J Intell Manuf (2011) 22:447–457
DOI 10.1007/s10845-009-0302-x

Measuring relative performance of wafer fabrication operations: a case study

Wen-Chih Chen · Chen-Fu Chien

Received: 4 July 2008 / Accepted: 6 July 2009 / Published online: 25 July 2009

Abstract This paper was motivated by a request to review relative operations performance for various fabrication facilities within a leading Taiwanese semiconductor manufacturer. Performance evaluation is important but often controversial. To dispel the controversy, we propose a two-stage fabrication process model to systematically analyze metrics currently adopted, and show that the commonly used wafer-based indices are biased for operations performance. Instead, they should be decomposed into productivity, representing true operations performance, and manufacturability. We suggest the use of data envelopment analysis because of its confirmed linkages to other widely used productivity measures and its overall performance via relative comparisons. The case study illustrates how the two-stage model evaluates and analyzes real-world operations, and the empirical results show the drawbacks of conventional methods.

Keywords Performance evaluation · Fab operations · Semiconductor manufacturing · DEA

Introduction

Since the global semiconductor industry is highly competitive, properly utilizing resources to provide products and services is essential for maintaining firms' competitive edges and survival. Various types of businesses regard relative performance analysis as a foundational tool for monitoring, diagnosing and improving activities and processes, determining future strategies and suggesting improvements. The performance comparison can be inter- or intra-firm. Inter-firm analyses compare different organizations which are not controlled or planned by a central decision-maker. In contrast, intra-firm comparison compares units within an organization and the results are monitored and used by the central decision makers. Particularly, intra-firm comparison is used for periodic review in the sense of competition and directly related to rewards including future promotion and compensation. Units under evaluation are thus very sensitive to the performance evaluation rules and methods; controversies often arise in response to the poor performance.

This paper was motivated by a request in real setting to evaluate (compare) the operations performance for various wafer fabrication facilities (fabs) within a leading Taiwanese semiconductor manufacturer. The evaluation is an intra-firm comparison, and is directly related to fab managers' reputations and compensations. This paper uses the term fab operations to refer to all activities in a fab that involve consuming resources to provide outputs such as products or services, and excluding finance, sales and technology development. Indeed, fab operations performance significantly interacts with capacity and resource/product allocations; these factors can influence the performance and evaluation result supplies feedback about past decisions. Although the operations performance evaluation is important, as addressed, it is controversial. Indeed, fab managers often refuse to accept the poor performance evaluation result and question the method. How to conduct a practically proper and convincing performance evaluation is the major challenge.

Performance evaluation processes include defining metrics and concluding the performance. A performance evaluation method should consider: (1) *Inclusiveness*—it should

W.-C. Chen (✉)
Department of Industrial Engineering and Management, National Chiao Tung University, 30010 Hsinchu, Taiwan
e-mail: wenchih@faculty.nctu.edu.tw

C.-F. Chien
Department of Industrial Engineering and Engineering Management, National Tsing Hua University, 30013 Hsinchu, Taiwan

consider all aspects of the organization (Beamon 1999); (2) *Reflectibility*—it should consider only the aspects responsible for the organization, (3) *Convergence*— it should provide consistent and simple information (Bhargava et al. 1994). Defining appropriate performance metrics and correctly collecting values underlie the success of the evaluation process regarding to inclusiveness and reflectibility. The metrics selected must be sufficient to represent all necessary important objectives of the target activities, in this paper, fab operations. There are trade-offs among objectives, and they often lead to inconsistent conclusions based on different indices. Thus, it is desirable to establish a single index effectively representing the overall performance, and then closely analyze the performance index and the possible drivers.

A literature survey reveals several methods for measuring equipment performance in semiconductor manufacturing. For example, overall equipment efficiency (OEE) is proposed as a comprehensive index for individual equipment performance. See SEMI E079-0200 (SEMI E079-0200), Jeong and Philips (2001) and Muchiri and Pintelon (2008). Since OEE focuses on a single machine, extensions have been created to measure performance in broader, more complex semiconductor manufacturing systems (e.g., Huang et al. 2003; Chien et al. 2007). Nevertheless, they are detailed and bottom-up approaches based on equipments and other relative policies.

Other studies of semiconductor industry performance evaluation look at macro-operational aspects. Leachman and Hodges (1996) provide the first complete benchmarking of the competitive semiconductor manufacturing (CSM) program; this program includes several of the leading companies around the world. The authors propose seven KPIs: cycle time per wafer layer, line yield, die yield, stepper productivity, direct labor productivity, total labor productivity, and on-time delivery. Associations between practices and performance are also presented. Although it is the first industry-wide benchmarking study with rich data, the performance is based on various KPIs, and the conclusions should be regarded as the correlations between practices and a particular KPI, and not the overall performance. Leachman et al. (2007) extend the studies in CSM program and use data envelopment analysis (DEA) to provide an overall relative efficiency measure. Total wafer starts, number of steppers, direct and indirect headcounts, and clean room size are the inputs, and die output and effective revenue the outputs. However, these two studies are inter-firm comparisons, not for relatively controversial inter-firm evaluation.

This paper aims to evaluate the performance of the fabrication site for period review, not a single production line or tools. To our experience, practitioners still utilize various wafer-based KPIs when assessing fab performance. The major hurdle in adopting methods proposed in the literature such as DEA is that early studies do not provide a clear rationale for the indices selected, and DEA is less intuitive than traditional wafer-based indices. Moreover, to dispel the controversy, it is necessary to develop: (i) a systematic way to express, model and organize performance metrics; (ii) a method to provide consistent performance conclusions. Therefore, this study aims to propose a two-stage fabrication process model that systematically analyzes metrics currently adopted, and show that using wafer-based metrics are problematic since it violates reflectibility. In particular, DEA is employed to compute the overall relative performance, in which we carefully link DEA theory and the currently adopted ratio methods. To estimate the validity of the proposed approach, we conducted a case study in real setting to demonstrate how the proposed model assists in diagnosing operations and suggesting improvements. The empirical results also reveal the bias of wafer-based KPIs and how it leads to poor decisions.

The paper is organized as follows. In the next section, we introduce a two-stage model for the fabrication production process. The model gives a clearer picture of fabrication and a better reasoning of the performance metrics. we then introduce DEA as the analysis technique and discusses some of the model's appealing properties. Next we apply the model to a real-world case study followed by conclusion.

