranges from 10% to 20%, deliberate workforce planning, maintenance and allocation by firms determines whether workforce productivity indeed becomes a key performance indicator (KPI).

Existing studies have investigated workforce decisions, including staff scheduling or rostering. Scheduling decisions determine work timetables in a manner that satisfies demand while optimising certain criteria. For example, Thompson and Goodale (2006) present a staff scheduling method for cases involving workers with different productivity levels. Staff scheduling problems are an extension of conventional scheduling problems and comprehensive reviews can be found in Aykin (2000), Burke *et al.* (2004), and Ernst *et al.* (2004). These decisions are typically operational and all detailed information is assumed to be obtainable.

Another line of research has focused on job assignment and reallocation, where assigning workers to perform certain jobs can be modified and modelled as a classic assignment problem (e.g., Holder 2005), or by other means such as simulation (e.g., Zulch *et al.* 2004). Additionally, long-term staffing optimises the workforce level for each category of staff hired during each period considering production ramping and technology migration. Some research on long-term staffing examines workforce planning optimisation under deterministic conditions (e.g., Mundschenk and Drexl 2007, Wirojanagud *et al.* 2007, Fowler *et al.* 2008). Other studies examine the stochastic nature of problems involving issues such as learning curve and turnover, which can be modelled as Markov decision processes and solved using various techniques (e.g., Gans and Zhou 2002, Ahn *et al.* 2005).

The applications of workforce decisions and planning have also been studied. Mundschenk and Drexl (2007) propose an integer programming model for long-run staffing in the printing industry, and Pesch and Tetzlaff (2005) study the interactions between staffing and scheduling decisions in the automotive industry. Bard *et al.* (2007) investigate workforce planning for US Postal Service mail processing and distribution centres. Finally, Holder (2005) studies the process of optimising job assignment to maximise satisfaction for US naval personnel.

There is a branch of the literature with a specific interest in semiconductor manufacturing industry human resource management. For example, Wirojanagud *et al.* (2007), and Fowler *et al.* (2008) propose a mixed integer programming model for workforce decisions considering worker differences. Chien and Chen (2007) investigate recruiting and retaining quality human capital. However, these optimisation-oriented methods lack sufficient application to today's knowledge-intensive high-tech industry.

This paper evolved from a practical need of a semiconductor firm in Taiwan to fairly allocate its indirect workforce among several fabs that have different configurations yet compete in performance within the company and then reallocate the workforce in light of changes of demands or tasks assigned. In contrast to the direct workforce that can be 'physically and conveniently associated with converting raw material into finished goods', the indirect workforce is identified because 'their efforts have no physical association with finished goods, or it is impractical to trace the costs to the goods produced' (Weygandt *et al.* 2005). For example, operators in a fab are the direct workforce while equipment engineers are classified as indirect workforce. Currently, on a quarterly basis the corporate planning unit and fab managers review previous performance, next-quarter task (output) assignments to fabs and overall workforce quota to make decisions about (re)allocation.

In this decision process, the real workforce requirement for each fab was unclear to the central corporate planning unit, and the fabs could request workforce levels in order to benefit their performance in the competition against other fabs. Consequently, the decision process was often controversial, and the workforce requirement estimation was crucial for solving the firm's allocation problem.

Based on problem diagnosis, a fair (re)allocation framework and the associated decision supporting mechanism for fair competition among fabs is more important than complex optimisation modelling. The proposed workforce estimation model is based on non-parametric frontier models that consider best experience performance, making adjustments to reflect anticipated productivity growth. We also suggest a (re)allocation decision process that can balance a firm's workload across different production sites by providing similar production environments and a fair and equitable basis for performance evaluation throughout a firm.

The remainder of this paper is organised as follows. Section 2 analyses and structures the problem rooted in a real world-case, and defines the problem scope. Sections 3 and 4 then describe how to estimate workforce requirements once the need to consider productivity improvement is considered. Section 5 presents a workforce reallocation that balances personnel shortages or excesses among different production units. Section 6 returns to the real-world example to demonstrate the proposed method, followed by conclusions.

2. A decision framework

This approach proposes a framework as illustrated in Figure 1 that describes the complex relationship of the decision issues and the information relevant to making the workforce allocation and reallocation decisions in a real setting. In particular, major decisions include:

(1) **Workforce planning:** the decision to set an appropriate corporate workforce level across certain time periods, such as two or three years, to meet future demand. In the high tech industry, the only constant is change. Long-term product demand generally exhibits significant variation, seasonality and uncertainty.

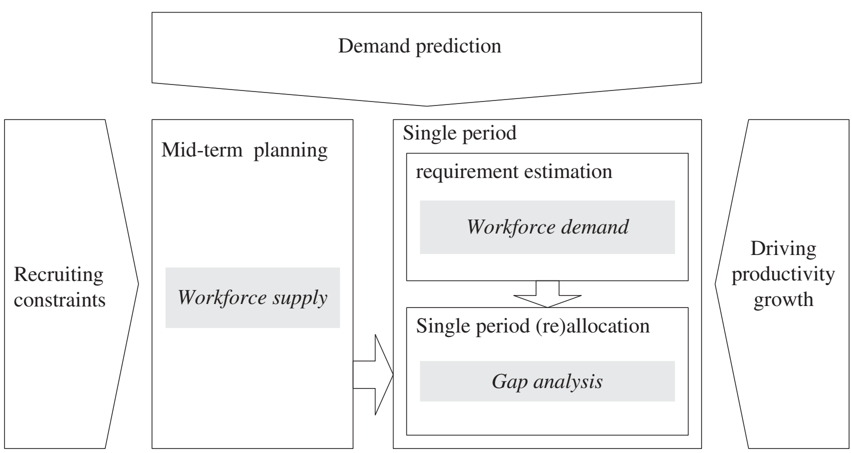

Figure 1. A framework for workforce decisions.

Therefore workforce planning determines future labour supply subject to some external influences affecting human resource availability, i.e., recruiting constraints, turnover rate, seasonality, and corporate strategic policy.

