Norwegian University of
Science and Technology

Technical Efficiency and Productivity in Incentive Systems

Gaute Bjørklund Gundersen

Industrial Economics and Technology Management
Submission date: June 2011
Supervisor: Einar Belsom, IØT

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

NTNU
Det skapende universitet

MASTERKONTRAKT
- uttak av masteroppgave

1. Studentens personalia

Etternavn, fornavn **Gundersen, Gaute Bjørklund**	Fødselsdato **23. jul 1984**
E-post **gautebjo@stud.ntnu.no**	Telefon **47322438**

2. Studieopplysninger

Fakultet **Fakultet for Samfunnsvitenskap og teknologiledelse**	
Institutt **Institutt for industriell økonomi og teknologiledelse**	
Studieprogram **Industriell økonomi og teknologiledelse**	Hovedprofil **Anvendt økonomi og optimering**

3. Masteroppgave

Oppstartsdato **17. jan 2011**	Innleveringsfrist **13. jun 2011**
Oppgavens (foreløpige) tittel **Technical Efficiency and Productivity in Incentive Systems** **Evaluate the potential for using technical efficiency and productivity as performance measures in incentive systems.**	
Oppgavetekst/Problembeskrivelse 1. Review and discussion of theoretical and empirical literature related to performance measures in incentive systems. 2. Presentation and systematic evaluation of various measures including technical efficiency measures and productivity measures based on technology estimated with parametric or non-parametric methods. 3. Overall assessment of the applicability of the different measures in different organizational contexts.	
Hovedveileder ved institutt **Førsteamanuensis Einar Belsom**	Biveileder(e) ved institutt
Merknader **1 uke ekstra p.g.a påske.**	

4. Underskrift

Student: Jeg erklærer herved at jeg har satt meg inn i gjeldende bestemmelser for mastergradsstudiet og at jeg oppfyller kravene for adgang til å påbegynne oppgaven, herunder eventuelle praksiskrav.

Partene er gjort kjent med avtalens vilkår, samt kapitlene i studiehåndboken om generelle regler og aktuell studieplan for masterstudiet.

Trondheim, 04.04.2011
Sted og dato

Cantu Gundersen
Student

Einar Belsom
Hovedveileder

Originalen oppbevares på fakultetet. Kopi av avtalen sendes til instituttet og studenten.

Preface

This Master's thesis is my final project for receiving the Master of Science degree in Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU). The topic of the thesis is selected as a part of my specialization in Managerial Economics and Operations Research, focusing on aspects related to performance evaluation in incentive systems.

The title is "Technical Efficiency and Productivity in Incentive Systems" for which the purpose is to evaluate the potential of performance measures derived from estimates of production technology in incentives systems.

Although the underlying theory in this thesis is well-established, limited work has previously been done on coupling the theory together. This thesis is therefore regarded to be one of the first of its kind, and hopefully it will bring additional insight on performance evaluation under strategic behavior. In conversations with business representantives representing different kind of businesses, it seems like little attention is spent on understanding possible pitfalls of performance evaluation. For this reason I believe the content of this thesis is appealing for those with limited knowledge about performance evaluation, and that experienced readers will feel enriched as well.

The subject of this thesis was introduced to me by my supervisor, associate professor Einar Belsom, and I am grateful that he brought it to my attention. I am also grateful for the feedback he has provided through the writing process. By reading articles related to performance evaluation I have increased my understanding of the topic substantially. I further believe that the effort invested in this thesis will be fruitful for my later career. I will therefore take the opportunity to thank my supervisor yet again.

Gaute Bjørklund Gundersen
Trondheim, June 12, 2011

Summary

In this thesis technical efficiency and productivity are evaluated for use in performance evaluation in incentive systems. In light of agency theory and the organizational context, these techniques have several promising attributes for use in incentive schemes, despite their limited occurrence in the incentive literature. The use of data envelopment analysis (DEA) for estimating technical efficiency limit subjective evaluations and eliminate unwanted Nash-equilibrium under comparative evaluation. The Hicks-Moorsteen index prove to be the preferable index for measuring productivity change, as it cope with technologies exhibiting globally variable returns to scale. By coupling DEA and Hicks-Moorsteen we get four linear programs, which are easy to solve with developed software. However, infeasibility might occur when estimating the index and no remedies to this problem exist in the literature. Infeasibility will not occur for continuous time indexes or when estimating technical efficiency with Stochastic Frontier Analysis (SFA). However, SFA is poor on other aspects and software incorporating continuous time indexes are yet to be developed.

The use of productivity as a method for performance evaluation might offset systematic bias for comparative evaluation in heterogeneous environments, and will in most cases give employees strong incentives to improve. Technical efficiency might induce efficient employees to only maintain their level of effort, but super efficiency models reduce this threat. When computing technical efficiency, environmental factors should be adjusted for through a stepwise regression procedure in order to reduce uncontrollable risk. Although the goal is to implement a model that minimize subjective evaluations that might lead to favoritism, a final expert judgment should verify or disprove the performance scores.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The development of employee performance is central to the continuous improvement of all organizations, which might be facilitated through the implementation of an incentive system. A successful incentive system relies on a balanced relationship between performance and rewards, inducing the employees to increase their effort on performing tasks. But how will individuals or groups behave when they are evaluated and this evaluation affects how they are rewarded? How do we know that an evaluation correctly assesses overall performance?

Empirical research has shown that biased and inaccurate performance evaluation reduces productivity by reducing the effectiveness of incentives in the organization (Baker et al. 1988), which in turn reduces organizational performance. As an unbiased and accurate performance evaluation technique is of key importance for a functional incentive system, this thesis is dedicated to an explicit evaluation of different evaluation techniques. In academic and empirical research some attention has been brought upon the applicability of techniques such as Balanced Scorecard and subjective assessments, but evaluations of measures derived from estimates of production technology seems to be missing. Because the latter class of measures are widely accepted and supported in the field of operations research and only has received limited attention in incentive systems, such measures will be the main focus in this thesis. The ultimate objective is to identify the most promising evaluation scheme to be incorporated in a variety of organizations — encouraging optimal decision making while limiting the effects of inaccuracy and bias.

1.2 Analytical Framework

In the following, an analytical framework for evaluating the applicability of performance measures in incentive systems will be established. An incentive system can be thought of as a scheme which communicates strategy, motivates employees, and reinforces achievement of organizational goals. In order to conceptualize this scheme, an intuitive representation is illustrated in figure 1.1. This cycle depicts

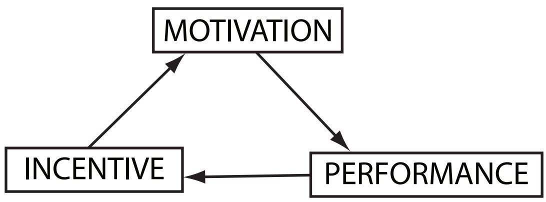

Figure 1.1: The incentive system: A schematic representation of the ideal world

the relationship between incentives, motivation and performance, but is limited in a practical sense, as it reflects an ideal world. In this ideal world we believe that a given level of motivation is perfectly converted to the employees effort on fulfilling his designated tasks. These tasks are believed to be optimally set up to capture the goals of the organization, and are perfectly measurable by a performance evaluation scheme. The performance score received is then optimally aligned with a justified incentive which stimulates the employees' motivation to increase effort.

In a real world application this simplistic representation fail to capture important aspects and a more realistic representation is suggested in figure 1.2. As opposed to the ideal world representation, motivation is believed to stimulate effort spent on gaming and effort spent on fulfilling tasks. Gaming will be covered in section 2.2, and point to the situation where an employee seek to maximize a performance score while minimizing effort. This situation is likely to occur if the performance evaluation is not aligned with the goals of the organization, and a gap exists between real and measured performance. Even though a performance evaluation

Figure 1.2: The incentive system: A schematic representation of the real world

distinguishes between gaming and effort, it might fail to capture the true performance of the employee. First, poorly communicated evaluation criteria might induce the employee to waste effort on tasks not contributing to value creation, even without gaming the system. Second, the performance of an employee might be affected by the states of nature. Nature reflects the environment's impact on performance, which can be both negative or positive, systematic or unsystematic. If not adjusted for, this might lead to an over- or underestimation of the performance of the employee, so that the incentive rewarded is unjust. Finally, the incentives provided do not necessarily lead to increased motivation.

The representation in figure 1.2 is more complex than figure 1.1 and reveals important relationships and pitfalls in the incentive system. These relationships will be investigated in further detail, except the link between performance and incentives and incentives and motivation.

A manager should be aware of the difference between intrinsic and extrinsic motivation, denoting the motivation for a task itself and the motivation from a separable outcome respectively. Intrinsic motivation is considered to be more powerful and is more likely to result in a desired outcome than extrinsic reward (Bates 1979; Motivation for learning), but it might still be corrupted by extrinsic rewards. In a discussion paper by Courty et al. (2008) the authors treat how the intrinsic motivation of employees in a Job Training Partnership Act (JTPA) program was corrupted by introducing pay for performance. While the employees initially found pleasure in providing jobs for the most needed, the introduction of an incentive program lead to cream-skimming where jobs were mainly given to those who were most likely to succeed. It is clear that considerable attention should be brought to the relationship beteween motivation and incentives as well as the effect of different incentives. The interested reader should look up an article by Clark and Wilson (1961) on a treatment on monetary and non-monetary incentives.

1.3 Outline

The conceptualized incentive system will form the analytical framework for evaluating performance measures. In chapter 2, the fundamentals of performance evaluation are treated, where section 2.1 regards performance evaluation in an organizational context and section 2.2 treat strategic responses and risk. The key observations from chapter 2 will provide the basis for the establishment of favorable properties which a performance evaluation technique should possess. These properties are summarized in section 2.3.2, where explicit criteria are selected and will be used to evaluate different techniques. Chapter 3 contains explicit evaluations of common techniques for performance evaluation in incentive systems, following with an evaluation of technical efficiency and productivity in chapter 4. Chapter 4 is introduced with a formal theoretical introduction to productivity and technical efficiency, including an evaluation of the conceptual difference between measuring efficiency and productivity change. In section 4.3 and 4.4 specific techniques for estimating technical efficiency and productivity will be treated respectively. Chapter 5 combine the virtues of chapter 3 and 4 in a simultaneous evaluation and a scheme for performance evaluation is proposed. Chapter 6 summarizes and concludes the past chapters as well as suggesting further work.

Chapter 2

The Fundamentals of Performance Evaluation

2.1 The Organizational Context

In the following section the focus is directed towards the link between organizational objectives and performance evaluation as depicted in figure 1.2. When evaluating performance in an organizational context one should pay attention to the employee's effort spent on fulfilling value adding actions, that conform with organizational goals. In this section, possible strategic choices to performance evaluation are neglected and employees are regarded as machines, only performing actions that are evaluated.

2.1.1 Capturing Value Adding Actions

In an incentive scheme we seek to measure and reward value adding actions, striving towards business excellence. This can be conceptualized by adapting the notation used by Baker (2002) and investigate the relationship between measured performance and created value. If V denotes the value of the organization and PM the performance measurement, the following formulation represent the performance evaluation problem:

$$V(a, \epsilon) = f * a + \epsilon$$
$$\text{and}$$
$$PM(a, \phi) = g * a + \phi$$
$$\text{where}$$

V : firm value

a : n-dimensional vector of actions an employee can take $\{a_1, a_2, \cdots, a_n\}$

f : n-dimensional vector of marginal product of actions in a with negative or positive contribution $\{f_1, f_2, \cdots, f_n\}$

g : n-dimensional vector of marginal product of actions in a on the performance measure $\{g_1, g_2, \cdots, g_n\}$

ϵ : random effects influencing the value of the firm

ϕ : random effects influencing the performance measure

The stronger correlation between V and PM, the better is our performance measure in capturing value adding actions. The random effects ϵ and θ might be correlated through common causes, both positively and negatively. For instance, let's assume that we are evaluating the performance of car salesmen and that there has been a recent increase in oil prices. This is likely to lower the expected sale of petrol cars while increasing the sale of hybrid cars. If salesmen were evaluated by the volume of sold cars, the performance of a hybrid car salesman would increase while the performance of a petrol car salesman would decline. The value of the entire firm is influenced by the expected value of future sales and would increase if the increased sale of hybrid cars surpassed the decline in sale of petrol cars. If it did, ϵ and θ would be positively correlated for the hybrid car salesman and negatively correlated for the petrol car salesman.

The correlation between V and PM will increase as the vector g is more aligned with f, or that $\text{corr}(\Delta V, g\Delta a) > 0$. Otherwise we could experience that the increase in performance actually leads to a decrease in value. This special case leads to another implication of performance evaluation: the fact that the performance of one unit might negatively affect the performance of another unit[1]. Therefore we have to evaluate if we should assess the performance of a group of employees or individuals. The dimension of time also complicates the relation between V and PM, as PM often measures past or present performance, while V more often evaluates future performance.