Fabrication production process

This section presents a two-stage model to describe (and understand) fabrication production process, and thus a rationale basis to discuss, model and organize essential performance metrics is provided to dispel the possible controversy. Particularly, we approach the problem by reviewing the all performance ratios including those based on wafer outputs, which are currently adopted. The model shows using wafer-output as a metric violates reflectibility since it also includes the aspect beyond the responsibility of fab operations.

The operations of a fab are the activities and decisions for transforming resources, e.g., labor, equipments, etc., into products (outputs). Productivity describes the resource-output transformation. A "good" production unit, i.e., one with high productivity, can provide more outputs by using fewer resources. Productivity is measured formally using ratios of output to input; in nontechnical contexts, ratios of input to output are also used (e.g., labor hours needed to produce a car). Both measures give identical conclusions with opposite improvement directions. Wafers are the physical final products produced and delivered. In practice, the typical productivity indices adopted, such as labor or equipment productivity, use total wafers produced as the output. However, the use of this output is questionable and may underestimate fab operations performance.

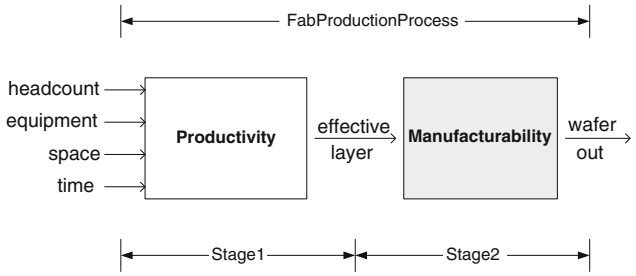

Fig. 1 Fab production process

This paper suggests a two-stage model to describe the transformation process from resources to final products; the model provides a rationale basis to define proper performance indices. Figure 1 shows the two-stage process. Stage 1 refers to *productivity*, the process that provides masking layers (outputs) by consuming labor, capital investments including equipments and space, and time. The first three resources are commonly found in other studies. Time, which is not as straightforward as the first three, is the total time used in the production corresponding to a particular volume of layers. Given the same level of labor, equipment and space, more layers require more time. Similarly, less labor (equipment) generally requires more time to generate the same required layers, and reveals the substitutability among resources. It is helpful to interpret the output of Stage 1 (number of layer processes) as a service provided rather than a product. A good performance indicated in Stage 1, namely providing more layers by using fewer resources, suggests that the production process is productive.

Stage 2 is the process of transforming layers to wafers. Different types of wafer products require different numbers of layers and the transformation itself is termed *manufacturability*. Design for manufacturability (DFM) is an engineering concept that includes a set of methodologies for designing products so that they are simple to manufacture. Manufacturability derives from both the engineering side (particularly in R&D activities of design or manufacturing, i.e., manufacturing recipes), and the strategic side (i.e., allocating products to specific fabs). A fab that is assigned products that are easier to produce will generally exhibit superior performance. In other words, a greater number of wafer outputs with fewer layers are always preferred since it means fewer steps in manufacturing.

The combinations of resource-output pairs in both stages are identical to the most commonly used productivity indices. Ratio (layer/headcount) is the layer labor productivity measure, and (layer/equipment) is the layer equipment productivity. Index (layer/space) corresponds to the effectiveness of space usage, and (time/layer) is the layer cycle time. With respect to wafer, combining wafer output and the four resources gives the same meaning as the four indices. For example, wafer output per headcount is the commonly used wafer labor productivity.

Although number of wafer outputs is commonly used as output in productivity indices, the two-stage model shows that this output may skew the performance of fab operations. Only Stage 1, productivity, stands for real operations performance. Poor performance using wafer outputs consists of activities determined and contributed by other functions (e.g., R&D in manufacturing recipes) and central production planning (e.g., in product allocation). Good operations (productivity) may yield less wafers due to poor manufacturability. In fact, the "real" responsibility of fab operations is to process various layers based on certain manufacturing recipes using the resources at hand. That is, the operations consume resources to provide services—the layer processing—so that the final product—the wafer—is produced via completing the required processes. The two-stage model provides a more accurate picture of this process, and thus the performance evaluation can be conducted with fewer arguments, a more effective diagnosis, and proper identification of departmental responsibilities.

For practical implementation, detailed definitions of resources and outputs for monthly performance reviews are listed as follows:

- Layers (L): total number of effective masking layers produced monthly. It is collected as (total layers produced) $\times$ (average layer yield rate).
- Wafer outs (W): total number of effective wafers produced monthly. It is collected as (total number of wafers produced) $\times$ (average wafer yield rate).
- Headcount (HC): total direct and indirect labor employed monthly. It also includes management and assistant staff.
- Equipments (MC): total installed capacity of steppers and scanners (exposure tools) used monthly. Since lithography equipment is generally the bottleneck among all machine types, it serves as the best proxy for the fab's equipment capacity. The literature uses total number of steppers and scanners, e.g., Leachman et al. (2007). Number of tools, however, ignores the difference in production capability of the machines. For example, new tools may have higher throughput than old machines, and/or a more expensive machine may have better throughput as well. The difference on resulting throughput can yield a significant impact on production performance. Hence, the weighted sum of installed tool capacity is a better proxy.
- Space (S): total floor space used and available for manufacturing. It reflects the infrastructure investment that becomes the limit of production capacity.
- Time (T): total time needed to produce the total number of layers. It is computed as (total number of layers) $\times$ (average cycle time per layer). It should be noticed that total layers, including effective layers and defects,

are considered. Time is a resource and defects result in resource waste; thus defects should be taken into account.

Service level is not considered in the proposed model. Good operations use fewer resources including time to provide more outputs and thus contribute to good delivery satisfaction levels. However, a fab operating efficiently will have a poor service level if it is allocated more outputs than it can produce.

Measuring operational performance

As noted earlier, there are up to four measures for any aspect in the proposed model. However, trade-offs exist among performance indices and they can lead to inconsistent conclusions. DEA, a mathematical programming approach introduced by Charnes et al. (1978), computes a single relative performance score while considering various resources and outputs simultaneously without assuming weights or functional form. Because of this attractive characteristic, DEA has been adopted for measuring performance in semiconductor manufacturing (e.g., Leachman et al. 2007). An introductory discussion on DEA can be found in Chen and Hong (2008), and Cooper et al. (2000) provide a comprehensive introduction.