(2) **Workforce requirement estimation:** the decision to estimate 'real' workforce demand in response to the actual task assignments for individual production sites. Different from workforce planning, typically this decision is decomposed to the production site level and is site-dependent to reflect inter-site differences in production conditions. The estimation corresponds to a single future period during which all tasks are assigned based on a firm's overall production plan. A good understanding of 'real' workforce requirements[1] to fulfil the expected tasks provides the foundation for optimal decision-making.

(3) **Workforce allocation:** the decision to solve for inconsistency, especially a workforce shortage, between currently available supply and demand. Given the discrepancy between workforce requirements and available supply, this decision tries to resolve the gap by reallocating the workforce to each fab. The decision can be viewed as a minor tuning based on the up-to-date workforce status and certain operational management objectives.

(4) **Driving productivity growth:** not a decision per se, but a process akin to setting targets or constraints. It is an attempt to drive the improvement of workforce-based performance to boost a firm's sustainability in competitive business environments. It may be subjective, and may have a top-down orientation.

Considering the process aspect of decision-making in the case studied, the first priority is to determine workforce supply, and follows estimating the labour demand, identifying the discrepancies between supply and demand, and developing a response. The supply side of the workforce is a mid-term decision based on long-term aggregate forecasting. Recruiting and training require a longer lead time. High uncertainty exists regarding long-term forecasting due to the nature of the industry's competitive market and rapid technology growth. Therefore, the workforce level functions to smooth fluctuating market demand to maximise a firm's long-term profits. In addition to profit-maximising, other objectives, such as the robustness of plans to manage for environmental uncertainty, are also considered in reality. The demand side of the workforce is a decision based on the assigned tasks which are known and obtained by arriving and 'in-hand' customer orders. Its timing occurs just before the execution and after the monthly or quarterly production plan, i.e., to discover a minimal requirement of the workforce to meet the production plan. Since there is always a gap between current workforce demand and the supply of suitable personnel due to business environment uncertainty, the first step is its identification. This paper does not consider certain solutions, such as production outsourcing, but focuses on reallocating the workforce to minimise difficulties in meeting production plans and to balance these difficulties across fabs. Subjective concerns related to productivity enhancement, generally originating from the executive level of the firm, are considered in addition to 'objective' requirement estimation and resource allocation.

Notably most conventional workforce planning studies formulated the problem as a workforce supply optimisation problem, in which precise demand information was provided, and the workforce supply-demand gap was handled as constraints together with other considerations. As illustrated in Figure 1, the existing approaches optimised workforce supply decision with detailed labour demand information properly and explicitly provided. For example, Wirojanagud *et al.* (2007), and Fowler *et al.* (2008)

optimise the workforce plan considering different skill levels. Their models assume that all task contents are properly transformed into labour working hours and taking labour productivity into account for different time periods and skill levels. However, in many practical cases including ours, the labour-task relationship is difficult to express and demand estimation is thus quite often challenging.

Indeed, for the executive level responsible for different business units (e.g., the fabs), workforce decision-making is not an optimisation problem but a political problem due to the absence of precise information. As mentioned the workforce-task relationship may be unclear and/or hard to quantify, especially when the production process is complex and there are various products/services, and the true labour demand is unclear to the decision makers. Most important, business units within a firm may be in competition and workforce allocations may affect performance. Under these circumstances, as we experienced in the fab industry problem, workforce requested units may request the level that will favour their performance rather than what is truly needed.[2] Consequently, the performance competition and asymmetric information make workforce (re)allocation, among all resource allocation problems, almost always controversial. The controversy arises over the requirement estimation rather than the allocation rule itself. In fact, estimating a firm's workforce requirement properly is perhaps the most critical factor in dispelling controversy.

Focusing on real needs, this approach develops a workforce estimation model in response to the challenges addressed, although the ultimate goal is to support the allocation decisions. The proposed methodology is especially suitable for cases with asymmetric information between planners and multiple agents, and/or when the labour-task relationship is hard to quantify, e.g., project-oriented or knowledge-intensive cases. Nevertheless, optimisation approaches (e.g., Wirojanagud *et al.* 2007, Fowler *et al.* 2008) can be employed when detailed labour demand information can be provided, e.g., direct labour for standardised mass production or with well-defined labour standards.

3. Requirement estimation

Early workforce allocation literature relies on precise labour demand information. The information is typically provided by means such as the bottom-up approaches based on labour standards. Such approaches require a detailed understanding of the labour-task relationship involved in simple and stable processes. However, complex production processes, in which various types of labour work together to provide multiple tasks (products/services) do not satisfy these underlying requirements, i.e., when the productivity rates with trade-offs exist. These challenges are particularly significant at the corporate level. Furthermore, the labour standard approaches adopted in practice typically use the average number from the past as a required workforce basis, and inefficient production units frequently require a larger workforce than efficient ones, hence the need to allocate more resources. This observation has yielded a number of criticisms, and the conventional approach cannot provide any incentive for improvements.

Instead of bottom-up approaches, this paper utilises non-parametric frontier models to determine minimal requirements based on past best experience and from an aggregate macro viewpoint at the corporate level. The non-parametric frontier approach can also handle multiple workforce types and tasks without *a priori* weights. The following sections introduce notions of non-parametric frontier models, followed by the workforce estimation model.

3.1 *Frontier model*

In conventional engineering disciplines, a function $y = f(x)$ is commonly used to represent the *ideal* relationship between an independent variable x and a dependent variable y. Typically, x denotes the resource (input) consumed while y represents the output generated by x, e.g., power $= f$(fuel). Considering that resource wastage occurs in real world situations, a resource-output bundle (x', y') in the area below $f(x')$, $f(x') \geq y'$, is feasible because x' can ideally produce $f(x')$, which is no less than y'. Namely, $f(x) \geq y$ defines the feasible region while $y = f(x)$ represents its boundary (or frontier).