Another useful and perhaps more intuitive representation of the problem of capturing value adding actions can be made through the Venn diagram in figure 2.1 combined with set theory. By using the same notation as earlier, we can see that

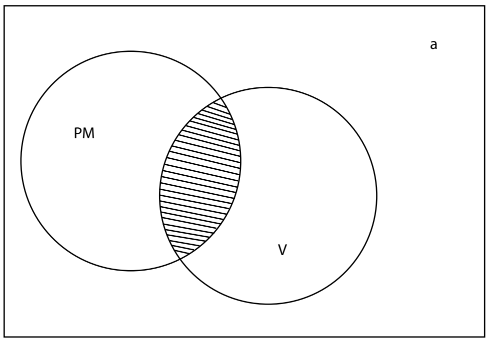

Figure 2.1: A venn diagram depicting the relation between PM and V

actions a denote the entire sample space while PM and V are subsets. The intersection between PM and V ($PM \cap V$) represents actions that are both adding actions and are captured by the performance measurement, which we seek to maximize. The area $PM \setminus V$ represents actions that are measured but that does not add any value to the organization. This area represents the worst case scenario where workers are encouraged to perform non-value adding actions, thus lowering the value of the organization. We will not pay much attention to this extreme case and rather focus on the Venn diagram in figure 2.2. Here PM is a subset of V, reflecting the case where we do not manage to capture all value adding actions.

Failing to capture important value adding actions might lead to sub-optimal decision making, thus failing to capture the goals of the organization. Piece-rate workers might for instance sacrifice quality for quantity in order to achieve a better performance score. In the Lincoln Electric Company, the performance of stenog-

[1]Such as market cannibalization or in value chains.

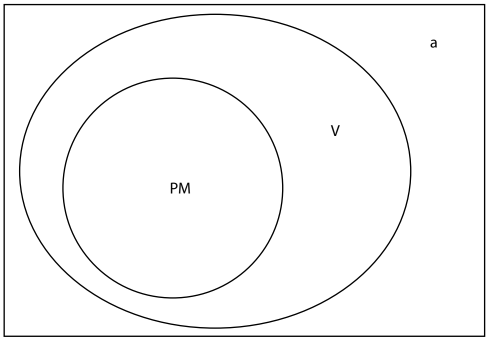

Figure 2.2: PM is a small subset of V

raphers was measured by counting the number of times the typewriter keys were operated. The company later discovered that one employee was earning much more than the others, and further investigation revealed that the worker ate her lunch at her desk, eating with one hand while punching the most convenient key as fast as she could (Berg and Fast 1975). Even though measuring the number of keys operated capture a desired action, the marginal production of g on PM was constant while the marginal production of f on V was decreasing. In this case it is likely that she was aware of the outcome of her actions, but in other cases it might not be so. For the sake of completeness we will therefore provide a framework for selecting appropriate measures ensuring organizational conformity. One final note is that many companies use stock prices performance for CEO evaluation, as stock prices should reflect unbiased estimates of firm value. However, this will impose a risk on the CEO as the stock price most probably will be affected by uncontrollable events.

2.1.2 Ensuring Organizational Conformity

As mentioned when introducing the incentive system, the implementation will not be successful unless the performance measurements are consistent with the objectives of the organization. Our ultimate goal when selecting measures, is to identify parameters that influence the value of the organization, favorable or unfavorable. This might be difficult in practice and sometimes it might be better identifying measures that conform with organizational goals. Folan and Browne

(2005) covered this topic in an excellent way, focusing on the importance of performance measurements as an integrated part of an organization's performance management. The book "Business performance measurement: theory and practice" by Neely (2002) is also a good source for answering this question. In order to establish parameters to evaluate the performance of employees, we can adopt the framework developed by Keegan et al. (1989). They separated the parameters in a two-by-two matrix with internal/external and cost/non-cost, which provides a balanced measurement of the performance of the employees. Kaplan and David (1992) extended this framework to capture subjective/objective and driver/outcome, establishing what we know as the Balanced Scorecard.

Driver versus outcome

While outcomes are observed products of an unit, drivers are the forces causing those outcomes. For example, customer satisfaction might be a driver for further sales and research and development might be a driver for innovation. It is therefore important to consider both drivers and outcomes when selecting measures for performance evaluation, as this will capture both causes and effects and balance long-run versus short-run performance.

Long-run versus short-run

If an organization seeks long-run rather than sort-run success, performance measures should reflect this goal. Even though an organization might have short run obligations which should be met, failing to focus on long-run indicators might put a premature end to its success. Long-run and short-run measures are highly correlated to drivers and outcomes respectively.

Internal versus external

While internal measures reflect operational performance, external measures reflect strategic performance. Because decision making on an operational level is considered short-run and strategic decisions are considered as long-run, these dimensions are related. The difference is the explicit formulation of external measures, which are focused against external stakeholders.

Subjective versus objective

While objective measures are easier to quantify and use in a performance evaluation, these measures might be easier to distort and their data might also be noisy. Subjective measures might offset distortion and risk through ex post evaluation, but are more difficult to quantify. The use of subjective measurements might also lead to favoritism resulting in a biased performance measure.

Financial versus non-financial
In an article by Ittner and Larcker (2000) the authors list advantages and disadvantages of non-financial over financial measures. They state that non-financial measures are more aligned with long-term organizational strategies and that financial metrics generally focus on short-term performance. By measuring both financial and non-financial performance, managers are provided incentives to focus on long-term strategy and intangible assets, intellectual capital and customer loyalty, which are assumed to be better indicators of the future value of an organization. On the other hand, non-financial measures might be more difficult and time-consuming to evaluate and might lack a common denominator. Non-financial metrics should therefore be organizational specific and be chosen in a dynamic process as strategies and competitive environments evolve.

2.1.3 Dynamic and Static Performance Evaluation

Apart from choosing the appropriate measures, one should determine whether performance should be evaluated status quo or by examining the development from one period to the other. But how will this choice affect the performance of an employee and what factors might influence the estimation of performance? These question will be repeated when productivity change and technical efficiency are compared in section 4.2, and will for now be treated on a conceptual level.

Evaluating performances status quo might also be denoted "temporal evaluation" or "static evaluation". The objective is to compare a unit against a benchmark or against other units in the same period, thus neglecting the dimension of time. Dynamic evaluation seeks on the other hand to measure change in performance, where a unit is measured against its performance in another period or how it has developed compared to its peers. Dynamic evaluation might also go by the term "trend study", where we try to spot the pattern in development. From an operational point of view, static performance evaluation is good for finding best practice and help in eliminating inefficient operations. From a strategic point of view, dynamic performance evaluation is more suited for capturing the development in performance and provides a better understanding of how an organization evolves.

Although dynamic performance evaluation is more attractive for capturing development in performance, it is more likely to be biased by inaccurate and inconsistent measures than static evaluation as more parameters are involved. A dynamic measure is also often constructed by two or more static measures, so that the variability is expected to be at least as large as for static measures. By decomposing evaluated performance into components which affects the calculation, possible sources to variability in cross-sectional and panel data are identified in figure 2.3.

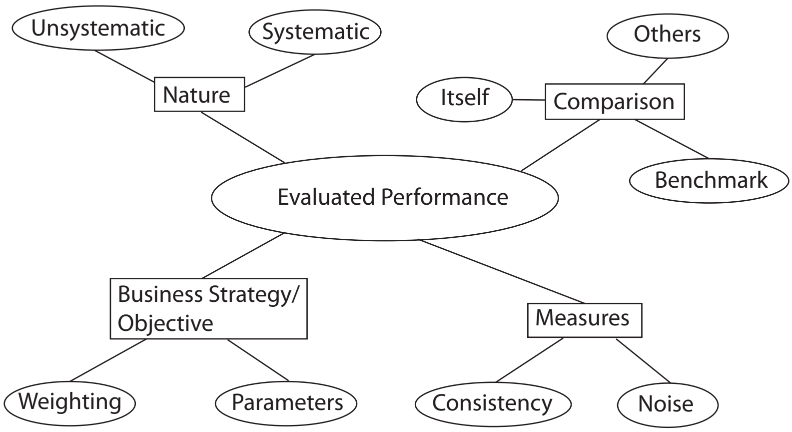

Figure 2.3: Aspects influencing performance evaluation

Some of these aspects are fixed for static performance evaluation, such as "business strategy/objective" and "comparison". "Business strategy/objective" reflect which parameters that are selected and how they are combined in an overall assessment. "Comparison" is how we anchor a performance in order to rank how good the performance is. For fixed evaluation this might either be compared to units in the same period or a benchmark or reference technology. In a dynamic performance evaluation, all aspects might change from one period to the other. Business strategy might have changed, so that other measures and weights are selected. Systematic and unsystematic influences of nature regard market fluctuations and systematic heterogeneity, and stochastic states of the operational environment respectively. E.g., a company evaluating the performances of different local stores may experience that seasonable market fluctuations have a large impact on trend analysis. Systematic heterogeneity reflects differences between operational environments, where some stores might operate under more stringent conditions than others. Unsystematic influence of nature and measurement noise will affect all units, both static and dynamic, but is likely to have a larger impact on dynamic evaluation when multiple static measures are combined. Measurement consistency points to consistency between units and periods, where evaluated performance might change between periods simply because we have adapted other methods for reporting data.

From the observations above, we might suspect performance evaluation based on dynamic evaluation to be more biased than those based on static evaluation. While this seems like an indisputable fact, dynamic measures might distinguishing systematic from unsystematic influences of nature, thus being capable of detecting heterogeneity between units. This is an important observation, as heterogeneity might cause performance scores to be over- or underestimated under peer evaluation. For instance, socio-economic factors that are geographically contingent might cause bank branches with a particular kind of customers to perform worse than others. This is not captured by static evaluations, but by using dynamic measures the effect of socio-economic factors on performance scores will diminish since a unit is measured relative to itself. Additional implications of dynamic and static measures will be treated in section 4.2. A final summary on the difference between dynamic and static measures is that static measures do not adjust for sources to bias and inaccuracy implicit, while dynamic measures might if bias is systematic. Dynamic measures might on the other hand introduce additional bias if the operational environment has changed from one period to another. Therefore, it seems wise to adjust static measures for the components in figure 2.3, while dynamic measures should be adjusted for fluctuations between periods and changes in business strategies.

2.1.4 Implementation Properties of Performance Evaluation

Even if the optimal parameters are selected ensuring organizational conformity while perfectly capturing value adding actions, the incentive system might fail if the implementation is poor. User acceptance and monitoring are amongst critical factors for success, in combination with top management support (Pinto and Slevin 1987). In order to satisfy these critical factors, one should stress the importance of an attractive and credible performance evaluation scheme. Attractivity and credibility are often used as traits of a person that are likely to influence how a message is perceived by its receiver, but will in the following be adopted to traits of performance evaluation techniques.

Attractivity is traditionally linked to how the physical likability of a person affects how we perceive their message. When redefined in terms of an incentive system, attractivity is used as a collective term for attributes affecting user likability of an evaluation scheme. E.g. a technique for performance evaluation is likely to be regarded as attractive if the methodology is easy to communicate to both managers and employees. It is also important that the methodology is transparent for those who are evaluated, so that they have a clear understanding of how their performance is assessed. As low transparency might negatively affect the employees' commitment to the incentive system, this is an important factor to stress.

Credibility denotes a collective term for perceived procedural justice and correspondence with production theory. Whether or not an evaluation scheme conform to production theory might be difficult to determine, but point to the fact that a reward should be non-decreasing in increased effort. When the optimal parameters are selected, this is equivalent to an evaluation scheme where a performance score improve as the fraction of output to input increase. Procedural justice reflects the evaluated employee's perception of influences of subjectivity and favoritism, which might lead to biased estimates. A high perception of procedural justice will increase the organizational commitment and participation in the incentive system, as stressed by Folger and Konovsky (1989). Recent research also support these predictions where inequity in procedural justice is associated with job satisfaction (Tremblay et al. 2000), absenteeism and turnover. A technique should also be consistent between units and evaluation periods to yield credible results.

2.2 Strategic Responses and Risk

As depicted in figure 1.2, gaming in incentive systems might lead to biased performance scores and undesirable results for an organization. In 1956, Ridgway wrote about dysfunctional consequences of performance measurements, and an empirical investigation of gaming responses to explicit performance incentives was conducted by Courty and Marschke in 2004. Figure 1.2 also contain the exogenous variable "Nature", which imposes risks and heterogeneity to the performance evaluation. Section 2.2 will introduce agency theory, where these issues are reviewed in detail.

2.2.1 The Incentive Game

The incentive system will be considered as a game with two or more players: The principal(s) and the agent(s). The principal(s) can be considered as stakeholders representing the interest of an organization, while the employees are agents. We assume that the agents are bounded rational, opportunistic and effort averse (Simon 1976, Williamson 1985), so that they will minimize their effort while maximizing their utility to their best knowledge. At the same time, the principal seeks to maximize the value of the organization or its outputs[2] (Jensen and Meckling 1976).