Following the proposed two-stage fabrication process model, the output-oriented CCR DEA model (Charnes et al. 1978) is employed to measure the performance on productivity, manufacturability and wafer-based performance as follows:

$$\begin{aligned} &\max_{\phi_k^t, \lambda_j} \phi_k^t \\ &\text{s.t.} \sum_{j \in S^t} x_{ij}\lambda_j \le x_{ik} \quad \forall i \in I^t, \\ &\quad \sum_{j \in S^t} y_{rj}\lambda_j \ge \phi_k^t y_{rk} \quad \forall r \in O^t, \\ &\quad \lambda_j \ge 0 \quad \forall j \in S^t \end{aligned} \tag{1}$$

where $t \in \{p, m, w\}$ represents models for productivity (p) and manufacturability (m) and wafer-based performance (w), respectively. Sets I^t and O^t are the resource set and output set associated with t, namely, $I^t = \{HC, MC, S, T\}$ for $t \in \{p, w\}$, $O^p = \{L\}$, $I^m = \{L\}$ and $O^t = \{W\}$ for $t \in \{m, w\}$. Set S^t stands for the collected records for model t. Moreover, x_{ij} and y_{rj} are the amounts of resource i and output r associated with record j, respectively, given the corresponding model for record set S^t. Subscripts k represent a particular record under evaluation, and $k \in S^t$. It should be noted that (1) is output-oriented, i.e., to maximize outputs using given resources, because in reality there is less flexibility in reducing resources. The result also hints at the maximum outputs, the production capacity.

Equation (1) evaluates record $k \in S^t$ by comparing against all records in S^t. Let ϕ_k^{t*} $t \in \{p, m, w\}$ be the optimal values in (1) corresponding to three different models. They are the record k's performance scores for different aspects. For model t, value ϕ_k^{t*} indicates that the record k can generate ϕ_k^{t*} times for current output level while using the same level of resources, because a best practice comprised by the records in S^t can be identified and used as the comparison reference. The comprised best practice derives from three simple assumptions: engineering interpolation, free disposal and constant returns to scale (CRS). Free disposal means that if a resource-output bundle is feasible, it is also feasible to use more resources or to produce fewer outputs. CRS assumes that any observed output-resource ratio will hold constant for different sizes of outputs and resources. (CRS is commonly used.) For example, assume that $(x, y) = (2, 6)$, i.e., $y/x = 3$, then x should be 4 when $y = 12$. Moreover, the reciprocal of the score, $e_k^t = \frac{1}{\phi_k^{t*}}$ $t \in \{p, m, w\}$, normalizes the performance score to be within 0 and 1. A best practice has a value of 1; thus e_k^t can be interpreted as the efficiency of record k, i.e., the ratio of record k's performance to the best practice performance. Clearly, $\phi_k^{t*} \ge 1 (e_k^t \le 1)$ $t \in \{p, m, w\}$. $\phi_k^{p*} = 1$ suggests that record k is *CCR-efficient in productivity*. Larger value of ϕ_k^{p*}, smaller e_k^p, reveals poorer (less efficient) performance in productivity, i.e., it has a more significant edge losing to the best practices. The record k is said to be *CCR-efficient in manufacturability* if $\phi_k^{m*} = 1$, i.e., has products that are the easiest to produce. A higher ϕ_k^{m*} (or smaller e_k^m) shows less manufacturability, i.e., more difficult to produce (because the reference, the easiest to produce, goes through the same number of layer processes to comprise more wafer outputs). Similarly, $\phi_k^{w*} = e_k^w = 1$ indicates that record k is *CCR-efficient in wafer-based performance*; the interpretation is identical to productivity ($t = p$), but with wafer as the output.

The dual of Model (1), which will give the same optimal value of Model (1), is (Charnes et al. 1978):

$$\begin{aligned} &\min_{u_i, v_r} \sum_{i \in I^t} u_i x_{ik} \\ &\text{s.t.} \sum_{r \in O^t} v_r y_{rj} \ge \sum_{i \in I^t} u_i x_{ik} \quad \forall j \in S^t, \\ &\quad \sum_{r \in O^t} v_r y_{rk} = 1, \\ &\quad u_i \ge 0 \quad \forall i \in I^t, \\ &\quad v_r \ge 0 \quad \forall r \in O^t. \end{aligned} \tag{2}$$

Model (2) is equivalent to the following problem:

$$\min_{u_i v_r} \frac{\sum_{i \in I^t} u_i x_{ik}}{\sum_{r \in O^t} v_r y_{rk}}$$

$$\text{s.t.}\ \frac{\sum_{i\in I^t} u_i x_{ik}}{\sum_{r\in O^t} v_r y_{rj}} \geq 1 \quad \forall j \in S^t,$$
$$u_i \geq 0 \quad \forall i \in I^t,$$
$$v_r \geq 0 \quad \forall r \in O^t. \tag{3}$$

Without loss of generality, taking productivity model ($t = p$) as an example, Model (3) can be rewritten as:

$$\min \frac{u_{HC}x_{HC,k} + u_{MC}x_{MC,k} + u_S x_{S,k} + u_T x_{T,k}}{v_L y_{L,k}}$$
$$\text{s.t.}\frac{u_{HC}x_{HC,j} + u_{MC}x_{MC,j} + u_S x_{S,j} + u_T x_{T,j}}{v_L y_{L,j}} \geq 1$$
$$\forall j \in S^p,$$
$$u_{HC} \geq 0,\ u_{MC} \geq 0,\ u_S \geq 0,\ u_T \geq 0,\ v_L \geq 0. \tag{4}$$

Let $\frac{u_{HC}}{v_L} = w_{HC}$, $\frac{u_{MC}}{v_L} = w_{MC}$, $\frac{u_S}{v_L} = w_S$, and $\frac{u_T}{v_L} = w_T$, Model (4) is rewritten as:

$$\min w_{HC}\frac{x_{HC,k}}{y_{L,k}} + w_{MC}\frac{x_{MC,k}}{y_{L,k}} + w_S\frac{x_{S,k}}{y_{L,k}} + w_T\frac{x_{T,k}}{y_{L,k}}$$
$$\text{s.t.}\ w_{HC}\frac{x_{HC,j}}{y_{L,j}} + w_{MC}\frac{x_{MC,j}}{y_{L,j}} + w_S\frac{x_{S,j}}{y_{L,j}} + w_T\frac{x_{T,j}}{y_{L,j}} \geq 1$$
$$\forall j \in S^p,$$
$$w_{HC} \geq 0,\ w_{MC} \geq 0,\ w_S \geq 0,\ w_T \geq 0. \tag{5}$$