A firm can utilise $f(x)$ to make decisions depending on different objectives or interests. For example, if a firm needs to determine its minimum resource usage to meet the required output y°, the problem can be formulated as:

$$\begin{aligned} &\min_x x \\ \text{subject to}\quad &f(x) \geq y^\circ. \end{aligned}$$

Moreover, if the unit cost and unit price for resource and output are available as p and c respectively, a firm can maximise its profits by solving the following problem:

$$\begin{aligned} &\min_{x,y}\ py - cx \\ \text{subject to}\quad &f(x) \geq y. \end{aligned}$$

Extending the above idea to multi-resource single-output cases, $f(\mathbf{x}) = y$ can be defined as the ideal resource-output function, in which $\mathbf{x} \in \Re_+^{|I|}$ is a value vector for resource set I. The feasible region for resource-output bundle $(\mathbf{x}', y')$ should satisfy $f(\mathbf{x}') \geq y'$. For example, consider a case with one output y and two resources, denoted as $\mathbf{x} = (x_1, x_2)$ (Figure 2). The x-axis denotes the value for the resource x_1 while the y-axis represents the value for another resource x_2. Figure 2 in particular represents the case of expected output being y°. Now suppose the ideal relationship between $\mathbf{x}$ and y° is given as $f(x_1, x_2) = y^\circ$ (Figure 2). The shadow area denoted as $L(y^\circ)$ is the region of $\mathbf{x}$'s such that $f(\mathbf{x}) \geq y^\circ$ because more resources are used and the corresponding ideal output will be no less. Namely, $L(y^\circ)$ is the collection of $\mathbf{x}$'s that can produce required output y°, and is

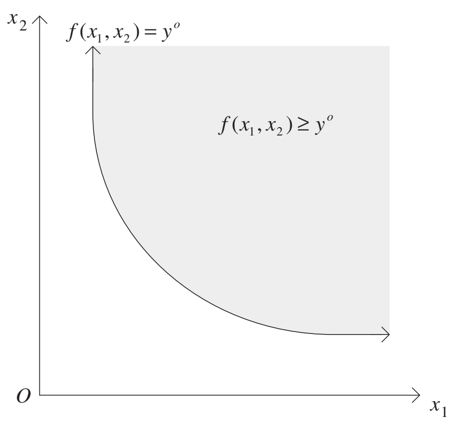

Figure 2. The frontier.

defined as $L(y^\circ) \equiv \{\mathbf{x}|f(\mathbf{x}) \geq y^\circ\}$. Similarly, $f(\mathbf{x})$ and $L(y^o)$ can help make decisions to achieve different, possibly competing objectives. For example, if the unit cost vector $\mathbf{c} \in \Re_+^{|I|}$ is given for all inputs, a firm can allocate resources to minimise its total cost as:

$$\min_{\mathbf{x}}\left\{\mathbf{c}^{\mathrm{T}}\mathbf{x}|f(\mathbf{x}) \geq y^\circ\right\} = \min_{\mathbf{x}}\left\{\mathbf{c}^{\mathrm{T}}\mathbf{x}|\mathbf{x} \in L(y^\circ)\right\}.$$

To generalise to the multi-resource and multi-output cases, we define the ideal function as $f: \Re_+^{|I|} \rightarrow \Re_+^{|J|}$ where J denotes the collection of different outputs. Denote $\mathbf{y}^o \in \Re_+^{|J|}$ as the value vector of output set J, and it is the outputs that need to be met. The corresponding feasible resource bundles are thus defined as:

$$L(\mathbf{y}^\circ) \equiv \left\{\mathbf{x}|f(\mathbf{x}) \geq \mathbf{y}^\circ\right\}.$$

In most engineering practices, function $f(x)$ is estimated by various regression techniques based on the collected data. However, conventional regression techniques estimate the central tendency of the data rather than its real 'ideal' performance, and the estimated function does not truly represent the ideal resource-output relationship. The estimation performed in this paper applies the frontier techniques based on the philosophy that a bundle is always achievable as long as it has previously been achieved. The estimated frontier function represents the real ideal condition. Function $f(x)$ is estimated using a piecewise linear function comprising the best possible ever in a given data set. Figure 3 shows an example of the estimation where the dots represent given historical records for a case with two resources and one output.

To generalise and implement the idea mentioned above, three assumptions were examined: (i) the engineering interpolation is adopted (if two bundles are achieved, any convex combination of them is achievable); (ii) if a resource-output bundle is feasible, it is also feasible to increase resources used or to reduce output produced; and (iii) the constant returns to scale (CRS) approach is employed (any observed output-resource ratio will hold constant for different sizes of outputs and resources; this property is commonly adopted in productivity indices). For example, suppose $(x, y) = (2, 6)$, i.e., $y/x = 3$, then typically x should be 4 when $y = 12$. Denoting S as the set of historical records, then $(\mathbf{x}^r, \mathbf{y}^r) \in \Re_+^{|I|+|J|}$

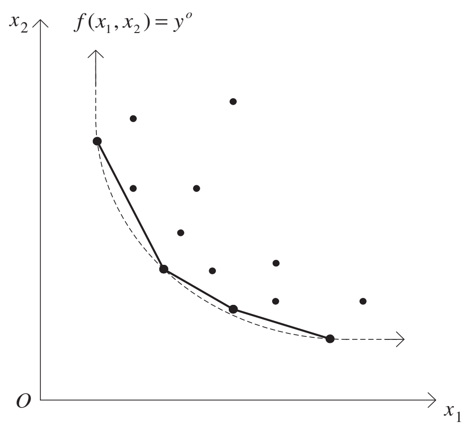

Figure 3. The estimated frontier.

for $r \in S$ are the value vectors of resources and outputs. For set $L(\mathbf{y}^{\circ})$, the $\mathbf{x}$ for which it is feasible to obtain $\mathbf{y}^{\circ}$ thus can be estimated according to S and our three assumptions as $\hat{L}(\mathbf{y}^{\circ})$:

$$\hat{L}(\mathbf{y}^{\circ}) \equiv \left\{ \mathbf{x} \middle| \sum_{r \in S} \mathbf{x}^r \lambda_r \leq \mathbf{x}; \sum_{r \in S} \mathbf{y}^r \lambda_r \geq \mathbf{y}^{\circ}; \lambda_r \geq 0, r \in S \right\}. \tag{1}$$

According to some popular objectives, such as cost minimisation and revenue maximisation, optimal solutions occur on the frontier, because it represents the ideal relationship. Typically, most engineering practices seek optimal solutions in relation to certain objectives regarding ideal functions.