Because the principal engages an agent to perform a task on their behalf by delegating decision making authority, it is likely that conflicting interests lead to suboptimal decisions (Jensen 1983). We typically assume that the agents have private information regarding their effort, unobservable to the principal, resulting in asymmetric information which creates a potential for moral hazard.

In order to mitigate this effect the principal can put effort into monitoring or compensating the agent to achieve desired behavior (Jensen and Meckling 1976). The latter instrument is known as optimal contracting, where the goal is to maximize the outcome for the principal. Performance can be signaled through actions or the result of these actions as captured by a performance evaluation. The principals objective is to balance behavioral-based pay and outcome-based pay without imposing to much risk and variability on the agent (Gibbons and Murphy 1990). The risk element becomes an important factor as the information about the nature is incomplete. Nature, as depicted in figure 1.2 refers to luck, ease of performing the task and other aspects that might skew the performance in a positive or negative direction. The nature imposes a risk on the agent, which we label "uncontrollable" if the agent cannot react to it and "controllable" otherwise (Gibbs et al. 2009).

The game described above can be recognized as a production game (Rasmusen 2007), which serves as a good starting point. In this game the monetary value of output is denoted by $q(e, \theta)$, increasing with the effort of the agent, e. $\theta \in R$ denotes the state of the world and is chosen by nature with assumed probability density $f(\theta)$. The agent's utility function, $U(e, w)$ is decreasing with effort and increasing with wage, w. In an output-based wage under uncertainty, the principal's

[2]The ultimate goal for a profit organization might be maximizing stock value. A non-profit organization might on the other hand seek to maximize the quality of their services or to maximize socio-economic welfare.

problem becomes:

$$\max_{w(\cdot)} EV\left((q(\tilde{e},\theta) - w(q(\tilde{e},\theta))\right) \tag{2.1a}$$

subject to

$$\tilde{e} = e_{argmax} EU\left(e, w(q(e,\theta))\right) \tag{2.1b}$$

$$\acute{U} \leq EU\left(\tilde{e}, w(q(\tilde{e},\theta))\right) \tag{2.1c}$$

The first constraint is the incentive compatibility constraint inducing the agent to choose desired effort, while the second is the participation constraint. These are included because an agent is free to reject participating in the incentive system and must be given an incentive to choose the desired effort.

While the principal is considered to be risk neutral, agents are most often thought of as being risk averse since it is assumed that they don't like a high variability in their compensation (Stiglitz 1987). This gives rise to the problem of balancing the agent's effort and risk aversion (Jensen and Meckling 1976), and agency theory states that the agents should be provided with some kind of insurance to accept risk. This insurance is often provided through base-pay. If the agent on the other hand had been risk seeking, he would accept a lower premium and would choose a mainly output-based wage. The presence of risk might induce the agent to place effort on less risky tasks in order to bring more certainty to their performance pay. By accounting for risk explicitly we might reduce its effects on performance evaluation so that we can lower the premium and make the agent less reluctant to execute risky tasks. For further treatment of risk in incentive systems, the reader is advised to look up an interesting paper written by Bloom and Milkovich (1997), covering theory and empirical results.

2.2.2 Asymmetric Information

When measuring the performance of employees, we should ask whether we should measure effort or outcome. This question might seem counterintuitive, as output rather than effort increase the value of an organization. While this is true for the organization as a whole, measuring output and neglecting effort gives the agent the favor of asymmetric information. As the agent's effort is unobservable by the principal, an opportunistic agent might choose to keep this information private and spend a minimum of effort to complete his tasks. This is especially the case if the performance measurements are not carefully selected, so that effort becomes measurable. If an agent is working as a car salesman and increase the number of sold hybrid cars simply because the price of petrol has gone up, it would not be efficient to provide an incentive. If this information is private to the agent and he benefits from receiving a higher bonus, we denote this moral hazard. In fact, if the agent had private information that indicated increased number of sold cars independent of his effort, the opportunistic agent could be induced to reduce his effort and still receive a bonus.

Issues regarding moral hazard are likely to increase as the monitoring of the agents effort become difficult, and with the use of absolute performance evaluation. In a scheme with absolute performance evaluation, the performance of one employee is not affected by others, as they are all measured against a benchmark. In the case of a multi-period game, the employees actually have an incentive to lower their effort hoping that the benchmark is lowered. If we on the other hand incorporate relative performance evaluation with multiple players, we get the classical prisoner's dilemma (figure 2.4). Clearly, the best option is to perform as good as possible,

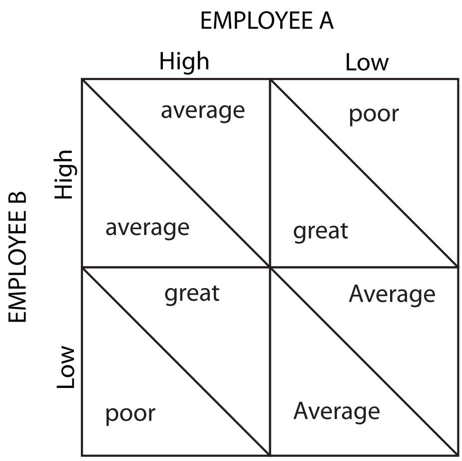

Figure 2.4: The pay-off matrix. Effort is given on the axis (High/Low) while pay-off is given inside the 2-by-2 matrix. Pay-off in L/L is higher than in H/H since the effort in H/H is higher than in L/L while the bonus remains the same.

as this will make the employee look good compared to the others. Unfortunately, measuring relative performance might affect knowledge-sharing and helping others in a negative direction, and we might get an unhealthy competition between the employees. The principal should therefore measure and emphasize knowledge-sharing in such a system (Lee and Ahn 2005). An important observation from figure 2.4 is that the effort spent in equilibrium H/H is higher than in L/L, so that collaboration between employees might offset the advantage of relative performance evaluation. Another issue with relative performance measurement is that the basis of comparison gets worse as the number of employees decrease. According to the law of large numbers, the sample mean of a large number of employees is more likely to reflect expected performance, than that of a small number of employees. With a small sample space, extreme performances and random outcomes may come more into play, causing outliers to have too much influence on peer evaluation. Increased monitoring of the employees is therefore a better, but more costly, solution as the number of employees decrease. Matsushima (2010) combined relative and absolute evaluation in intergroup competition and prove that this eliminate unwanted Nash equilibria. Note that the use of comparative evaluation undermines the agent's information advantage as the principal gets more information, thus lowering the information rent captured by the agents.

Moral hazard is normally at hand when one of the players have private information about the operational environment, and cooperate with a less informed part. While constructing a performance evaluation in the incentive system, this is actually equivalent to not modeling the impact of nature[3], visible to both principal and agent. A good performance evaluation should therefore be able to adjust for environmental factors as far as possible.

2.2.3 Distortion and Dysfunctional Responses

Even though we are aware of how moral hazard might corrupt the incentive system, it is difficult to completely eliminate the threats. The main treats while implementing a performance measurement, is that it might not conform with the goals of the organization and that an agent might manipulate the measurement. For instance, performance measures elicit dysfunctional and unintended responses because the employees, through their daily work, gain a superior understanding of how the measurement system works and how performance outcomes can be manipulated (Marschke and Courty 2003). If an piece-rate worker is only measured

[3] alternatively "macro shock"

by the number of outputs, he will have an incentive to decrease quality in order to increase speed. This behavior is denoted dysfunctional response. The organization can fight this behavior by introducing multiple measurements, but this will give raise to the multitasking problem where the agent have to split its effort over multiple tasks (Holmstrom and Milgrom 1991a;b). How the agent split its effort may be influenced by several factors, preferably by top management decisions. A risk-averse agent might on the other hand choose to focus on performing tasks with low uncontrollable risk.

To ensure that employees spread their effort, it is common to use some kind of aggregation where the principal put different weights on each of the measures. Placing optimal weights is difficult in practice and the employees are induced to lobby about the specification of the weights.

2.2.4 Agency Cost and Transaction Cost

Agency cost incur in any situation involving cooperative effort as a result of impersonal exchange. This cost was defined by Jensen and Meckling (1976) as the sum of:

1. the monitoring expenditures by the principal
2. the bonding expenditures by the agent
3. the residual loss

The term "agency cost" is often used interchangeably with "transaction cost", which occur under contracting (Williamson 1979) and is divided in ex ante and ex post costs. In our case, we can think of ex ante costs as costs related to getting an incentive scheme into place and communicate its applicability to its users. Ex post costs are on the other hand related to enforcing, monitoring and coordinating the incentive system.

Even though agency and transaction cost have slightly different applications, both terms point to the fact that collaborative effort give rise to additional costs. These costs do not bring any value to the organization and we can think of them as waste, which we seek to minimize. In the following, the term "transaction cost" will be designated all costs related to developing, implementing and enforcing an incentive system. As the focus of this thesis is performance evaluation, transaction costs will emphasize on costs related to measuring performance. Effort and time spent on specifying weights and handle lobbyists are two examples of transaction costs that are likely to occur.

2.3 Criteria for Performance Evaluation

As discussed so far in chapter 2, evaluating performance is not a straightforward task. We have reviewed how employees might behave when they are evaluated in an incentive system, and shed light on the objective of balancing multiple measures assessing overall performance. The past chapters have established a theoretical and empirical foundation for evaluating the applicability of different evaluation schemes, and in this section criteria for performance evaluation will be established. First, a short guide for selecting suitable measures will be given. Secondly, and more importantly in this thesis, favorable characteristics of a performance evaluation technique will be proposed.

2.3.1 Criteria for Selecting Suitable Measures

Since our intention is to evaluate the performance of a specific employee or group, the parameters should measure the result of actions carried out by those who are evaluated and not by some external event. The number of sold units might for example be influenced by fluctuations in demand, which impose an uncontrollable risk on the sales manager. We might not always be able to distinguish controllable from uncontrollable risk, but a careful selection of parameters might mitigate the effect of uncontrollable risk. It is also important to select measures that are less likely to be manipulated or distorted by the employee. We might argue that the performance of a teacher is reflected by the grades of students, but this might motivate the teacher to teach to the test or even alter the test scores. Monitoring the teachers in order to detect this dysfunctional behavior might provide satisfactory results, but this will increase the transaction cost. Finally, the parameters should be measurable so that they can easily be quantified. Low data noise and high consistency is also important for decreasing bias.

A summary of favorable properties:

- Employee or group specific

- Mitigate effect of uncontrollable risk
- Undistortable
- Measurable
- Low data noise and high consistency

2.3.2 Criteria for Techniques Estimating Performance

"Techniques" referes to systematic procedures to combine measures in order to evaluate the overall performance of an employee or team, relative or absolute. In order to evaluate the applicability of different techniques, important criteria will be established. These criteria are believed to be the key observations from chapter 2 and will be listed below with a short summary.

1. **Attractive**
   A technique is considered attractive if it is easy and intuitive to implement. High degree of transparency increases user acceptance. (Section 2.1.4)
2. **Credibility**
   Credibility reflects how believable or trustworthy the technique is thought to be. The degree of subjectivity and favoritism might introduce bias and affect the perception of fairness. A technique should also conform with production theory and yield consistent measures. (section 2.1.4)
3. **Balanced measurement**
   A technique should result in balanced measurement of employee performance. Both for capturing all value adding actions and mitigate the multitasking problem. (Sections 2.1.2 and 2.2.3)
4. **Non-corruptible**
   Non-corruptible techniques are robust against dysfunctional responses and other gaming responses. The technique should also distinguish controllable and uncontrollable risk, and adjust for heterogeneity if this is present. (Sections 2.2.2 and 2.2.3)
5. **Low transaction cost**
   Transaction cost is the sum of ex ante and ex post costs related to implementing and sustaining the technique. This include time spent communicating

the evaluation scheme and estimating the performance scores. (Section 2.2.4)

A technique satisfying all criteria is more likely to capture true performance and will ease the implementation and stimulate user acceptance.

Chapter 3

Common Practice in Incentive Systems

In this chapter, the most common techniques for performance evaluation in incentive systems are evaluated. These techniques were identified through examining numerous articles regarding performance appraisal in incentive systems[1], in addition to findings in a survey conducted by Nankervis and Compton (2006).

3.1 Subjective Performance Evaluation

Subjective performance evaluation is an evaluation technique where the overall performance of an employee is subjectively assessed by a manager. This evaluation often seeks to determine if a performance is under, at or over par, or in more general terms: "exceptional", "satisfactory" or "unsatisfactory". To determine which category a performance belongs to, explicit criteria and to what extent they should be met are established. These criteria ease the rather complex evaluation and serves as anchoring, which might be used for ordinal ranking of performances. Behaviourally Anchored Rating Scales (BARS) is one commonly used evaluation scheme which make use of this approach.