Model (5) evaluates record k and has a straightforward interpretation. Indeed, $\frac{x_{HC,j}}{y_{L,j}}$ is the number of headcount needed per layer associated with the commonly used labor productivity in the semiconductor industry. The index used here is the reciprocal of the formal academic labor productivity measure, which is another form used in practice to represent productivity from the resource consumption viewpoint. Analogically, $\frac{x_{MC,j}}{y_{L,j}}$ associates with the equipment productivity, and $\frac{x_{S,j}}{y_{L,j}}$ associates with the return rate on space. $\frac{x_{T,j}}{y_{L,j}}$ is the layer cycle time. Therefore, an overall index is provided and used for all records by weighting productivity values on four aspects in the objective function and constraints. Without assigning subjective predetermined weights of four individual indices, weights are selected in favor (to minimize the weighted index) of record k, while normalizing all records' score being no < 1 using the same weights.

The manufacturability model is more straightforward since there is only one resource and one output. The conventional single productivity ratio approach can be easily adopted. Model (1) will give the same results as using wafer per layer—"how many layers needed to become a complete wafer?"—for the comparison metric.

Model (1) is a useful tool to measure the overall performance based on resources and outputs. Applying (1) to the proposed two-stage model gives three overall performance scores for three different purposes. It should be noted that the conclusion consistency referring to the performance of a particularly aspect (or responsibility), e.g., productivity or manufacturability. The commonly adopted wafer-based performance is biased and should be decomposed into productivity and manufacturability. Manufacturability and productivity performance have different managerial meanings and are the responsibility of different functions within a firm. For fab operations performance, productivity is the proper aspect to be considered, and this is explained by the two-stage model addressed in section "fabrication production process".

Further, there are some concerns and limitations when implementing DEA. For examples, DEA is a weighting scheme in response to the need of consistency. Its evaluation results vary upon different resource-output sets, which should be carefully defined such as our two-stage model. Another important pitfall is that the records should be sufficiently larger than number of resources and outputs to have reasonable results. There are some useful references when applying DEA. For example, Golany and Roll (1989) suggest a standard procedure to apply DEA, and Dyson et al. (2001) provide a comprehensive review on the pitfalls of DEA.

Case study

This section presents the performance analysis results for a real-world firm in Taiwan. The original goal of the study is to find a way to properly review fab operations by comparing various fabs within the same firm, i.e., the intra-firm comparison. The four 8-inch fabs operating under similar environments are studied (reviewed together) via peer comparison. One hundred fifty-six records collected in a 19-month period in 2000s represent the monthly resources consumed and outputs produced. To maintain confidentiality, all records are transformed and then pooled to compute three different performance scores (wafer-based performance, productivity and manufacturability) based on Model (1). The efficiency scores $e_k^t \quad t \in \{p, m, w\}$ are presented because the values are between 0 and 1 and easier to read.

To estimate the validity of the proposed approach, we conducted a case study in real setting to demonstrate how to use the proposed performance evaluation model; show how wafer-based performance is biased. We also discuss the ways in which the bias affects the decisions. In addition, sections "scale vs. productivity" and "resource slacks" present byproducts of the tool we proposed, which include the scale issue and possible ways to monitor the resources slacks.

Performance analysis

Figure 2 is the box plot for the wafer-based efficiency scores for the four fabs. The y-axis represents the efficiency scores obtained by (1), that is, the reciprocal of optimal value of Model (1), where value 1 is the best and the smaller values indicate poorer performance. The x-axis associates with the fabs. Figure 2 shows the fabs' very different performance

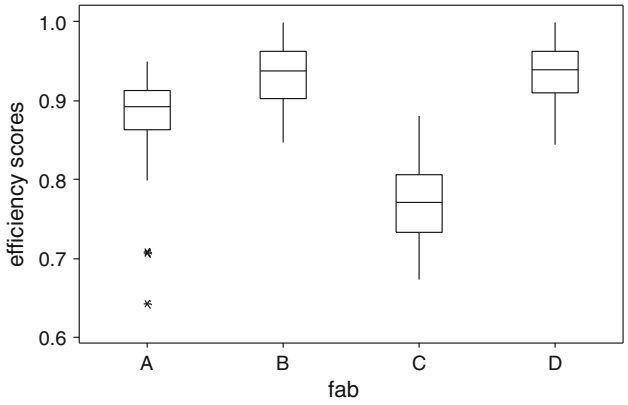

Fig. 2 Wafer-based efficiency

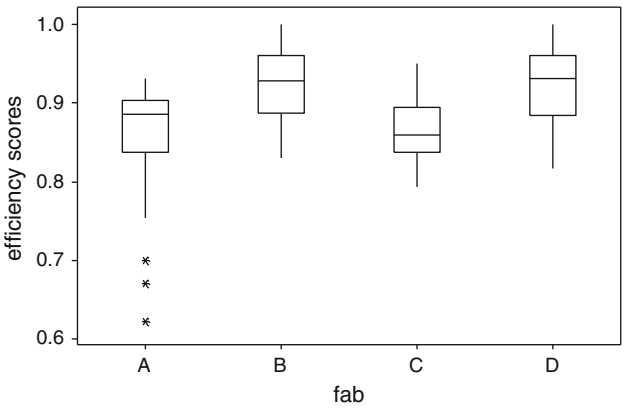

Fig. 3 Productivity efficiency

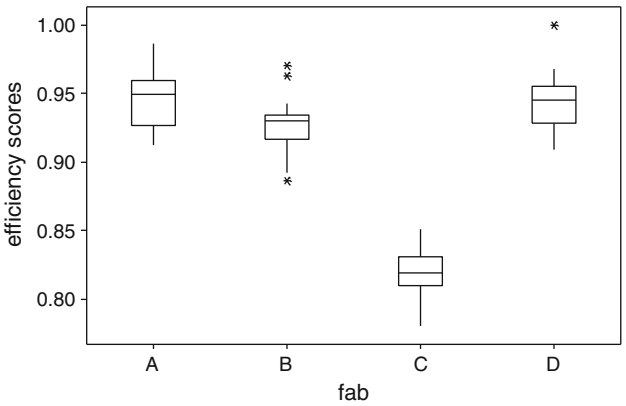

Fig. 4 Manufacturability efficiency

ranges. Fabs B and D perform slightly better with average scores around 0.93, while fab C is the most inefficient with an average score around 0.77. The difference suggests that fab C does not transform resources to wafer outputs efficiently; there is roughly a 20% performance gap compared to fabs B and D.