3.2 *Estimating total indirect workforce requirement*

Considering a real setting, this section presents a method to estimate quarterly indirect workforce requirements for 8-inch fabs. Estimating indirect workforce needs is more difficult than direct workforce needs because work content is more complex, irregular and not routine. As suggested above, the proposed frontier model is more appropriate than traditional labour standard workforce demand estimation.

We consider three resources as three different indirect labour categories and functions (process engineer, process integration engineer, and equipment engineer). From a central resource allocation viewpoint, it is necessary to determine the minimum total workforce requirements, but not the detailed quantities in each category. In practice, the total amount of workforce is preferred because indirect labour is sufficiently flexible (e.g., cross-trained) to respond to different types of tasks through training, and fab managers can gain reallocation flexibility.

Total wafer output volume, customer service loading, and total technological difficulty are considered outputs generated by the indirect workforce. Detailed definitions are as follows:

- Process engineers (*PE*) and equipment engineers (*EE*): oversee routine maintenance and troubleshoot. Their numbers are calculated by averaging headcount over the planning period. Here, the planning horizon is three months (quarter).
- Process integration engineers (*PIE*): responsible for product yield and customer service, such as answering questions related to process recipe and experiments. Their numbers are also measured as the average headcount for each quarter.
- Total outputs (*Q*): the quarterly equivalent of 8-inch wafer outputs, the key product of the fab.
- Customer service loading (*SL*): an output that directly uses equivalent customer numbers. Equivalent customer numbers are the sum of all customers with weights assigned according to the orders and values. A fab provides customised products based on customer needs. The effort required for customer service and set-up increases with customer numbers.
- Technology difficulty (*TD*): an output that represents the extra loading due to technology complexity. In practice, technical difficulty weight (*TDW*) is used as a relative measure to adjust the complexity change compared to the previous year. For example, $TDW_{\mathrm{Y99}}^{\mathrm{Y00}} = 1.2$ indicates that the technical complexity in 2000 is 1.2 times that in 1999. Given year to year *TDW*, we use 1999 as a basis for

calculating the technical difficulty index (TDI) for a particular time period of interest. For example, if $TDW_{\mathrm{Y99}}^{\mathrm{Y00}} = 1.2$ and $TDW_{\mathrm{Y00}}^{\mathrm{Y01}} = 1.1$, then $TDI^{01} = 1.2 \times 1.1$ for the year 2001 using 1999 as the base period. In reality, the technical difficulty of production should consider output volume, because increasing output while retaining constant TDI is more difficult to produce. Therefore, TD is expressed as $TD = Q \times (TDI - 1)$.

Suppose set S contains historical records on workforce and outputs measured over a three-month period. We can estimate possible workforce level to meet expected outputs $\mathbf{y}^{\circ}$ using Model (1), where $I = \{PE, EE, PIE\}$ and $J = \{Q, SL, TD\}$. The minimum total amount of the indirect workforce $THC(\mathbf{y}^{\circ}, S)$ can be obtained as follows:

$$
\begin{aligned}
THC(\mathbf{y}^{\circ}, S) &= \min_{\mathbf{x}} \left\{ \sum_{i \in I} x_i \middle| \mathbf{x} \in \hat{L}(\mathbf{y}^{\circ}) \right\} \\
&= \min_{\lambda, x} \sum_{i \in I} x_i \\
\text{subject to:} & \sum_{r \in S} x_i^r \lambda_r \le x_i, \quad \forall i \in I = \{PE, PIE, EE\}; \\
& \sum_{r \in S} y_j^r \lambda_r \ge y_j^o, \quad \forall j \in J = \{Q, SL, TD\}; \\
& \lambda_r \ge 0, \quad \forall r \in S.
\end{aligned}
\tag{2}
$$

$THC(\mathbf{y}^{\circ}, S)$ denotes the minimum total workforce requirement based on the experience of S.

3.3 *Link with single factor productivity indices*

Based on the feedback of domain experts, we note that fab managers may resist adopting the frontier Model (2). The hurdles include the difficulty to understand the technical content and managers' fears that use of the proposed model will hurt their performance; the term 'past best performance' tends to produce the most anxiety. This section addresses the link between the proposed frontier model and conventional productivity indices; the discussions provide a way to promote the new model and to convince managers that fabs may even gain some benefits in performance from applying Model (2).

In practice single factor productivity indices, ratios of one output to one resource, are tracked to evaluate and monitor the performance of production units, such as fab operations. Notably, it is frequently observed that no consistent conclusions can be made among a set of productivity values. Trade-offs always occur among different performance indices. The concept of 'best performance' in Model (2) indeed is based on pairwise comparisons. If each index of record A is no better than the corresponding index of B and at least one index is strictly worse, A is said to be dominated by B while B is the dominant of A. A record that is found to be dominated by any record has no chance to be the best. After completing all possible pairwise comparisons, the remaining non-dominated records are indifferent to draw any conclusion since all of them are better-off in some aspects and worse-off in others.[3] It should be noted that the definition of 'best' performance used in this study is conservative in the sense of disqualifying some resource-outputs records from being the best. Detailed decomposition of single factor productivity indices can be found in Chen and McGinnis (2007).

Figure 4 presents a simple example with two productivity indices, Y1/X and Y2/X, where Y1 and Y2 are two outputs generated by a resource X. In Figure 4(a) E is dominated by B and C (comparing E with C, they both have the same performance in Y2/X, but C is 'better' than E in Y1/X). C also dominates D since it is better in Y2/X although the same in Y1/X. No conclusions regarding dominance can be drawn for the pairwise comparisons related to A, B and C. Therefore, $\{A, B, C\}$ are all labelled as best performance according to the definitions. In Figure 4(b) all units are dominated by F in both Y1/X and Y2/X. The only non-dominated unit is F and therefore F has the best performance. The proposed method can handle both cases and identify the proper best performance records based on the definition.