[1] Informal conversations with business representatives have also contributed to the identification of common practice.

Examples of use
Subjective performance evaluation is often used when evaluating non-financial and not easily quantified performance. This performance evaluation scheme is therefore often used when setting grades, and for supplementing decision making when promoting employees.

Pros and cons
Pure subjective evaluations can easily be biased and distorted. The "halo effect" is described as a cognitive bias whereby the perception of one trait is influenced by the perception of another trait, supported by Thorndike (1920) in his empirical research. This effect leads to implicit personality theory (IPA) (Stricker et al. 1974), which concerns the general expectations that we build about a person after we know something of their central traits. Subjective evaluation may also result in a compression of rating and rewards, as the supervisors often are reluctant to give poor ratings to subordinates (Prendergast and Topel 1993).

Subjective evaluations might on the other hand enable an evaluator to capture characteristics that would otherwise be difficult to account for. As organizations often are complex, the performance of employees is often too complex to only be accounted for by objective evaluations. Subjective evaluation might also adjust for noisy and inaccurate data in contrast to pure objective and deterministic models. A subjective evaluation is also rather intuitive to apply, although the technique becomes complex when evaluating multiple tasks.

Use in incentive system
Subjective evaluation is a relatively intuitive and easy evaluation scheme to implement. This yields high attractivity but at the same time low transparency as it relies on non-quantified and highly subjective data. The agent will probably more often than not argue that a score was unjust when receiving a low score, especially if he does not observe the actions of other agents or receive a proper feedback. The credibility of the technique will foremost rely on the expertise, devoutness and objectivity of the principal, as supported by Albright and Levy (1995). Empirical research has documented how the "halo effect" introduce bias and show that a principal's evaluation is highly corruptible[2]. In this evaluation scheme, the agent has an incentive to spend effort on ingratiating itself with the evaluator. This dysfunctional response will lower the credibility of the evaluation scheme and increase the transaction cost. The transaction cost will also increase with the level

[2]Other psychometric characteristics: leniency, interrater agreement and ratee discriminability.

of complexity, number of tasks to balance and with increasing number of principals collaborating in setting the score. One way to increase the transparency, and in turn credibility, is provide feedback to the evaluated employee but this is a trade-off to increased transaction costs.

If the evaluators manage to keep their independence and not be influenced by factors such as the "halo effect", subjective performance evaluation might offset effects related to uncontrollable risk. The evaluation might also offset dysfunctional responses, as the principal(s) might use their expertise to judge an outcome or action against the true objective of the organization. Because we are able to evaluate tasks that otherwise would be difficult to quantify, subjective evaluation might also provide a balanced measurement.

Subjective evaluation is preferable in complex organizations where the focus is to evaluate performance on long-run measures and drivers instead of quantifiable outcomes. In organizations with a high degree of quantifiable measures, subjective performance evaluation might foremost be used in implicit contracting (Baker et al. 1993) and for reinforcing objective performance measures.

3.2 Key Performance Indicator (KPI)

KPIs are multiple quantitative or qualitative measures, used by organizations to gauge or compare performance in terms of meeting their strategic and operational goals. This is done by comparing against an internal or external target to give an indication of performance. In order to assess the performance of an organization, it is common to construct multiple KPIs, each measuring a particular activity. For this reason KPIs goes by the term partial productivity measures. It seems to be no established framework for selecting the appropriate KPIs, and the selection is rather industry and firm specific. Common metrics are often expressed in ratios (e.g. output per employee, Cycle time ratio), relative (e.g. mean time to repair, average customer satisfaction) or absolute values (e.g. number of sold units, EBITDA[3]).

Examples of use
Every organization reporting their financial statement, quarterly or yearly, in order to compare performance, use KPI. Most organizations use KPIs in a wider extent

[3]Earnings before interest, taxes, depreciation and amortization.

as well, but might call it by other names. KPI is therefore regarded as the most common way to measure performance. For examples of KPI used in the construction and health care industry, I suggest reading articles by Cox and Ahrens (2003) and Sheldon (1998). See appendix A for examples of KPIs at different levels.

Pros and Cons

The main advantage with KPI is the ease of communication and implementation through its simplistic and intuitive representation. A KPI seldom include more than two different metrics, so that the result is transparent and understandable. KPIs are easy to calculate on a simple calculator, which make them reliable and yield a high attractivity and credibility. The downside to this simplistic measure is that it fails to provide a balanced measurement. As the organization become more complex and the number of KPIs increase, the multitasking problem grows increasingly. While KPIs capture each particular task in an excellent way, the technique fail to assess the overall performance of an employee.

Use in Incentive Systems

If KPIs are to be used in a incentive system, incentives will be provided on the basis of each KPI which might lead to gaming. For instance, in a pay for performance scheme with the measures quality and quantity, an employee might put all his effort into reaching the benchmark for quality. This will result in no effort spent on meeting the demanded quantity, but the employee receives a bonus nevertheless. The worker could be penalized by this dysfunctional response, but as soon as we do so, we are indirectly placing weights on the different KPIs which is not in the nature of this technique. If we on the other hand had published a list, displaying all KPIs, the agent would himself have an incentive to spread his effort[4]. The employee could then use the same level of effort to look mediocre on most KPIs, or choose to increase his effort and perform better overall. KPIs are not suitable to avoid gaming and to assess overall performance under tangible rewards, but might be satisfactory when intangible rewards are used. If the calculated KPIs are communicated to all employees, each employee will subjectively and perhaps unknowingly place weights on the KPIs to find his ranking. Employees who score the highest on all KPIs are easily identified as the most "successful", but those who only outperform their peers on certain KPIs might be ranked differently as a result of individual preferences[5].

[4] Assuming that the employee cares about what the other employees might think of him.

[5] An employee who outperform others on a certain KPI is likely to rank himself the highest as he might consider his field of interest as the most important.

KPIs yield high attractivity through high transparency and the ease of calculating and communicating the measures. KPIs are free from subjectivity if quantifiable data exist and are credible as they conform to production theory. The transaction cost is low, as KPIs are usually easy to estimate and the costs related to implementation and sustaining the technique is low. However, KPIs fail to provide a balanced measure and assessment of the overall performance of an employee, and are corruptible as they are absolute measures and do not adjust for risk and heterogeneity. By creating a composite measure of multiple KPIs, we will on the other hand achieve a more balanced measure and mitigate the multitasking problem.

3.3 Balanced Scorecard and Composite Measurements

Balance Scorecard (BSC) was introduced in 1992 by Kaplan and David, and came as a solution to how companies could improve their management of their intangible assets. The original thought was to use financial metrics as the ultimate measures for company success, and supplement with the three additional perspectives: customer, internal process and learning and growth. This thought was inspired by Lewis (1955) who based divisional performance on one financial and seven non-financial metrics. Balance Scorecard did not provide any new insight on balancing measurements, but served as a framework that put everything together. In this thesis we will think of Balance Scorecard as a framework for including all relevant KPIs in a composite measure, and that other similar methods for overall assessment are deviations from this framework.

In order to express employee performance as one single score, we have to create composite measurements through explicit weighting. The weights placed on each single measure ought to reflect the relative importance of that measure. Put in other words, the greater the importance, the greater the weight. Determining the optimal weights is perhaps the most difficult task after selecting appropriate measures or parameters, and is often carried out by subjective evaluations. The weights can also be determined by the Delphi method (Dalkey and Helmer 1963), in order to reach group consensus. The final score then becomes the summed product of all KPIs and their relative weight. The weights placed on each measure can either be fixed or flexible. Fixed weighting is often used in formula-based plans, where the weighting is held constant under the final performance evaluation. Flexible weighting allows the weights to be adjusted when calculating the final performance score, which increases the subjectivity. In an excellent article by Ittner et al. (2003), the authors treat how weighting of different kind of measures should be conducted. In the same article, the authors also evaluate the use of flexible weighting in a Balanced Scorecard based incentive system.

Examples of Use

Balanced Scorecard is perhaps the most popular framework for assessing overall performance of an employee or to aid decision making when multiple aspects are treated. In 12 annual surveys (1996-2008) conducted by Bain & Company with 9,933 respondents from more than 70 countries, they report the total usage and overall satisfaction of Balanced Scorecard:

Figure 3.1: Total usage and overall satisfaction of Balanced Scorecard (Bain & Company; 2011)

Pros and Cons

The Balanced Scorecard coupled with composite measurements manage to balance financial and non-financial metrics, short run and long run. When the weights are fixed, this approach yields an intuitive and transparent representation of performance which is easy to communicate to its users. The score is easy to calculate when the weights are fixed and the framework is widely accepted for balancing multiple measures. The downside of this approach is the subjective method for placing weights, which require great skill and experience. Mispecified weights might be adjusted by flexible weighting.

Use in Incentive Systems

Balanced scorecard ensures that effort is spread out on tasks incorporated in the scorecard by placing weights on financial and non-financial metrics. The balanced measurements yield an overall assessment of performance evaluation, which mitigate dysfunctional responses. To determine whether or not an employee should receive an incentive for his performance, his score must be compared against a set (absolute) benchmark or relative to other employees.

The specification of weights is the most difficult task after selecting appropriate measures, and in a real world application agents have incentives to lobbying about the weights. Agents might criticize the selection of weights as they seek to maximize the weight on the task which they perform the best. The use of flexible weights might adjust for uncontrollable risk, but will at the same time increase subjectivity and might lead to favoritism and bias. In fact, in a paper by Kaplan et al. (2007), based on empirical research, the author notes that "subordinate likability influences evaluators' judgments even when the performance measurement instrument is structured using the BSC" [p. 107].

Even though BSC initially yields high attractivity, the extended use of subjective evaluations lower its transparency and attractivity. In a paper by Ittner et al. (2003) the authors conducts a survey amongst 572 North American managers, and find that 14,7% of the respondents consider BSC as a "black box". In terms of credibility Ittner et al. found that 12 % felt that favoritism and bias came too much in play in BSC. While composite measures are less corruptible than KPIs, as they mitigate the multitasking problem by balancing multiple parameters, the subjective weighting seems suboptimal. Composite measures and BSC score average on transaction cost, as the specification of weights is likely to initiate time-consuming discussions.

3.4 Summary of Common Practice

Balanced Scorecard seems to be more suitable than KPIs in most cases, as the methodical framework seeks a balanced performance evaluation. BCS is also less corruptible than KPIs as KPIs are unable to cope with the multitasking problem. Balanced Scorecard and composite measures are also more suitable than subjective evaluations if all important parameters are measurable and if we seek a pure deterministic method. Balanced Scorecard has on the other hand been criticized for its rather subjective methodology for placing weights on the parameters and might be difficult to incorporate in complex organizations. Subjective evaluations might be more appropriate in complex environmnets, as a performance evaluation solely based on objective measures is likely to be incomplete. Subjective evaluations are also more suitable for offsetting bias related to measurement errors and gaming as ex post evaluations may be used to identify dysfunctional responses.

A short summary is displayed in table 3.4, ranking each techniques fulfillment of the criteria in section 2.3.2. The asterisk indicates that a score is highly dependent on the principal, while the respective score is thought to be the most likely.

	Atr.	Cred.	Balanced	Non-corruptible	Transaction cost
Subjective	2	2*	3*	2*	1
KPI	3	2*	1	1	3
BCS, CM	3	2*	3	2	2

*Table 3.1: The higher score (3-1), the better. * Relies on the principal*

Productivity and technical efficiency are little exploited as performance evaluation techniques in incentive systems. Empirical applications seem to be lacking and only a few articles have been found with a direct link between technical efficiency and incentive systems. This is quite surprising giving the fact that the literature on technical efficiency and productivity is extensive and well-established in the field of operations research. One might wonder why these techniques are so little exploited in performance evaluation under strategic behavior, and if technical efficiency and productivity are unsuitable or simply neglected.

The remaining chapters of this thesis will therefore be designated an explicit evaluation of the potential for using productivity and technical efficiency as techniques for evaluating performance in incentive systems. To my knowledge, no previous work has been done on providing a simultaneous evaluation of different techniques for evaluating the performance of agents.

Chapter 4

Technical Efficiency and Productivity

4.1 Theoretical Introduction To Efficiency and Productivity

Section 4.1 will be devoted to a theoretical introduction to technical efficiency and productivity as the techniques are somewhat intertwined. The introduction will provide a basis for treating technical efficiency and productivity in detail in section 4.3 and 4.4. When covering productivity and technical efficiency, I will adapt the set representation following the notation of Färe and Primont (1995), apart from substituting q for y. This provides a compact and consistent formulation and will be applied for all techniques.