Figures 3 and 4 are the box plots regarding productivity and manufacturability, respectively. Figures 2 and 3 have the same interpretation. Manufacturability efficiency score 1 indicates that it is the easiest to produce, while a smaller score suggests a more complex product (Fig. 4). Efficiencies related to productivity have fewer differences among the fabs, particularly for fab C (Fig. 3). Although fab C is still the worst on average, the gap of the mean is significantly reduced to <10% compared to fabs B and D. Fabs A and D have average manufacturability efficiency scores around 0.94, and fab B is slightly worse than A and D (Fig. 4). Fab C has the poorest manufacturability; half of its records have efficiency scores from 0.81 to 0.83. The comparison suggests that fab C produces products with special or complicated processes (Fig. 4). In fact, fab C mainly produces memory products that typically require more layers than logic products, i.e., memory fabs have poorer manufacturability than logic fabs. Comparing Figs. 2, 3 and 4 shows that fab C has poor wafer-based performance and manufacturability, but with a much smaller gap in productivity than the other fabs, again because it produces more complex products.

Figure 5 shows the efficiency scores for fab A over time. The x-axis represents the time stamp in month and the y-axis the efficiency scores. Squares are scores for productivity, triangles are for manufacturability, and circles are for wafer-based performance. There are no significant trends, either improvement or decline, on productivity and wafer-based performance overall, but there is a decline in manufacturability (Fig. 5). Figures 6, 7 and 8 represent efficiency across 39 months for fabs B, C and D. The interpretations are identical to Fig. 5. All four figures show relatively stable manufacturability. Manufacturability for fabs B and D is slightly worse, while fab C trends in a positive direction (as noted, it has worse manufacturability than the other fabs). Figure 7 shows significant gaps between wafer-based performance and productivity compared to Figs. 5, 6 and 8. The differences are due to fab C's poor manufacturability.

In addition, a comparison of the three different efficiency scores finds that wafer-based performance and productivity have almost identical patterns, yet with different magnitudes. One the other hand, manufacturability is relatively independent on the growth or decline pattern change. Periods 10–15 in Fig. 5 provide a significant example, where wafer-based and productivity performance drop tremendously in period 11 from about 0.85 to below 0.65 while manufacturability remains stable in score. Although manufacturability has relatively stable trend over time, larger gap between wafer based performance and productivity arises when average manufacturability efficiency is low (e.g., Fab C in Fig. 7). Therefore, we can conclude that wafer-based performance is mainly contributed by productivity and manufacturability plays a role of multiplier.

Table 1 presents the results of simple linear regression analyses for three models in which the independent variable

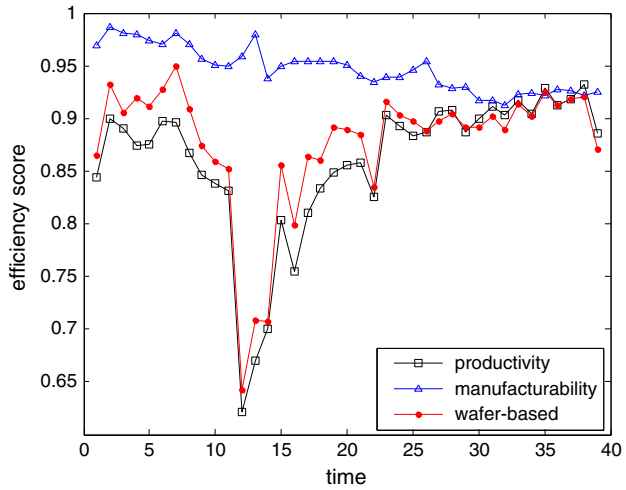

Fig. 5 Changes of efficiency scores (Fab A)

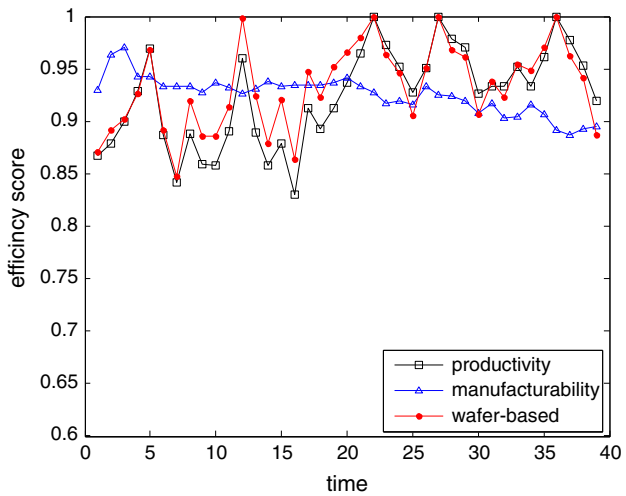

Fig. 6 Changes of efficiency scores (Fab B)

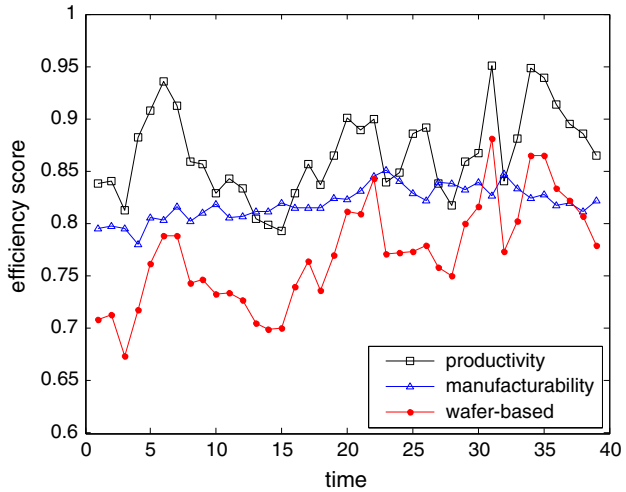

Fig. 7 Changes of efficiency scores (Fab C)