In summary, we can use the non-parametric frontier approach to estimate the workforce required of a single fab in response to the delegated tasks for a single period. Model (2) should apply to each fab to estimate its own requirement based on the task assigned and its past experience. Notably this method imposes minimum detailed assumptions on the relationship of type of workforce and tasks, but is a deterministic empirical method. The proposed method approaches the problem from the aggregate macro viewpoint, and thus implicitly assumes all labour is cross-trained in some sense for the functions considered. It estimates a fab's need for one period but does not determine the overall optimal solution over multiple periods.

Indeed, the method can be viewed as a compromise from the perspective of both the central planning unit and the fabs such that controversy is dispelled. It uses the best experience from the past as the estimating basis which satisfies the interest of the corporate level and provides flexibility to fabs by selecting a single productivity index in their favour.

4. Productivity improvement

It is reasonable to assume, and thus to require, that any previously achieved level of productivity is continuously achievable in the future. However, the composition of S with

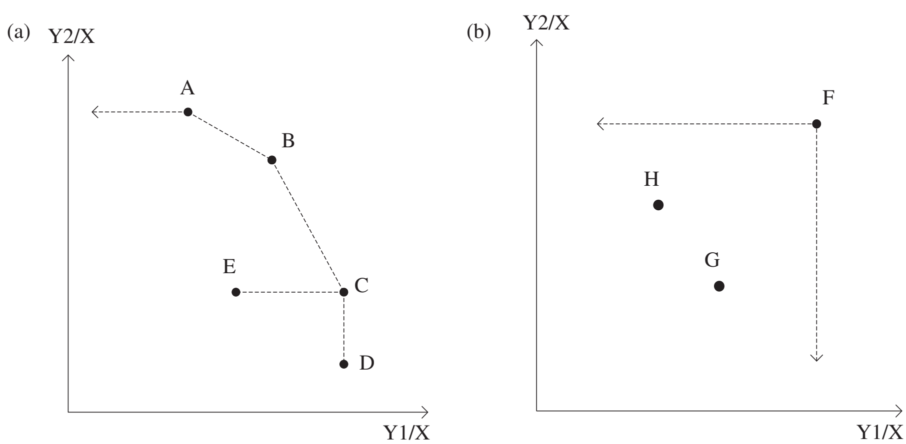

Figure 4. Examples with two productivity indices.

raw historical records leads to a conservative estimation. S does not consider the real or target growth of productivity indices for each fab, despite the fact that growth exists. Growth may come from some well-known production properties, such as increasing returns to scale and economies of scale, or the technology improvement and experience accumulation. On the other hand, the target growth is set to drive productivity improvement, as shown in Figure 1. Consider the dataset containing y_j^r for $j \in J$ and $r \in S$, and where y_j^r denotes output j produced by $\mathbf{x}^r$ during a given base time period b. Consider productivity growth with a predetermined annual growth rate α_j for output $j \in J^4$, and denote output j produced at time t using the same $\mathbf{x}^r$ as $y_j^r(t)$. y_j^r should be adjusted to $y_j^r(t)$ using α_j as follows:

$$y_j^r(t) = y_j^r \times \left(\alpha_j\right)^{t-b}. \tag{3}$$

Depending on the decision horizon of the analysis, α_j can be an annual, quarterly or monthly growth rate, but since an annual rate is typically given by the top executives, adjustment is needed for monthly or quarterly decision-making. For example, suppose the annual growth rate is 5%, the monthly rate $\alpha_j = (1.05)^{1/12}$. If $y_j^r = 100$ denotes the quantity of output j for record r occurring at time $b = 5$, e.g., the base month is May 1997, one should adjust using Equation (3) and the adjusted level for the target month December 1997, $t = 12$, is $y_j^r(t = 12) = 100 \times (1.004)^{12-5} = 102.887$, where $\alpha_j = (1.05)^{1/12} = 1.004$. This means that using the same input level and providing annual growth rate of 5%, output j should be 102.887; i.e., 102.887 is achievable if it is found to be 100 seven months ago for this particular record r. All records in S are modified to S^* using Equation (3) with respect to their time stamps. The total workforce considering productivity growth thus is calculated using Model (2) given substituting S by S^*.

We observe that the value of α_j is output-type dependent and assumed to be constant over time in Equation (3). In fact, saving resources can also result in productivity growth rate α, at least mathematically. In our real-world case only the output-oriented approach is adopted, because there is a possibility of low demand since the business culture of Taiwan generally discourages laying off workers. The only short-term workforce reduction action usually possible is instituting a recruitment freeze, and thus labour turnover. Moreover, the discount rate does not need to be constant, and an indexing system like the consumer price index (CPI) can be developed.

5. Workforce allocation

This section presents the proposed method to allocate the up-to-date total available workforce (*GHC*) to individual fabs based on their estimated workforce needs. As addressed in Section 2, workforce allocation is defined in the sense of responding to a single-period inconsistency between supply and demand but not the pure supply side planning based on determined labour demand (e.g., Wirojanagud *et al.* 2007, Fowler *et al.* 2008). The supply, *GHC*, is determined through long-term workforce planning and subjective top-down enforcement that drives productivity growth. The demand requested units (fabs) compete with each other so that fair (re)allocation is the major concern, ignoring the issue of workforce sufficiency. Perceived fairness among the stakeholders is critical when the shortage of workforce exists, and sharing shortage loading evenly places all fabs on the same basis for fulfilling tasks and performance evaluation.

We define the suffering index (SI) to measure the magnitude of suffering inconsistency between workforce supply and demand by normalising allocated workforce level as:

$$SI \equiv \frac{\text{required workforce}}{\text{allocated workforce}}.$$

The ideal condition is clearly $SI = 1$. $SI > 1$ indicates that the corresponding fab suffers a labour shortage. Larger SI indicates a more serious shortage. Further, $SI < 1$ indicates that a fab has a workforce exceeding requirements. This situation may negatively impact productivity and is undesirable for fab managers although the workforce is sufficient. Note the required workforce in SI is given by Model (2).