Performance evaluation related to productivity and technical efficiency are rooted in production theory, in which we consider decision making units (DMUs) as entities transforming inputs into outputs. A DMU's input-output vector is denoted production mix, while all DMUs under evaluation can be represented by the technology set, also named production possibility set:

$$S = \{\ (x, y) : x \text{ can produce } y\ \}$$

where

x :$n * 1$ input vector of non-negative real numbers

y :$m * 1$ output vector of non-negative real numbers

This set contains all feasible input-output vectors representing the underlying technology for a given period. We may equivalently define the technology by the output set, $P(x)$, denoting all output vectors technological feasible using input vector x:

$$P(x) = \{\ y : (x, y) \in S\ \}$$

or input set $L(y)$:

$$L(y) = \{\ x : (x, y) \in S\ \}$$

The input and output sets are assumed to satisfy certain axioms, such as the property of convexity[1].

4.1.1 Distance Functions

The output set $P(x)$ is a basis for constructing a production possibility frontier (PPF), representing various combinations feasible at a given input level. The input counterpart to the PPF is the isoquant. If we assume that one input, x_1, is needed to produce two outputs, y_1 and y_2, we can construct a two dimensional PPF. This construct is depicted in figure (4.1), which also illustrate the effect of technical change by an outward shift of the PPF. As we can see from this figure, the new technology will able us to produce a higher level of outputs for the same level of input. The outward shift of the blue PPF depicted in figure (4.1) is named neutral technical change, as the ratio of y_1 and y_2 is unchanged. The red PPF is a non-neutral technical change which skew the curve, so that the marginal production rate of one of the goods has increased relative to the other.

The construction of a PPF gives us the foundation for describing output and input distance functions. The two latter terms were introduced independently by

[1] See Battese et al. (2005) for a full summary.

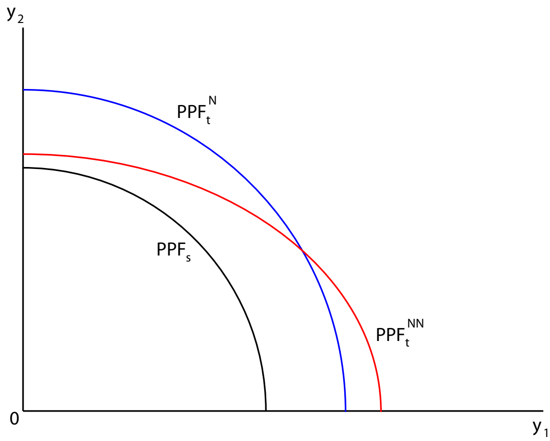

Figure 4.1: PPF_t^N depicts a neutral technical change, while PPF_t^{NN} depicts non-neutral technical change. PPFs is PPF for base period, s.

Malmquist (1953) and Shephard (1953), which enable a description of a multi-input or multi-output production technology without specifying any behavioral objective (maximize profit or minimize cost). The output distance function is defined on the output set, $P(x)$ as:

$$d_o(x, y) = \inf_{\sigma} \ \{ \ \sigma : (y/\sigma) \in P(x) \ \} \tag{4.1}$$

where σ is a scalar denoting the minimal radial expansion of the output vector to reach the frontier. Likewise, the input distance function is given in equation 4.2, ρ being the maximal radial contraction of x to reach the frontier.

$$d_i(x, y) = \sup_{\rho} \ \{ \ \rho : (x/\rho) \in L(y) \ \} \tag{4.2}$$

Under constant returns to scale it holds that $d_o(x, y) = [d_i(x, y)]^{-1}$.

The distance functions will be used for defining index numbers and to introduce technical efficiency. For further treatment and discussion of distance functions, the reader is advised to read articles by Färe and Primont (1995) and Russell (1998).

4.1.2 Technical And Allocative Efficiency

Based on the input distance function assuming constant returns-to-scale (crs), I will introduce the work of Farrell (1957). Based on the contribution from Debreu (1951) and Koopmans (1951), Farrell established the foundation for studies of efficiency and productivity on a micro-level. Farrell described economic efficiency as a combination of two measures: Technical efficiency and allocative efficiency. The first term measure a firm's ability to maximize output at a given input level, while allocative efficiency reflects the optimal input-mix which minimize the costs of an efficient production. This is depicted in picture 4.2. In figure 4.2, the curve Y

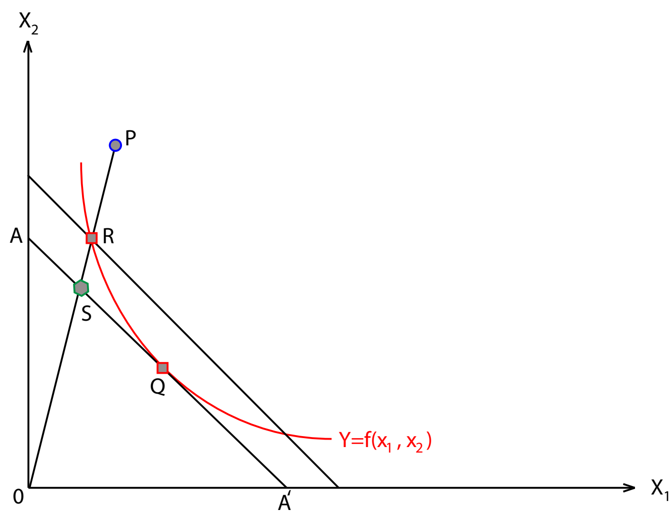

Figure 4.2: Technical and allocative efficiency

represent the isoquant where two inputs, x_1 and x_2 are consumed ($Y = f(x_1, x_2)$), alternatively defined by a distance function isoq$(y) = \{x : d_i(x, y) = 1\}$. The point P refer to one observed DMU producing one unit, which could reduce its inputs and still reach the isoquant. The ratio between used inputs (OP) and optimal input mix (OR) gives a measurement for technical efficiency ($TE = OR/OP$) or defined using distance functions: $TE = \frac{1}{d_i(x,y)} = d_o(x, y)$. By this, we can see that distance functions are reciprocals of Farrell's efficiency measure.

If we have information on input prices, we might also be able to define allocative efficiency. In figure 4.2 the line AA' being a tangent to Y in point Q represent the isocost curve, so that Q is the point where the price of the input mix is minimized. Allocative efficiency (AE) then become OS/OR.

This leads us to the final overall cost efficiency (CE), defined as:

$$\frac{OR}{OP} * \frac{OS}{OR} = \frac{OS}{OP}$$

We have now seen how the output distance functions are used for describing technical efficiency as the distance to the isoquant. The importance of technical efficiency will be stressed in chapter 4.3, but we have to be aware of a few important issues related to efficiency. As we will see in the next section, two firms A and B might both be technical efficient, but B still have a higher productivity ($\frac{y}{x}$) than A.

4.1.3 Decomposition of Total Factor Productivity

When we consider productivity change it is common to use total factor productivity (TFP). TFP is often decomposed into sources of productivity change, where each component has its distinct effect and causes.

For now, we have focused on firms operating under a global crs production technology, implying that all firms are automatically scale efficient. If the underlying production theory on the other hand is variable returns-to-scale (vrs), a firm might be too small or too large in its scale so that it operates at an increasing returns-to-scale (irs) or decreasing returns-to-scale (drs) respectively. This is depicted in figure (4.3) for a one output, two input vrs production technology as given in (Battese et al. 2005; page 59). By becoming scale efficient a firm will be able to increase its productivity and the problem of operating at the technically optimal productive scale (TOPS) is given as:

$$TOPS = max\ \{\ y/x|(x,y) \in S\ \} \tag{4.3}$$

This can be described as finding the feasible production point that maximizes productivity, illustrated in figure (4.4). Apart from improving productivity through scale efficiency change (SEC), decompositions into technical change (TC) and technical efficiency change (TEC) are well-established in the literature. Technical change (TC) is a result of change in production technology, measured as the change in ability to produce a level of outputs with a given input vector in period

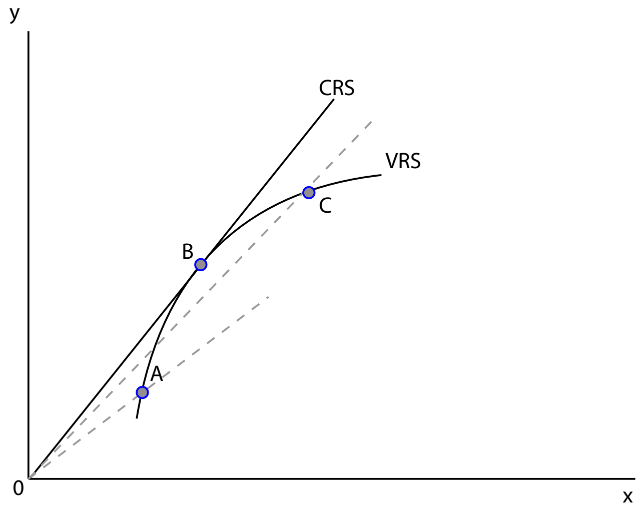

Figure 4.3: All firms are TE, but the rays from the origin show differences in productivity given as the slope y/x. This inconsistency is due to the effects of scale. A is operating at irs, C at drs and B is operating at TOPS.

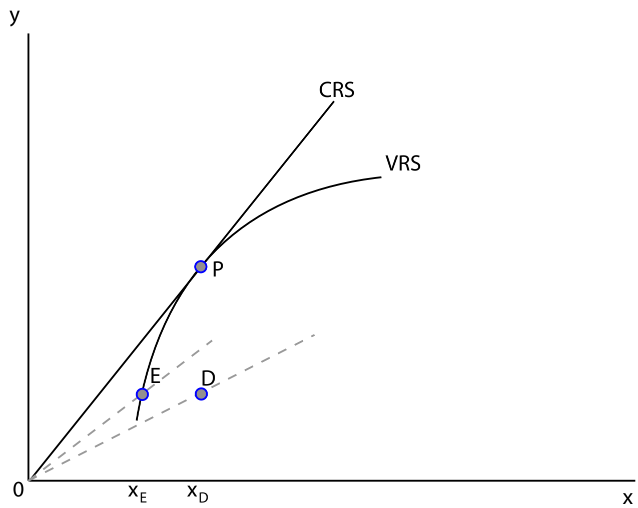

Figure 4.4: D can become efficient and increase productivity by moving to E. P is a feasible production point and is more productive, but as efficient as E.

t in comparison to levels feasible in period s (see figure 4.1):

$$TC_o^{s,t} = \frac{d_o^t(x,y)}{d_o^s(x,y)} \tag{4.4}$$

TC is equivalently the change in feasible input levels with a given output level. Positive TC is either illustrated by an expansion of the PPF-curve or contraction of the isoquant.

Technical efficiency change (TEC) is given by how much an observed output or input vector can be radially expanded to reach the frontier of the production possibility set:

$$TEC_o^{s,t}(x_t, y_t, x_s, y_s) = \frac{d_o^t(x_t, y_t)}{d_o^s(x_s, y_s)} \tag{4.5}$$

In his paper "Scale Efficiency and Productivity Change", Balk (2001) introduced yet another source for productivity change, named the output mix effect (OME) or equivalently input mix effect (IME). These two effects measure how changes in the composition of the input or output-vector effects scale efficiency. A general measure of OME can be defined for a technology in a given period (e.g. t) and input or output vector:

$$OME^t(x, y_s, y_t) = \frac{SE_o^t(x, y_t)}{SE_o^t(x, y_s)} \tag{4.6}$$

In other words, OME capture the effect on scale efficiency when moving from output vector y_s to y_t. Note that $OME = 1$ under CRS. This effect is on the other hand not well established in the literature and is less intuitive than the other sources of TFP change.

When we combine all factors the decomposition of TFP change become:

$$\text{TFP Change } = \text{TC} * \text{TEC} * \text{SEC} * \text{OME} \tag{4.7}$$

While the decomposition of TFP change is an important topic when assessing the sources of productivity change, it might be less relevant in case of assessing the overall performance of an employee for determining rewards. The decomposition of TFP change might give us an idea of why an employee has become more productive, thus being potential important in sake of adopting best practice. In the case of providing rewards we are on the other hand not interested in why an employee has become more productive, apart from revealing gaming or risk, but rather measuring the extent of change in productivity. Although, in some industries an increase or decrease in scale efficiency might result in lower profits and make the firm worse off. The most important lesson from the decomposition of TFP change, hereinafter productivity change, is that technical efficiency is a component of productivity change, and that the terms should not be used interchangeably. In the following we will focus on how measures of technical efficiency and productivity change might affect how agents behave under performance evaluation in incentive systems.