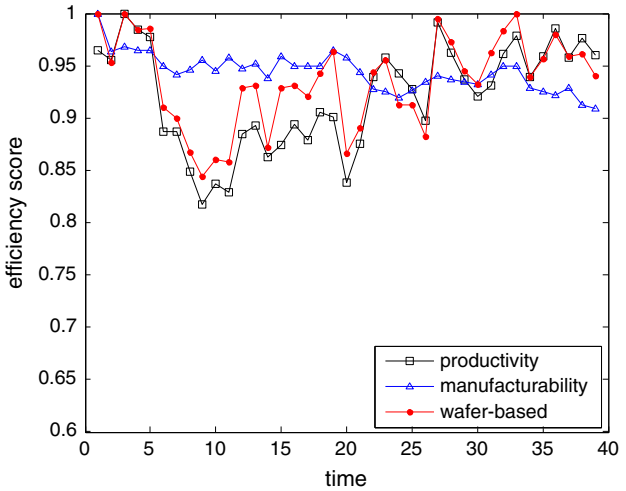

Fig. 8 Changes of efficiency scores (Fab D)

is time and the dependent variables are efficiency scores. Instead of slope coefficient, we present the annual growth rate, namely, 12 × (slope coefficient). For wafer-based performance, fab C has the best growth rate (3.7%); fab B's growth rate is 2%. Wafer-based growth rates associated with fabs A and D are not statistically significant. Fab A has the best growth rate (3%) in the productivity model, followed by fab B (2.9%), and fabs C and D (below 2%). For manufacturability, fabs A, B and D have negative annual rates, suggesting that the products are becoming more complex to produce over time. However, fab C's positive manufacturability growth rate (1.1%) shows that the products assigned to it are becoming easier to produce.

These observations demonstrate both the biases of wafer-based performance and how manufacturability affects wafer-based performance. Thus, a detailed decomposition of wafer-based performance is recommended, and it is preferable to use layer as the output to evaluate fab operations.

We note that fab C is the worst in wafer-based performance with significant loss (Fig. 1), mainly due to product allocation so that fab C also has the poorest manufacturability (Fig. 2). Interviews with company personnel reveal that product complexity is always a key challenge in reviewing performance using traditional wafer-based indices. Consequently, the production planning department has made modifications in product allocation. We observe that the manufacturability for fab C is improving (Fig. 7); fab C is the only one with improvements in manufacturability (1.1% in Table 1). The adjustment made for fab C results that fab C has the best annual wafer-based performance growth rate (3.7%), which has significant edge to others, but the worst annual growth rate for productivity (1.5%). Therefore, we surmise that there may be an over-adjustment. Using conventional methods, manufacturability is a masking effect for argument and so that the real problem is possibly ignored. Moreover,

Table 1 Simple linear regression over time

Fab	Productivity		Manufacturability		Wafer-based	
	Rate (%)[a]	p Value	Rate (%)[a]	p Value	Rate (%)[a]	p Value
A	3.0	0.01	−1.6	0	1.3	0.232
B	2.9	0	−2.0	0	2.0	0.003
C	1.5	0.032	1.1	0	3.7	0
D	1.9	0.024	−1.4	0	1.1	0.152

[a] 12 × Slope coefficient

manufacturability may have significant multiplier effect to wafer-based performance, namely wafer-based performance is sensitive to change in manufacturability. This effect should be taken into account when make resource allocation decisions.

Scale vs. productivity

The following two sub-sections investigate relative productivity performance ($t = p$) comparisons among different fabs under the well-studied DEA framework. This section analyzes the relationship between productivity and scale in the data set.

Another popular DEA model proposed by Banker et al. (1984) is:

$$\begin{aligned} &\max_{\theta_k^p, \lambda_j} \theta_k^p \\ &\text{s.t.} \sum_{j \in S^p} x_{ij}\lambda_j \le x_{ik} \quad \forall i \in I^p, \\ &\quad \sum_{j \in S^p} y_{rj}\lambda_j \ge \theta_k^p y_{rk} \quad \forall r \in O^p, \\ &\quad \sum_{j \in S^p} \lambda_j = 1, \\ &\quad \lambda_j \ge 0 \quad \forall j \in S^p. \end{aligned} \tag{6}$$

Model (6) has additional convexity constraint, $\sum_{j \in S^p} \lambda_j = 1$, compared to (1). The difference of optimal values for a particular record k between (1) and (2), $\frac{\phi_k^{p*}}{\theta_k^{p*}}$, measures the scale effect, named *scale efficiency* (Banker et al. 1984). Clearly, $\frac{\phi_k^{p*}}{\theta_k^{p*}} \le 1$; $\frac{\phi_k^{p*}}{\theta_k^{p*}} = 1$ indicates that record k is at the proper production scale, the *most productive scale size* (MPSS). $\frac{\phi_k^{p*}}{\theta_k^{p*}} < 1$ reveals that k is scale inefficient, namely k is too large or too small (Banker et al. 1984). Smaller $\frac{\phi_k^{p*}}{\theta_k^{p*}}$ shows more differences from the correct size. Moreover, the optimal solution of (1) related to k, $\sigma_k = \sum_{j \in S^p} \lambda_j^*$, provides information on k regarding the relative scale position to the corresponding MPSS. $\sigma_k = 2$ suggests that the size of k is twice that of MPSS, and scaling down in production scale may increase the performance in productivity. Similarly, $\sigma_k = 0.8$ says that the size of k is only 80% of the correct size.