The allocation is intended to minimise the largest (worst) suffering index of all fabs. F denotes the set of fabs under consideration, where the objective function is expressed as $\min\max_{k \in F} SI^k$. Hereafter, the superscript represents fab k. Notably, the minmax objective function can be rewritten as:

$$\begin{aligned} &\min_{v} v \\ &\text{subject to: } v \geq SI^k, \quad \forall k \in F. \end{aligned} \tag{4}$$

Having a significant change of the workforce level for each fab is undesirable. Associated problems include costly learning curves and damage to morale. Therefore, it is preferable for the newly allocated workforce to be minimally changed in terms of volume and mix, or to be within the pre-specified tolerance. Given that fab k currently is allocated workforce CHC^k, the allocated workforce amount HC^k should result in a percentage change in the predetermined tolerance ε_k, e.g., $\varepsilon_k = 20\%$. That is:

$$\frac{HC^k - CHC^k}{CHC^k} \leq \varepsilon^k.$$

Combining the factors above, the allocation of workforce to each fab can be determined by HC-allocation as:

HC-allocation:

$$\min_{v, HC^k} v.$$

Subject to:
minmax constraints:

$$v \geq \frac{THC^k(\mathbf{y}^k, S^{k*})}{HC^k}, \quad \forall k \in F;$$

individual capability constraints:

$$HC^k \leq (1 + \varepsilon^k) CHC^k, \quad \forall k \in F \tag{5}$$

$$HC^k \geq (1 - \varepsilon^k) CHC^k, \quad \forall k \in F \tag{6}$$

$$\sum_{k \in F} HC^k = GHC$$

$$HC^k \geq 0,$$

where $THC^k(\mathbf{y}^k, S^{k*})$ is computed by Model (2). It should also be noted that imposing Constraints (5) and (6) may lead to infeasibility of the problem, and that infeasibility provides feedback for improper task (product) allocation or other managerial issues. HC-allocation is a non-linear programming problem, but its optimal solutions can be obtained without solving HC-allocation as follows. Without considering Constraints (5) and (6), it is possible to optimise the allocation of the workforce to fab k:

$$HC^{k*} = GHC \times \frac{THC^k(\mathbf{y}^k, S^{k*})}{\sum_{k\in F} THC^k(\mathbf{y}^k, S^{k*})}. \tag{7}$$

While Constraints (5) and (6) are being considered, if HC^{k*} in Equation (7) violates (5) or (6), HC^{k*} is $(1+\varepsilon^k)CHC^k$ or $(1-\varepsilon^k)CHC^k$, respectively; the workforce assignment of fab k is complete. Equation (7) is used to reallocate the remaining workforce (by resetting GHC to remaining level) to the remaining fabs, and summarised and implemented using *allocateHC* (Figure 5).

6. Case study

To estimate the validity of the proposed method, a case study was conducted in a leading foundry in Taiwan that has four 8-inch fabs with different capacity scales and configurations. The corporate planning unit has assigned tasks to the fabs for the next quarter, and now needs to individually (re)allocate the predetermined (may be insufficient) overall workforce level. Real fab-level operational data is collected. The detailed definitions of workforce and task types collected are identical to those addressed in Section 3.2 but on a monthly basis. Because the real-world decision is on a quarterly basis, we aggregate information from three successive months to represent a quarterly record. For each fab there are a total of 28 records in the data set representing 28 different quarterly resource-output data which is used in Model (2). We use the first month of the

```
procedure allocateHC(GHC, ε^k, CHC^k, THC^k(ỹ, S^k*), ∀k ∈ F)
while GHC > 0
begin
    HC^k := GHC × THC^k(ỹ, S^k*) / Σ_{k∈F} THC^k(ỹ, S^k*), ∀k ∈ F
    for k ∈ F
    begin
        if HC^k > (1+ε^k)CHC^k
            HC^k := (1+ε^k)CHC^k,
            F := F − k,
            GHC := GHC − HC^k
        if HC^k < (1−ε^k)CHC^k
            HC^k := (1−ε^k)CHC^k,
            F := F − k,
            GHC := GHC − HC^k
    end
end
return HC^k, ∀k ∈ F
```

Figure 5. The pseudocode of the procedure *allocateHC*.

quarterly record as the time stamp of the record, e.g., record 3 is the aggregate of months 3, 4 and 5.

Table 1 illustrates the effectiveness of the entire decision framework. Due to business confidentiality, the data presented has been somewhat altered. There are four fabs and the overall corporate target level is 407 set by the top executive. Q, SL and TD are tasks assigned to each fab next quarter; applying Model (2) for each fab gives the estimated workforce requirement (required THC), where no productivity growth is required in this illustration. According to the target corporate quota 407, the current level of total workforce, and the reallocating tolerance (5%), we can reallocate the workforce to the four fabs based on HC-allocation, resulting in $SI = 1.08$ for all fabs. The allocated results are not integers but can be simply rounded off. Only fab D increases its workforce by 3.57% while fab C needs to reduce 4.5% of its workforce.

As Section 2 has explained, the challenge is estimating the workforce requirements. Recall that our estimation model must convince fab managers to dispel the controversy. Therefore, we specifically compare performance based on commonly used productivity indices, between with and without using Model (2). The results provide evidence for the discussion in Section 3.3, namely that our new method will not hurt performance, and thus can be safely adopted.

The information for workforce requirement estimation is the historical records collected from one of the four fabs. Figure 6 presents the productivity trends of the fab under study. To simplify presentation and ensure data confidentiality, all indices are normalised using corresponding values in period 1 as the base, while the y-axis represents the values. Each data point represents a normalised three-month productivity index value based on estimated and actual total workforce (THC), sum of three workforce types, and realised outputs. Circles, squares and triangles connected with dotted lines are the normalised actual values of the productivity indices Q/THC, TD/THC and SL/THC, respectively. Apparently, not all three indices have the same trend. Index SL/THC grows significantly during all time periods, possibly because of reductions in total workforce or increases in service loading (Figure 6). Indices Q/THC and TD/THC are less than one in periods 11 to 18 (Figure 6), primarily because of low demand in those periods. Overall, the SL/THC index exhibits the most significant growth even in the current economic downturn, while the Q/THC and TD/THC exhibit declines in some periods. This observation derives from the growth in number and variation of customers. As a result the role of the fab changes from quantity-oriented to service-oriented.