4.2 Efficiency and Productivity in Incentive Systems

Assume that we regard employees as DMUs, who carry out processes transforming input(s) into output(s), tangible or intangible. If we had knew the most efficient production mixes (underlying technology), the performance of an employee could be measured by the distance to the efficient frontier, receiving an efficiency score from 0 to 100 %. However, in most cases we do not know the underlying technology and must compare against an empirical estimated frontier. The most efficient DMUs in the reference technology set will construct this frontier, and the performance of a DMU is measured relative to other DMUs and/or established benchmark(s). For now the technical details on estimating this frontier are omitted, but will be the main focus in section 4.3. When measuring productivity change, we can do so against a set benchmark, relative to other DMUs or relative to the DMU under evaluation in another period. The technical details for estimating productivity change will be treated in section 4.4. and for now the focus is directed at the fundamental differences between measuring productivity and efficiency. In the following an evaluation on how the choice between technical efficiency fixed in time and productivity change might affect the outcome and behavior of agents.

When conceptualizing the difference between measuring efficiency and productivity a hypothetical firm with multiple plants will be utilized. These plants are producing identical products but at different geographical locations and to different markets. These plants will be subjected to differences in market fluctuations and organizational culture, giving birth to heterogeneity and environmental differences. In this firm, the CEO is considered being the principal delegating decision making authority to plant managers who are considered agents. In order to lower the ex ante transaction costs, the CEO has decided to incorporate the same performance evaluation scheme of plant managers at all plants. Output is total revenue (products are sold locally at the plant), while inputs are man-hours, operating costs (deducted for wages) and investments in tangible assets[2]. In the following, extrinsic rewards are limit to performance based pay.

4.2.1 Technical efficiency

When measuring technical efficiency DMUs are compared against an efficient frontier. For this comparison to be valid, the units under investigation must be able to reach the frontier, in other words, not be limited by a different technology than the one they are measured against. When comparing against a frontier, we have two distinct frontier cases: 1) The most efficient plants determine the frontier (relative), 2) the plants are compared against a known frontier (absolute).

If two plants, Alfa and Bravo, are producing the same good but Alfa uses a capital intensive technology while Bravo uses a labor intensive technology, they will not follow the same production function. If we on the other hand introduce a third plant, Charlie, identical to Bravo but with a more skilled and motivated workforce, heterogeneity might be less observable to the principal. In this case Charlie will most probably be more efficient than Bravo, thus the production manager at Charlie will receive a higher performance score than the manager at Bravo.

If we follow frontier case 1, Charlie will establish the frontier and become technical efficient. If Charlie always turn out to be efficient, the plant manager will only have weak incentives to increase his efforts as he always receive a reward. Only plant manager at Bravo have a strong incentive to increase his effort, although he might chose to not participate in the incentive system if he never receives a bonus. Weak incentives might be offset by limiting the total amount of performance pay, so that the entire budget for bonus payments is split amongst the players. The better plant Charlie operates, the worse plant Bravo will look and plant manager at Charlie captures more of the incentive pay.

[2]These are perhaps not the most useful inputs, but are selected for illustrative purposes.

Technical efficiency used as a performance evaluation scheme is desirable for an efficient manager, although it might only give weak incentives for increasing effort. The inefficient manager will on the other hand always have a strong incentive to increase its effort, but he might be limited by environmental factors. Therefore, technical efficiency violates the credibility criterion if a plant manager operates in a more difficult environment and the performance evaluation will seem unjust.

If we calculate efficiency against a known production frontier (frontier case 2) with deterministic data, the evaluation become absolute and will be unable to capture shocks. In this case, the performance of an agent is not affected by the performance of another, and the strength of relative performance evaluation is lost. At the same time, a pure deterministic model might induce the agent to blame the environment if he is inefficient, as a result of asymmetric information. Absolute performance evaluation has the advantage that no employees receive a bonus if everybody perform poorly, as opposed to relative evaluation where at least one will be rewarded. For empirical estimated frontiers it might therefore be advisable to include virtual units representing expected performances.

When measuring technical efficiency, we do so for a defined period. This period might consist of a single measurement point or aggregated data for several points. Either way, technical efficiency measures are status quo and do not tell us anything about the development in performance. Such temporal performance measures are inadequate to measure trends and thus lacks the ability to measure whether the organization has improved from one period to another. Temporal comparisons also lack the ability to capture the effect of price changes and comparability due to changes in production.

4.2.2 Productivity change

When using productivity change in performance evaluations, we measure the productivity of current period against the productivity in a base period. Again consider the plant managers at Bravo and Charlie, where manager at Charlie was the most efficient in the previous period. The CEO has decided to incorporate a scheme measuring productivity change from previous period and is indifferent whether the cause is change in scale efficiency or technical efficiency.

We remember from section 2.1.3 that several factors might cause variability in panel data, where unsystematic and systematic influences of nature imposes a risk on the agents. While unsystematic variability is difficult to account for, systematic fluctuations should receive great attention when base year is selected as it might affect the calculation of productivity change. If the base year is selected when nature has chosen favorable conditions, a plant manager might experience lower productivity in later periods simply because environmental factors (uncontrollable risk) have worsen. This might even be true for an agent who is not effort averse and who genuinely seeks to maximize the value of the organization. If the previous period in our thought experiment was dominated by high demands and has decreased for current period, the productivity at Bravo and Charlie is likely to decrease if the plant managers were unable to adjust inputs proportionally. This would result in a productivity score below 1, thus indicating a decline in productivity. If rewards only would be provided to plants with a TFP change exceeding 1, it would be cruel to omit seasonal market fluctuations in the equation. By comparing periods with similar characteristics, plant managers would always have an incentive to improve their productivity. If the evaluation periods rather were dominated by unsystematic fluctuations, it would be a better strategy to provide rewards on the basis of productivity change relative to other plants. Relative productivity change might be the best evaluation scheme during recessions. Another observation about base periods, is that they may become obsolete, as strategic choices and technologies are likely to change in time.

Productivity change as a performance evaluation always gives a strong incentive to increase evaluated performance, but state of nature give rise to moral hazard and adverse selection. Productivity change is therefore foremost a desirable framework if the market is stable, or when the principal and agent share the same information about the state of nature. But should increased productivity result in a bonus? Let's assume that plant Bravo and Charlie have identical production environments, but the manager at Bravo has invested a higher level of effort than manager at Charlie over the past years. When introducing productivity change as performance evaluation, the most effort averse plant manager is likely to have greater possibilities for improvement, assuming they have the same potential. If this is true, the manager at Charlie will receive the best performance score, thus contradicting the credibility criterion. Nevertheless, the objective of a incentive scheme is to increase the effort spent on value adding actions, thus rewarding bonuses to those who improve the most seems like a good strategy afterall.

4.2.3 Summary and comparisons

Before giving a simultaneous evaluation of technical efficiency and productivity in incentive systems, a summary of the findings so far will be given.

Technical Efficiency

- Efficient plant managers have weak incentives to improve
- Inefficient plant managers have strong incentives to improve
- In homogeneous environment, the best manager is likely to receive strongest incentives
- In heterogeneous environments, theoretically the worst manager might receive strongest incentives
- Temporal comparisons

Productivity Change

- Strong incentive to increase evaluated performance (unless private and perfect information about state of nature)
- Selection of base year increase complexity and might lead to gaming responses
- A base period might become obsolete in time
- Do not necessarily provide incentives to the most efficient plant manager
- Dynamic comparisons

It was noted that plant manager at Bravo would receive a lower efficiency score than manager at Charlie due to heterogeneity, thus being unjust as a performance evaluation. By adapting an evaluation scheme based on productivity change, the managers will be compared against themselves so that possible effects of heterogeneity diminishes. This would also provide a strong incentive to improve for the efficient manager, unless he has private information about market fluctuations which might give rise to moral hazard. But how may decision-making amongst plant managers be affected by the choice of evaluation scheme? If the efficient plant manager at Charlie had a choice whether to invest in a capital intensive technology or remain labour intensive, he could be reluctant to do so if he was assessed by TE. This is because he would then be evaluated against manager at

Alfa, who most likely would be more efficient due to superior expertise. Since the dominated strategy would be to not invest, the CEO would have to compensate the agent for investing in machines if this was desirable. If he rather was evaluated by productivity change, the manager would be more likely to invest in machines. Charlie should still be evaluated against Alfa, as plants with capital intensive technologies are more are likely to experience higher start up costs as well as low productivity under low demands. Let's also assume that the CEO was searching for a new manager at plant Charlie and Bravo and that a candidate was considering which job opening to apply for. If the evaluation scheme utilized technical efficiency, the applicant would apply for plant Charlie if he had knowlede about the differences between the workforces. By basing the evaluation scheme on productivity change, he would on the other hand be less reluctant to apply for the opening at Bravo.

Productivity might, opposed to measures of efficiency, detect and adjust for heterogeneous environments if the state of nature is somewhat fixed through time. Under both techniques we see how controllable and uncontrollable risk might distort the measures, and that environmental factors should be accounted for explicitly. Note that this is true for both KPI and composite measures as well. The perhaps most important observation is that technical efficiency is a static measurement, while productivity is dynamic. As treated in section 2.1.3, a high productivity is often more satisfactory than a high efficiency when evaluating how organizations evolve. Technical efficiency is important for evaluating present state of performance so that managers might detect e.g. best 25% and worst 25% performers. Technical efficiency and productivity change have different applicabilities and should be regarded as complements, as they tell different tales about the performance of an employee.

4.3 Technical Efficiency Estimation

Technical efficiency was introduced in section 4.1.2 through Farrell's isoquant diagram, which can equally be derived through distance functions. Estimating TE is difficult in practice as we do not know the shape of the isoquant, or similarly the production function. In order to estimate a production function, Farrell constructed a scatter plot of the input mixes for different DMUs, each producing one unit. Assuming that it was possible to define new DMUs through convex combinations of observed homogeneous[3] units, Farrell then constructed the empirical production possibility frontier (depicted in figure 4.5).

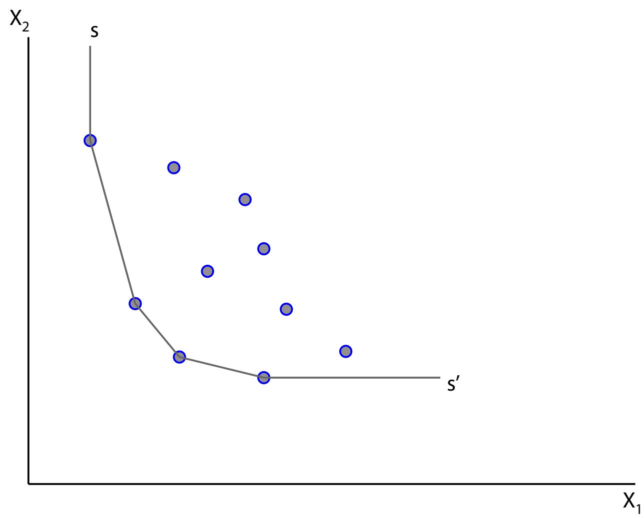

Figure 4.5: Farrell's estimation of the empirical PPF

The four DMUs on the frontier are all technical efficient by the definition of Pareto-optimality (Warburton 1983), while the other firms inefficiency are given by their distance to the frontier.

Farrells method for deriving the PPF is fundamental for several empirical estimation techniques, and in this thesis I will cover the most used techniques, Stochastic Frontier Analysis (SFA) and Data Envelopment Analysis (DEA).

4.3.1 Data Envelopment Analysis (DEA)

DEA is a non-parametric and deterministic technique for estimating the production function. The first DEA-model was developed by Charnes et al. (1978), also known as the CCR-model. The purpose of the technique was to calculate the relative efficiency of a DMU through a fractional linear programming formulation on the

[3]Producing the same kind of products with identical inputs under similar conditions.

form:

$$\text{maximize } \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \tag{4.8a}$$

s.t.

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \qquad \forall j \in I \tag{4.8b}$$

$$u_r, v_i \geq 0 \qquad \forall r, i \tag{4.8c}$$

where

n : number of DMUs

s : number of outputs

m : number of inputs

y_{rj} : output r for DMU j

x_{ij} : input i for DMU j

u_r : weight on output r

v_i : weight on input i

DMU_k : DMU to be evaluated

I : reference set with all units

When estimating the efficiency for the DMUs, the objective function is calculated for one unit at a time (DMU$_k$). The objective function (4.8a) maximizes the fraction of weighted sum of outputs over the weighted sum of inputs, with the weights as decision variables. For each optimization problem, constraint (4.8b) ensure that no unit in I achieve more than 100% efficiency (objective function is equal 1), while (4.8c) ensure non-negative weights. If one of the DMUs in I achieve an efficiency score of 1 with the optimal weights of DMU$_0$, the DMU under evaluation cannot be efficient unless DMU$_0$ is a scaled version of one of the units in I. It is worth mentioning that a DMU might have several sets of optimal weights for the same efficiency score.

By performing a linear transformation of formulation (4.8), the problem of estimating efficiencies can be formulated as a set of n linear programs (LP). The distance to the efficient frontier can either be measured as a proportional reduction in input-usage or proportional increase in outputs. These are denoted input-oriented or output-oriented models respectively, and should be selected on the basis of which orientation the DMUs have the most control over. E.g., if outputs are more or less fixed by certain demands one should adopt the input orientation.