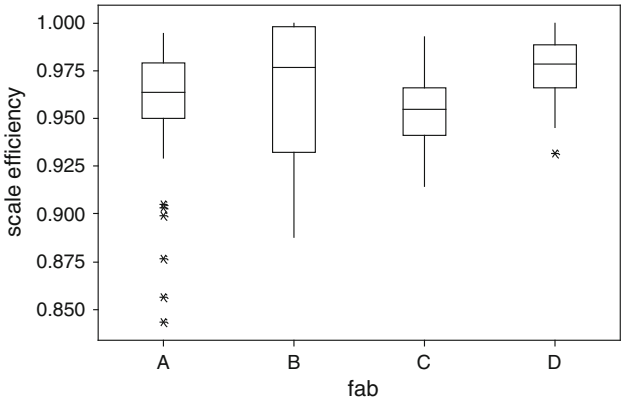

Fig. 9 Scale efficiency

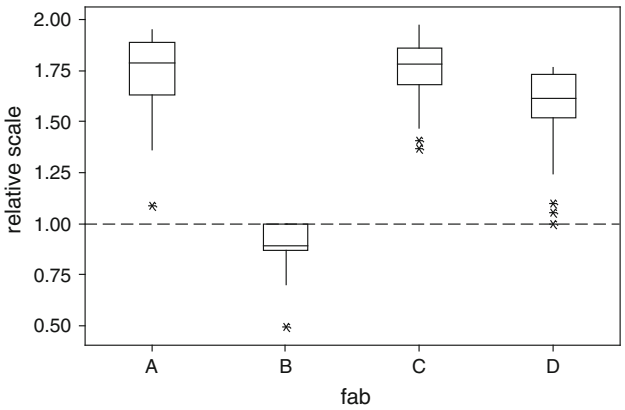

Fig. 10 Relative scale

Figures 9 and 10 are box plots for scale efficiency and the relative scale to the MPSS. Figure 9 shows that all records of fabs A and C are scale inefficient, unlike in MPSS. Fab B is only half the size with respect to MPSS, while fabs A and C can be twice as large as the MPSS (Fig. 10). In order to balance the fab utilization, resources can be transferred from a fab operating at decreasing return to scale to a fab operating at increasing return to scale.

Resource slacks

It is important to monitor the resource slacks to provide feedback on resource allocation. Model (7) further investigates the resource slacks, i.e., the exceeded amount of resources needed to provide the maximum outputs:

$$\begin{aligned}
&\max_{\lambda_j, s_i} \sum_{i \in I^p} s_i \\
&\text{s.t.} \sum_{j \in S^p} x_{ij}\lambda_j + s_i = x_{ik} \quad \forall i \in I^p, \\
&\quad \sum_{j \in S^p} y_{rj}\lambda_j \geq \theta_k^{p*} y_{rk} \quad \forall r \in O^p, \\
&\quad \sum_{j \in S^p} \lambda_j = 1, \\
&\quad \lambda_j \geq 0 \quad \forall j \in S^p; \qquad (7)
\end{aligned}$$

where θ_k^{p*} is the optimal solution from (6). Optimal solution s_i^* indicates exceeded unnecessary amount for resource i, namely record k can produce the maximum outputs $\theta_k^{p*} y_{rk}$ $r \in O^p$ without this amount of resource. We note that (7) is not unit invariant; different optimal solutions will be obtained with scaling on different resources. However, we can gain useful insights about resource allocation and management by observing the trends from the results.

Figure 11 shows the changes of resource slacks over time for fab A. y-axis is the percentage of slack for a particular resource of the record under evaluation, $\frac{s_i^*}{x_{ik}}$. Squares represent slack percentages for headcount; triangles, circles and crosses represent slack percentages for equipments, space and time. We find that equipment slacks are significant among four resources. Fab A gets more headcount slacks as time moves. Most of time, space and time slacks are <5%. In particular only three of the 39 time periods show time slacks, suggesting that time is the bottleneck resource for fab A in most situations.

Figures 12, 13 and 14 have the same interpretation as Fig. 11, but represent slacks for fabs B, C and D, respectively. Fab B has the least slacks among the four fabs; there are no slacks on space and time (Fig. 12). However, headcount and equipment slacks start to exist with vibrations as time passes. This is a negative sign for fab B's headcount and equipment allocation and usage. Fab C is the worst in terms of slacks among the four fabs, but Fig. 13 also shows significant improvements for fab C. There are improvements in resource slacks for fab D (Fig. 14); the space slacks decrease over time from >15 to 0% in the last four periods.

Comparing Figs. 11, 12, 13 and 14 shows no slacks for time. This observation may suggest that time is typically the bottleneck resource and is intensively controlled and monitored, although we cannot identify the causes and/or effects. We also note that two fabs improving their exceeded resource

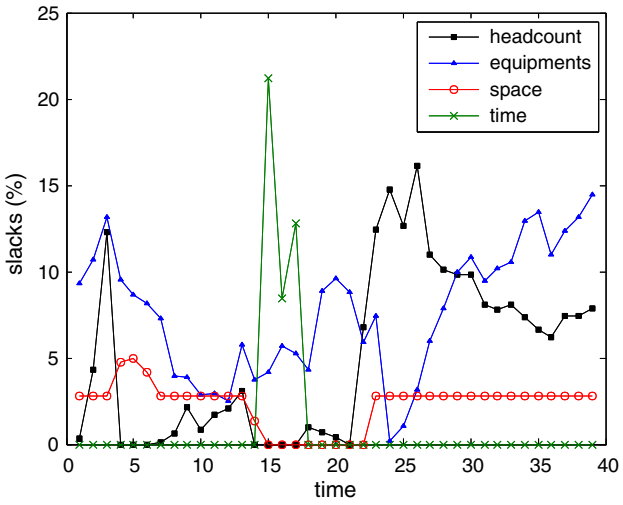

Fig. 11 Change of resource slacks (Fab A)

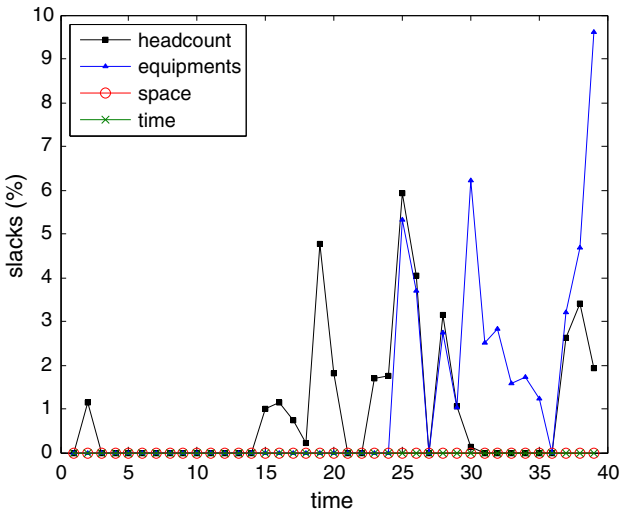

Fig. 12 Change of resource slacks (Fab B)