Table 1. Workforce allocation for four fabs.

fab	Q	SL	TD	Current THC	Required THC^a	Allocated THC^b	SI	ΔTHC (%)
A	1448	54	444	58	62.07	57.58	1.08	−0.34%
B	2370	84	733	99	105.47	97.85	1.08	−1.23%
C	2521	92	782	113	116.43	108.01	1.08	−4.54%
D	2511	91	757	139	154.74	143.56	1.08	3.57%
Total				409	439	**407**[c]	1.08	

Notes: [a]estimated by Model (2) without productivity growth;
[b]tolerance = 5% for all fabs;
[c]total target.

Model (2) is applied to determine required workforce level for the periods after period 13. The workforce in period 13 is estimated using data from periods 1 to 12; each record is added to the data set S to estimate the requirements for the next period. Solid circles, squares and triangles connected by solid lines are the normalised values of productivity indices Q/THC, TD/THC and SL/THC respectively, using actual output values and estimated workforce (labelled 'Frontier' in Figure 6). The normalised base is identical to the actual data. Generally, the estimated productivity values exhibit similar patterns to the real values; however, the actual productivity values outperform those suggested by the model, because the actual values include productivity improvement, while the estimations accurately reflect the best past experiences. The productivity values based on the estimation lag the real values especially in situations involving growth. Figure 6 also shows how the resulting productivity values correspond to the best past values as discussed in Section 3.3. All index values estimated are no better than the best from the past. In six out of 16 records, SL/THC captures the best past values with a lag, e.g., period 15 has the same SL/THC index value as period 14, which itself is the best value prior to period 15. The last three periods demonstrate that SL/THC fails to reach a new best level while Q/THC and TD/THC record best-ever levels. This suggests a possible change in output mix, where the growth rate of SL/THC stabilises while the other two growth rates increase steeply.

Figure 7 presents the productivity values based on estimation with productivity adjustments. Figures 6 and 7 are interpreted identically. The only difference is that records used for estimation are adjusted according to Equation (3). The expected service loading (SL) productivity growth in period 28 is 1.56 compared to period 1, namely $\alpha_{SL} = (1.56)^{1/27}$. Similarly, Q and TD are expected to exhibit overall growth of 1.4. The growth rates are obtained from the interview with the studied firm based on current observations of firm productivity. Figure 7 further indicates that the estimated

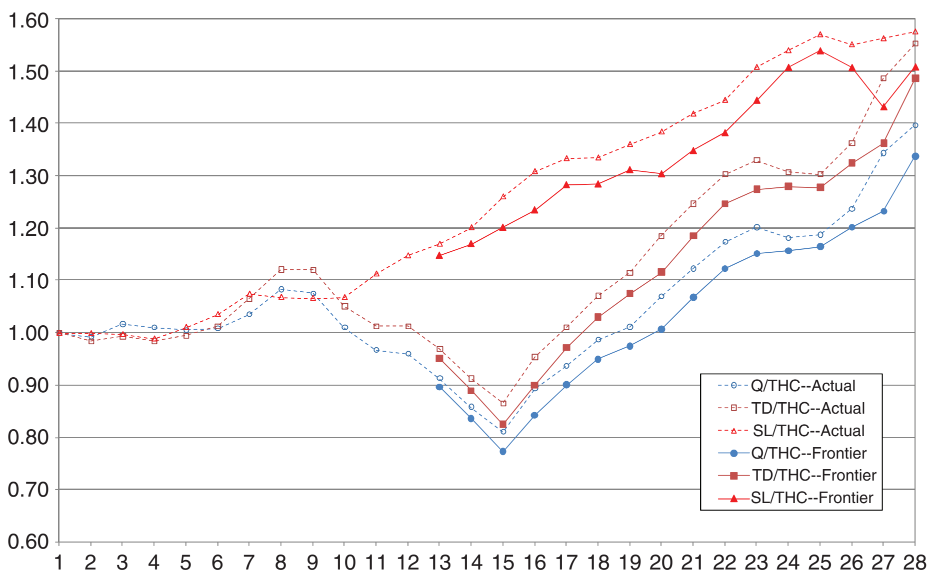

Figure 6. Values of workforce productivity.

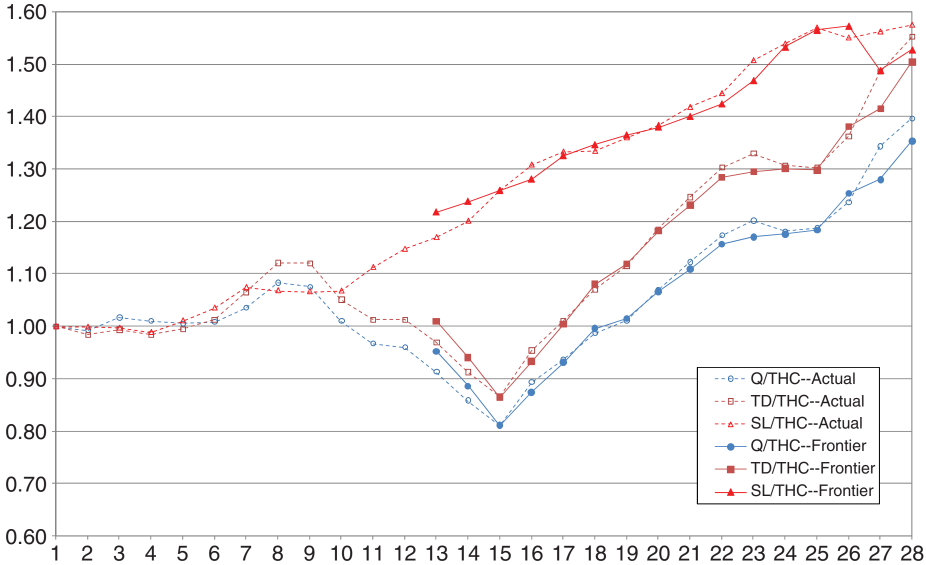

Figure 7. Values of workforce productivity (with adjustment).

productivity values are quite close to the actual ones. Those time periods with value gaps (e.g., period 23 and the last two periods) reflect relatively significant changes in growth rates or possible output mix changes, and the growth rates adopted in Equation (3) are constant over time.