When performing a linear transformation of formulation (4.8) and assume that the DMUs is capable of reducing their inputs, the CCR-model is transformed on the form:

$$\max \sum_{r=1}^{s} u_r y_{r0} \tag{4.9a}$$

s.t.

$$\sum_{r=1}^{s} u_r y_{r0} - \sum_{i=1}^{m} v_i x_{i0} \leq 0 \qquad \forall j \tag{4.9b}$$

$$\sum_{i=1}^{m} v_i x_{i0} = 1 \tag{4.9c}$$

$$u_r, v_i \geq 0, \qquad \forall r, i \tag{4.9d}$$

This formulation is the CCR-model on multiplier form, which can be transformed to the envelopment form by taking its dual:

$$\theta^* = \min \theta \tag{4.10a}$$

s.t.

$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{i0} \qquad \forall i \tag{4.10b}$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0} \qquad \forall r \tag{4.10c}$$

$$\lambda_j \geq 0, \qquad \forall n \tag{4.10d}$$

The objective function (4.10) calculates the efficiencies of DMUs by assuming that all units are operating at an optimal scale, thus measuring TE that are confounded by scale efficiencies (Charnes et al. 1985). The introduced dual variables consist of one $1 * n$ vector λ and a scalar θ. A DMU is defined as efficient if $\theta = 1$, while being increasingly inefficient as θ approach zero. For an inefficient DMU, θ is the minimum radial reduction of inputs to become efficient.

The λ-vector contain the convex combination of DMUs constructing the local efficient frontier for an inefficient DMU. If we go back to figure 4.5 and draw a straight line from the origin to the evaluated DMU, the line will be the radial contraction path for its inputs. The units that constructs the convex local frontier, which is intersected by the contraction path, constitutes the efficient reference DMUs.

Because units may not operate at optimal scales, Banker et al. (1984) seeked to separate scale efficiencies and TE. This was accomplished by adding the restriction $\sum_j^n \lambda_j = 1$ to equation (4.9), resulting in the BCC-model assuming variable returns to scale (VRS):

$$\max \sum_{r=1}^{s} u_r y_{r0} + u_0 \tag{4.11a}$$

mhp

$$\sum_{r=1}^{s} u_r y_{r0} - \sum_{i=1}^{m} v_i x_{i0} + u_0 \leq 0 \qquad \forall j \tag{4.11b}$$

$$\sum_{i=1}^{m} v_i x_{i0} = 1 \tag{4.11c}$$

$$u_r, v_i \geq 0,\ u_0 \text{ free }, \qquad \forall r, i \tag{4.11d}$$

While CCR define a linear front BCC define a piece-wise linear front of non-dominated units, as we can see in figure (4.6).

Under CCR, the efficiency of DMU$_B$ is calculated as $\frac{MN}{MB}$, while BCC efficiencies are calculated as $\frac{MO}{MB}$. From this we can derive scale efficiency as the efficiency score under CCR over the efficiency score under BCC. From this follows that the efficiency scores calculated under BCC will be equal or higher than efficiency scores derived from CCR. In addition, we will expect BCC to identify more efficient DMUs than CCR (DMU$_A$ is defined efficient under BCC, but not CCR), while all CCR-efficient DMUs will automatically be BCC-efficient (DMU$_C$). BCC and CCR make a priori assumptions about the global returns to scale (grs) for the technology, but in cases where the choice of grs is not obvious one should adopt the two-stage hypothesis approach proposed by Eopold Simar and Wilson (2002). In empirical applications it is most common to report efficiency scores from both BCC and CCR models, enabling us to determine whether efficiencies are scale size dependent. In the DEA-literature, a great deal of attention has also been brought to the determination of whether a DMU operates at a locally crs, drs or irs. Three

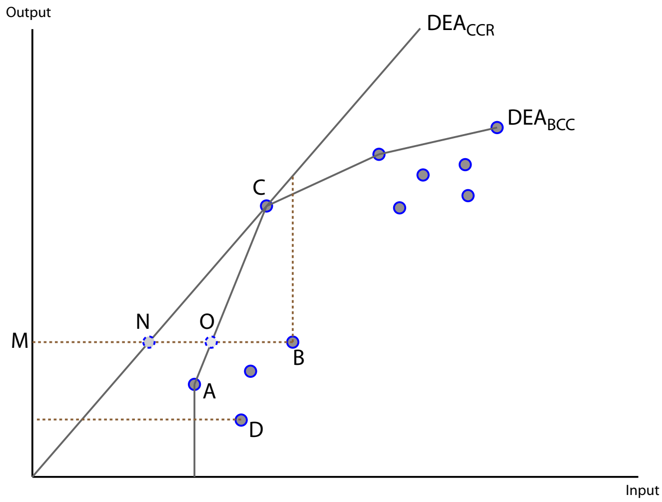

Figure 4.6: CCR define a linear front, while BCC define a piece-wise linear front.

basic methods are summarized in an article by Seiford and Zhu (1999), but all have in common that they evaluate the sum of λs for each DMU. In short, if the sum of λ is constant in any alternate optima, $\sum_{j=1}^{n} \lambda* = 1$:crs, $\sum_{j=1}^{n} \lambda* > 1$:drs and $\sum_{j=1}^{n} \lambda* < 1$: irs. A last theoretical observation on the difference between the BCC and CCR model is that the choice of orientation does not affect the efficiency scores for CCR. Under BCC on the other hand, the input and output-orientation will not yield identical efficiency scores for inefficient units, but as Battese et al. (2005)[p. 181] note: *"...output - and input orientated DEA models will estimate exactly the same frontier and therefore, by definition, identify the same set of firms as being efficient."*.

DEA is under continuous methodological development, where extensions and new approaches facilitates the use of DEA in increasingly complex problems and new applications. In a recent paper by Cook and Seiford (2009), the authors summarizes the development of DEA over the past 30 years, which also might serve as an additional introduction on the topic.

Examples of use

DEA is under continuously development and has been applied on a large variety of applications across different organizations and industries. It has become common practice for evaluating performance in health care (Huang and McLaughlin 1989, Parkin and Hollingsworth 1997) and schools (Alexander et al. 2007, Coelli 1996), to mention a few applications. DEA was also used by Barr et al. (1991) for an early identification of troubled banks, with high accuracy. By adding units from other periods to the reference set, DEA has been used in trend studies where technical efficiency is compared between periods. This goes by the term "window analysis" in the DEA-literature and it is often common to use a rolling window of three periods (Charnes et al. 1984; example). DEA might also be used for cost minimization and revenue or profit maximization when input and output prices are available.

Pros and cons

DEA is superior in its ability of simultaneously handling multiple inputs and multiple outputs and ease of formulating. Because DEA is non-parametric, we do not have to make any a priori assumptions about the shape of the frontier. This yields great flexibility, but free weighting of parameters might be undesirable because a DMU will under-emphasize weak performances and overemphasize on strong performances. This can be resolved by setting an explicit interval for the relative weights to be placed on each parameter by adding constraints to the linear optimization problem on multiplier form. This method goes by the name assurance region (Dyson and Thanassoulis 1988), but the cone ratio method Kornbluth (1991) is also a well established method. Assurance regions will ensure a balanced evaluation of a DMU, conform with the criterion about a balanced measurement. In order to implement an assurance region we simply add conditional constraints to equation (4.9):

$$lb \leq \frac{u_i}{\sum_{r=1}^{s} u_r} \leq ub$$

$$\text{resulting in two linear constraints:} \tag{4.12}$$

$$u_i - lb \sum_{r=1}^{s} u_r \geq 0$$

$$ub \sum_{r=1}^{s} u_r - u_i \geq 0$$

,where lb is lower bound and ub is upper bound on the relative weight u for parameter i. Alternatively we might apply ordinal ranking of the relative importance of each parameter. E.g. $u_{i+2} \geq u_i \geq u_{i+1}$:

$$u_i - u_{i+1} \geq 0$$
$$u_{i+2} - u_i \geq 0$$

Another issue is the increasing degree of freedom - increase in number of optimal product mixes - as ratio of DMUs to parameters get low. This will make more DMUs look efficient and lower the discriminating power. To increase the discriminating power, one might reduce the number of parameters by combining and eliminate highly correlated variables or apply the super efficiency model. By implementing the super efficiency model (Banker and Gifford 1988) DMUs can on the other hand receive a score greater than 100%, thus increasing the discrimination power (see figure 4.7). Note that efficient units are not indifferent whether a BCC or CCR-model is applied, as the super efficiency scores based on BCC and CCR will differ. Also note that super efficiency models might be infeasible as it may not be possible to radially expand or contract a DMU to the frontier when the evaluated DMU is removed from the reference set.

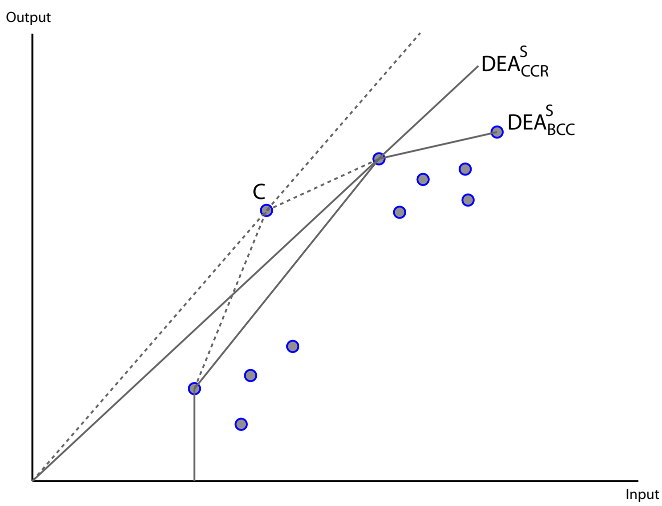

Figure 4.7: The figure depicts how the efficient frontier is altered when DMU_C is excluded from the reference set. Super efficiency for DMU_C may be estimated as the distance to either DEA^s_{BCC} or DEA^s_{CCR}

Most limitations of the standard DEA model might be overcome by extending the model, but there are a few issues which are more difficult to cope with. First, DEA is a deterministic method and for this reason it will loose credibility when noisy data and heterogeneous units are used. The dataset should therefore be of high quality, which is important for the calculation of KPIs and composite measures as well. Secondly, DEA might be perceived as a complex technique if one is unfamiliar with optimization procedures and mathematical expressions. The most important implementation issue is perhaps to gain user acceptance, so that the calculation seem transparent and the results are credible. Finally, DEA might identify DMUs as efficient even though they might not be efficient by the definition of Pareto-Koopmans efficiency. DEA do not distinguish non-dominated DMUs from weakly non-dominated DMUs, and both appear to be on the frontier even though the last class of DMUs could reduce their inputs and still produce the same level of outputs. Some practitioners adopt a multistage DEA model where output or input slacks are captured, but other practitioners state that the importance of slacks are overstated and view it as an artifact of DEA (Ferrier and Lovell 1990).

Use in incentive systems

Even though DEA has become common practice for performance measurement, it seems to be less common in incentive systems. For one example on how DEA might be incorporated in a incetive system, the reader is adviced to read an article by Sexton et al. (1994), who used DEA for implementing a pupil transportation funding process encouraging operational efficiency. Banker (1980) was the first to publish a paper on a game theoretic approach to measuring efficiency, which lead way to applications for optimal contracting. In two papers by Bogetoft (1994, 1995) the author propose a solution to the incentive problem, formalized as a contract design problem. While Bogetoft in his first paper focused on cost minimal implementation, he focuses on the problem of implementing desired production plans in his second paper. Together with Bogetoft, Jørgen Tind and Per J. Agrell (2000, 2002) are regarded to be the most active individuals in agency theory and DEA to this date, although their focus is on incentive regulatory mechanisms and not performance evaluation of groups or individuals.

When implementing a DEA-model for use in incentive systems, the principal have a choice whether a CCR (crs) or BCC (vrs) model should be applied and whether to select an input or output orientation. Employees who are efficient under CCR are automatically efficient under BCC and are indifferent about the choice of global returns to scale or orientation. Note that if we apply the super efficiency model or the entire budget under performance pay is split amongst the employees, the efficient players are no longer indifferent. To guide a principal to choose between the global returns to scale, one may either adopt the two stage hypothesis or examine the nature of the production environment. In the CCR model we believe that all firms are scale efficient and compare units with different scale sizes. In practice, factors that put constraints on input- or output quantities such as imperfect competition, regulations and restrictions on finance might cause DMUs to operate at sub-optimal scale sizes. Hence we should use the BCC model where such effects prevails, i.e. where employees have different quantity constraints[4].