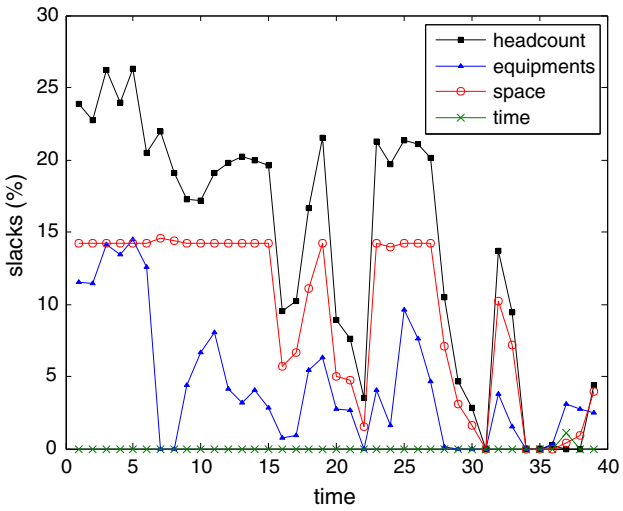

Fig. 13 Change of resource slacks (Fab C)

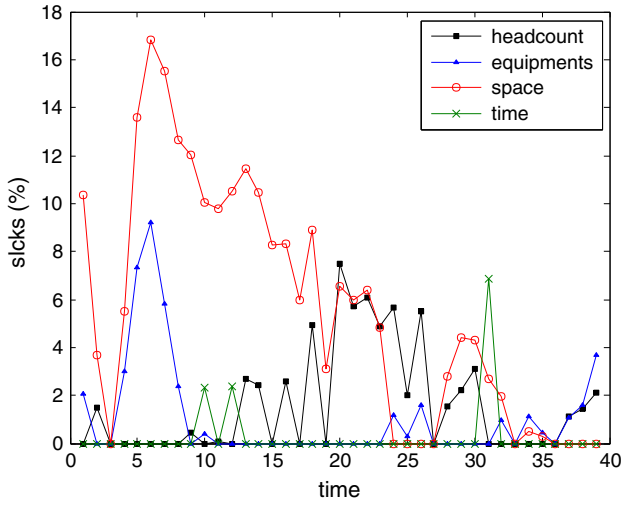

Fig. 14 Change of resource slacks (Fab D)

level while the other two show declining performance may signal an important message for future improvement.

This case study demonstrates how a proper (and/or inappropriate) performance analysis influences the decisions, e.g., resource allocation and management. The results suggest that, first, productivity indices using wafer as output are biased. The wafer-based performance should be decomposed as productivity and manufacturability. Wafer-based metrics may mislead the adjustments in response to the critics. The ultimate evaluation metrics for individual fabs should be carefully defined. Second, the scale effect "bigger is better" does not always exist. Third, exceeded resource usage must be controlled. Typically more than 10% slacks are for physical resources such as headcount and equipments; however, time is the largest binding constraint.

Conclusion

Performance evaluation and analysis are important for maintaining competitive edges and survival in the globally competitive semiconductor industry. This paper is motivated by a real case in Taiwan to evaluate operations performance by comparing various fabs within the same firm. The intra-firm competitive performance comparison has always been controversial. We present a two-stage fabrication process model to provide a reasoning platform to discuss, organize, and explain performance metrics. The model suggests that conventional wafer-based metrics are biased for operations performance since it includes the manufacturability aspect that is not the responsibility for fab operations. The case study provides evidence to support our suggestion for evaluating fab operations. The empirical results show that wafer-based performance is mainly contributed by productivity while manufacturability plays a role of multiplier that results in masking effects that may cause misinterpretation of the real problem. The results also suggest that exceeded resource usage must be properly controlled, especially since there are typically more than 10% slacks for physical resources such as headcount and equipments. Surprisingly, "bigger is better" is not true in our case, which suggests further investigation of scale decisions. The proposed model with DEA provides an overall performance measure through relative comparisons. In addition, we present its strong linkages to various, widely used productivity ratios, and offer a mechanism to promote its adoption in practice.

Acknowledgments This research is partially supported by National Science Council, Taiwan (NSC97-2221-E-007-111-MY3 and NSC97-2221-E-009-113), Taiwan Semiconductor Manufacturing Company (96A0279J8), and the Faulty Empowerment Award of National Tsing Hua University (97N2521E1) from Ministry of Education, Taiwan. The authors appreciate the invaluable comments from anonymous reviewers and the Guest Editors Professor Mitsuo Gen and Professor Hark Hwang.

References

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*, 1078–1092.

Beamon, B. M. (1999). Measuring supply chain performance. *International Journal of Operations and Production Management, 19*(3), 275–292.

Bhargava, M., Dubelaar, C., & Ramaswami, S. (1994). Reconciling diverse measures of performance: A conceptual framework and test of a methodology. *Journal of Business Research, 31*, 235–246.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444.

Chen, W.-C., & Hong, I.-H. (2008). Selecting an e-scrap reverse production system design considering multi-criteria and uncertainty. *IEEE Transactions on Electronics Packaging Manufacturing, 30*(4), 326–332.

Chien, C.-F., Chen, H.-K., Wu, J.-Z., & Hu, C.-H. (2007). Constructing the OGE for promoting tool group productivity in semiconductor manufacturing. *International Journal of Production Research, 45*(3), 509–524.

Cooper, W. W., Seiford, L. M., & Tone, K (2000). *Data envelopment analysis: A comprehensive text with models, applications, references, and DEA-solver software*. Berlin: Springer.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research, 132*, 245–259.

Golany, B., & Roll, Y. (1989). An application procedure for DEA. *OMEGA, 17*(3), 237–250.

Huang, S. H., Dismukes, J., Shi, J., Su, Q., Razzak, M. A., Bodhale, B., & Robinson, E. (2003). Manufacturing productivity improvement using effectiveness metrics and simulation analysis. *International Journal of Production Research, 41*(3), 513–527.

Jeong, K., & Philips, D. T. (2001). Operational efficiency and effectiveness measurement. *International Journal of Operations & Production Management, 21*(11), 1404–1416.

Leachman, R. C., & Hodges, D. A. (1996). Benchmarking semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing, 9*(2), 158–169.

Leachman, R. C., Ding, S., & Chien, C.-F. (2007). Economic efficiency analysis of wafer fabrication. *IEEE Transactions on Automation Science and Engineering, 4*(4), 501–512.

Muchiri, P., & Pintelon, L. (2008). Performance measurement using overall equipment effectiveness (OEE): Literature review and practical application discussion. *International Journal of Production Research, 46*(13), 3517–3535.

SEMI E79-0200. (2000). *Standard for definition and measurement of equipment productivity*. Mountain View, CA: Semiconductor Equipment and Materials International.