7. Conclusion

Motivated by the practical need of a semiconductor firm in a real setting, this paper has proposed a new decision-making framework for (re)allocating an indirect workforce to individual fabs to enable the completion of delegated tasks. Rather than optimising workforce supply decisions, the proposed approach provides a decision framework and mechanism to reduce the gap between workforce demand and available supply, particularly when multiple, competing units require workforces, in which estimating workforce requirement in response to the tasks assigned is critical to ensure the sense of fairness among the stakeholders. In particular, the proposed approach is two-stage: approaching the problem from a macro viewpoint, workforce requirements can be estimated based on historical experience. The optimal historical experience is set as the 'ideal' requirement. Given the total available workforce at the corporate level and the estimated requirements of individual fabs, the workforce is (re)allocated among individual fabs to balance the workload. The model also incorporates productivity growth.

Indeed, the proposed model can be extended to other industries, especially those with the following characteristics: (1) a central decision-maker allocates workforce to several agents after assigning them different tasks (outputs); (2) agents compete in their performance; (3) the workforce-output standards do not exist or are not clear especially when there are multiple outputs; (4) subjective productivity growth is enforced; and/or

(5) the real need of workforce is asymmetric information between the central decision-maker planner and the agents.

Notably similar application of the proposed model to the real-world firm occurred prior to the economic 'crash' in October 2008. Absent the global downturn and using the model for workforce decision-making, the Taiwanese firm was on track to increase its profits and to expand its capacity. We suggest further research to develop applications for recessionary and highly volatile periods. The proposed method relies on the best past performance method, ensuring the feasibility of the estimation, and reducing subjective, top-down judgments and challenges from fab managers. However, careful attention should be devoted to creating appropriate incentives to encourage production sites to become self-motivated for productivity improvement. Otherwise, workforce determination based on past performance may result in little or no improvement. Another side-effect is that employees and managers may hide their optimal capabilities to obtain a better resource allocation and perhaps achieve more significant productivity improvements in the future. This behaviour is observed and named the *ratchet effect* in many applications. However, these are topics for additional research.

Acknowledgements

This research is supported by National Science Council, Taiwan (NSC 97-2221-E-009-113) and Taiwan Semiconductor Manufacturing Company (96A0279J8).

Notes

1. For the purpose of this decision estimate, 'real' means feasible and optimised, indicating the requirements under the ideal situation absent inefficiency.
2. In fact, this type of problem is typically due to the fact that central planners and agents do not have the same interests, and has been studied in the context of agent theory (see, e.g., Laffont and Martimort 2002).
3. This is the concept of Pareto (in)efficiency in economics (McGuigan *et al.* 1999).
4. More precisely, it is one plus the growth rate and should be no less than one, i.e., no technology recession is allowed.

References

Ahn, H.-S., Righter, R., and Shanthikumar, J.G., 2005. Staffing decisions for heterogeneous workers with turnover. *Mathematical Methods of Operations Research*, 62 (3), 499–514.

Aykin, T., 2000. A comparative evaluation of modeling approaches to the labor shift scheduling problem. *European Journal of Operational Research*, 125 (2), 381–397.

Bard, J.F., Morton, D.P., and Wang, Y.M., 2007. Workforce planning at USPS mail processing and distribution centers using stochastic optimization. *Annals of Operations Research*, 155 (1), 51–78.

Burke, E., *et al.*, 2004. The state of the art of nurse rostering. *Journal of Scheduling*, 7 (6), 441–499.

Chen, W.-C. and McGinnis, L.F., 2007. Reconciling ratio analysis and DEA as performance assessment tools. *European Journal of Operational Research*, 178 (1), 277–291.

Chien, C.-F. and Chen, L., 2007. Using rough set theory to recruit and retain high-potential talents for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 20 (4), 528–541.

Ernst, A.T., *et al.*, 2004. Staff scheduling and rostering: a review of applications, methods and models. *European Journal of Operational Research*, 153 (1), 23–27.

Fowler, J.W., Wirojanagud, P., and Gel, E.S., 2008. Heuristics for workforce planning with worker differences. *European Journal of Operational Research*, 190 (3), 724–740.

Gans, N. and Zhou, Y.-P., 2002. Management learning and turnover in employee staffing. *Operations Research*, 50 (6), 991–1006.

Holder, A., 2005. Navy personnel planning and the optimal partition. *Operations Research*, 53 (1), 77–89.

Laffont, J. and Martimort, D., 2002. *The theory of incentives: the principal-agent model*. Princeton, NJ: Princeton University Press.

McGuigan, J.R., Moyer, R.C., and Harris, F.H., 1999. *Managerial economics: applications, strategy, and tactics*. 8th ed. Cincinnati, OH: South-Western College Publishing.

Mundschenk, M. and Drexl, A., 2007. Workforce planning in the printing industry. *International Journal of Production Research*, 45 (20), 4849–4872.

Pesch, E. and Tetzlaff, U., 2005. Scheduling personnel for press machines in the automotive industry. *Pacific Journal of Optimization*, 1 (3), 545–564.

Thompson, G.M. and Goodale, J.C., 2006. Variable employee productivity in workforce scheduling. *European Journal of Operational Research*, 170 (2), 376–390.

Weygandt, J.J., Kieso, D.E., and Kimmel, P.D., 2005. *Managerial accounting: tools for business decision making*. 3rd ed. New York: John Willey & Sons.

Wirojanagud, P., *et al.*, 2007. Modelling inherent worker differences for workforce planning. *International Journal of Production Research*, 45 (3), 525–553.

Zulch, G., Rottinger, S., and Vollstedt, T., 2004. A simulation approach for planning and re-assigning of personnel in manufacturing. *International Journal of Production Economics*, 90 (2), 265–277.