When choosing between orientations, one option is to select the orientation for which the manager has most control but we should also consider strategic choices. When it comes to the efficiency of employees, the principal are more likely to emphasize on increasing the outputs, especially if work-hours is the only input. We might argue that decreasing work-hours might open up a new job position or reduce overtime but traditionally, and particularly in many public jobs, an employee is employed for a fixed number of work-hours. Also, in order to increase market share, a firm might be more focused on increasing production so that an output orientation should be adopted. A firm should on the other hand adopt an input-orientation if the focus is on resource minimization and lay-offs.

The use of relative efficiency evaluation is both a strength and weakness for use in incentive systems. As pointed out in section 2.2.2, the use of relative performance evaluation might induce the agents to maximize their effort. On the other hand, when the number of parameters get high relatively compared to number of DMUs, the discriminating power will diminish. In order to increase the discriminating power, the principal should seek to reduce parameters or introduce virtual DMUs which serve as standards or expectations. When virtual DMUs are constructed, the principal should avoid constructing linear combinations and focus on generating possible corner points. By adapting the super-efficiency model proposed by Andersen and Petersen (1993) we might allow units to achieve more than 100% efficiency, so that a ranking is made possible. This will also aid to eliminate unwanted Nash-equilibrium, as employees always can improve their efficiency score. Super efficiency models should be used with caution as it might encourage further specialization, resulting in dysfunctional responses and give rise to the multitasking

[4]The clusters in figure 4.6 witnesses of such restrictions.

problem.

Perhaps the greatest advantage of DEA is that it is non-parametric, so that the principal does not have to specify the weights. The specification of weights is perhaps the most difficult part of assessing overall performance when composite measures are used, and is a source to lobbying. When the weights are set mathematically by the maximization procedure, lobbying becomes nearly impossible and an agent cannot rightfully complain about the weights placed on the different parameters. In fact, every employee will automatically be assigned a set of weights that are most favorable to them. The use of assurance regions might invite to some discussion, but to a far lesser extent. The use of dynamic weighting of parameters seems more appealing than fixed weights as different employees possess different skills and expertise, so that an organization accept several ways to be efficient.

Even though DEA has gained acceptance as a technique for evaluating performance and for identifying best practice in the research literature, it has not been found evidence of extensive use in incentive systems. DEA will most likely be regarded as an unfamiliar technique and great effort should be spent on communicating its applicability and methodology. Some practitioners are likely to be overwhelmed by the compact mathematical formulation, and DEA will at first seem less intuitive than Balanced Scorecard. The technique might therefore be considered as a "black box", contradicting the transparency and attractivity criterion. If carefully explained, the underlying objective of DEA should be easy to grasp if the technical details are omitted and one communicate to the agents that *"the weights you are assigned will make you look as good as possible compared to your peers"*.

The ex ante cost of communicating and implementing DEA increase overall transaction costs, but will increase user acceptance. In fact, DEA might be perceived as a reliable technique which avoids favoritism through its optimization procedure and objective weighting. DEA seems therefore promising as a technique for assessing overall performance in an incentive system, as it limit the effect of corruptible responses. Both by using relative performance evaluation, or semi-relative by adding virtual DMUs, and through the limited interference of a principals subjective evaluation.

4.3.2 Stochastic Frontier Analysis (SFA)

The foundation for SFA was established by Aigner and Chu (1968), who assumed that a production possibility frontier could be estimated through an underlying functional form. This functional form was supposed to reflect the production properties of the DMUs, and they assumed a Cobb-Douglas function on the form:

$$\ln y_i = \beta_0 + \beta \ln x_i - u_i \qquad i = 1, \cdots, I \tag{4.13}$$

where

I : Number of units

y_i : output from unit i

x_i : K*1 input-vector

β : K*1 vector with unknown parameters

u_i : a non-negative stochastiv variable with assumed probability density function

From this formula, we can see that SFA estimate a PPF for a production mix with one output an multiple inputs. This might be equally transformed to a single-input multiple-output case by changing the sign of the inefficiency term u (inefficiency increase the use of input). SFA might also be extended to a multiple input- output case by using distance functions, but this calculation is not straightforward and imposes several computational issues.

In order to estimate the unknown parameters Aigner and Chu took use of linear programming, alternatively quadratic programming if a linear functional form is assumed, while Richmond (1974) took use of modified ordinary least squares (MOLS)

A drawback with formulation (4.13) is that it neglects measurement error and other statistical noise, so that the distance to the frontier is strictly due to technical inefficiency. For this reason Aigner et al. (1977) and Meeusen and van Den Broeck (1977) introduced independently of each other a new stochastic variable v_i, capturing noise with a positive or negative contribution:

$$\ln y_i = \beta_0 + \beta ln x_i + v_i - u_i \qquad i = 1, \cdots, I \qquad (4.14a)$$

alternatively

$$y_i = \underbrace{exp\left(\beta_0 + \beta ln x_i\right)}_{\text{deterministic component}} * \underbrace{exp\left(v_i\right)}_{\text{noise}} * \underbrace{exp\left(-u_i\right)}_{\text{inefficiency}} \qquad (4.14b)$$

The newly introduced variable enable an inefficient unit to lie above the frontier because the noise gives a positive contribution, exceeding the inefficiency ($v_i > u_i$). This is illustrated in figure 4.8, where a deterministic Cobb-Douglas function with decreasing returns to scale for one input is assumed. Figure 4.8 depicts two units,

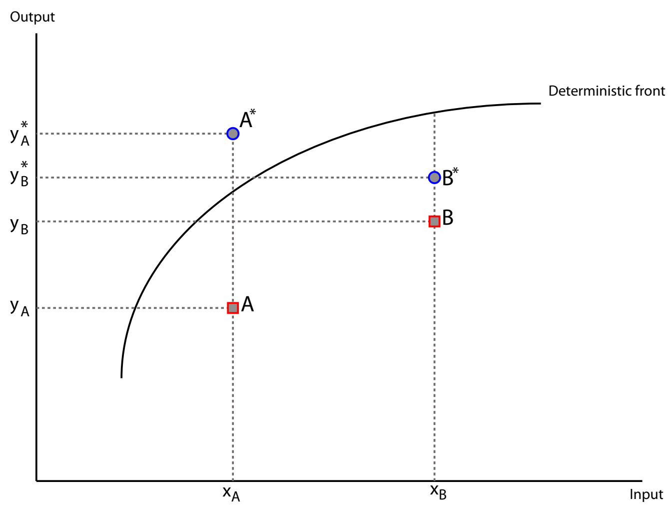

Figure 4.8: q_i : equation (4.14) and q_i^* : equation (4.15)

A and B, represented on the graph by their reported production mix, while A* and B* gives the production mix if the units had been technical efficient:

$$\ln y_i^* = \beta_0 + \beta ln x_i + v_i \qquad i = 1, \cdots, I \qquad (4.15)$$

As both units are assumed to be technical efficient, the deviation of A* and B* from the deterministic frontier is a result of a positively contributing noise element for A and negative element for B. In fact, B is more efficient than A, as A is further away from the frontier.

In order to estimate the technical efficiency for a DMU, we calculate the observed output over the corresponding output from formula (4.15). Or in other words, the observed output level over the optimal output level with identical input vector:

$$TE_i = \frac{y_i}{y_i^*} \tag{4.16a}$$

$$= \frac{exp\left(\beta_0 + \beta ln x_i\right) * exp\left(v_i\right) * exp\left(-u_i\right)}{exp\left(\beta_0 + \beta ln x_i\right) * exp\left(v_i\right)} \tag{4.16b}$$

$$= exp\left(-u_i\right) \tag{4.16c}$$

To be able to calculate the stochastic variable, we have to assume a probability density function for v and u. Aigner et al. assumed that the v_is were independently and identically distributed normal random variables with zero means and variances σ_v^2, and that the v_is were independently and identically distributed half-normal random variables with scale parameter σ_u^2:

$$v_i \sim iidN\left(0, \sigma_v^2\right)$$
$$u_i \sim iidN^+\left(0, \sigma_u^2\right)$$

By parameterizing the log-likelihood function of equation (4.14) and use $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\lambda^2 = \sigma_u^2/\sigma_v^2 \geq 0$, they arrived the following equation:

$$\ln L(y|\beta,\sigma,\lambda) = -\frac{I}{2}\ln\left(\frac{\pi\sigma^2}{2}\right) + \sum_{i=1}^{I}\ln\Phi\left(-\frac{\epsilon_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\epsilon_i^2 \tag{4.17}$$

where $\epsilon = v_i - u_i$ and $\Phi(x)$ is the cumulative distribution function for x.

Aigner et al. then estimated the parameters by maximizing the log-likelihood function (4.17) through an iterative optimization procedure, as the equation is not possible to solve analytically. The iterative optimization used OLS-estimation to initialize the parameters β, σ and λ, and then a MLE-estimation to finalize the determination of the parameters. This calculation is rather complex, and practitioners often take use of the program Frontier v. 4.1 [5]. See appendix B for example code.

Another frequently used method to estimate the parameters is Bayesian markov-

[5] Frontier v. 4.1 use MLE in an iterative Davidon-Fletcher-Powell Quasi-Newton process

chain monte carlo (MCMC), introduced by Smith and Roberts (1993), van den Broeck et al. (1994) and Koop (1994). This method take use of a Gibbs sampling algorithm for generating parameters, which converge toward final values after a sufficiently large number of iterations[6]. For a more thorough review of the techniques, the reader is advised to read an article by Koop and Steel (2007) as well as the mentioned articles above. For empirical applications, I strongly recommend to use WinBUGS 1.4[7].

The most challenging task when implementing SFA is to choose an explicit functional form of the production function. In the literature, Cobb-Douglas and the translog function is often used, but linear and quadratic functional forms are also common. Cobb-Douglas is often used as it exhibit convenient mathematical properties, such as ease of estimation and analyzation. At the same time the Cobb-Douglas function yield constant elasticity for returns to scale and is therefore poor for estimating the efficiency of DMUs with variable returns to scale. The translog function is more flexible, but need more observations to provide reliable estimations (Kuenzle 2005). The task of identifying the best functional form is therefore a difficult task which was addressed in a work by Lau (1986) where he proposed several conditions to be met:

1. Theoretical consistency
2. Domain of applicability
3. Flexibility
4. Computational facility
5. Factual conformity

For a full treatment of these conditions, the reader is advised to look up the article cited above. It is worth mentioning that the flexibility criterion is a trade off to an increase in the degree of freedom. At the same time, flexible production functions might be inconsistent with economical theory of a monotonic and quasi-concave production function (Feng and Serletis 2009).

Apart from choosing the functional form there are several probability density functions (pdf) proposed for the inefficiency term. In addition to half-normal, the most

[6] Kim and Schmidt (2000) compare the two methods and find that both are consistent.
[7] `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`

common are gamma- (Greene 1990), exponential (Meeusen and van Den Broeck 1977) and truncated-normal pdf's (Stevenson 1980).

Examples of use

SFA has been used for evaluating efficiency in many of the same applications as DEA: health care (Rosko 2001), schools (Stevens 2005) and banks (Kraft and Tirtiroglu 1998), to mention a few.

Pros and cons

As SFA is a parametric technique, the most important question is what functional form to impose. If we identify a functional form satisfying the criteria established by Lau, SFA is an exceptional technique for evaluating performance. A perfect fit is on the other hand far from common in practice and usually several functional forms must be tested. If the different functional forms yields significantly different results in TE and we do not have any clear indications of which is the most promising form, the technique loses its credibility. The stochastic nature of the error term is also both a strength and a weakness of SFA. The error term might capture some sources to heterogeneity as well as measurement errors, but it might also introduce model generated noise, even when the dataset is free of measurement errors.

The mathematical procedure of estimating the parameters through OLS and MLE has been reported to result in unexpected results because of multicollinearity, even for medium to weak correlation between the parameters (Gundersen 2010). The most common solution to overcome this problem is to respecify the model or to extend or reject parts of the dataset. This problem is known as the "skewness issue" and might cause problems even for correct specified functional forms (Simar and Wilson 2010). The best known solution is to use Bayesian MCMC instead of OLS and MLE, but this might introduce monte carlo generated noise and the parameters estimated might deviate from production theory where output is assumed to be nondecreasing with increased outputs (Gundersen 2010). The occurrence of multicollinearity is even greater for SFA formulations with multiple inputs and outputs.

The obstacles described above makes the estimation of TE difficult, and the technique require extensive skills from the user. Event hough SFA requires extensive skills the possibility of fine-tuning the model is low.

Use in incentive system

SFA seems to be little exploited for performance evaluation in incentive systems and the literature is concentrated around regulatory methods of firms, especially for the electricity industry (Knittel 2002, Jamasb and Pollitt 2001) and optimal structures of CEO compensations (Habib and Ljungqvist 2005).

SFA has its strength in balancing multiple inputs or outputs, and might cope for heterogeneity to some extent through the error term. While satisfying the criterion of giving a balanced measurement, the technique is less satisfactory on the other criteria. As stated above, estimation difficulties and choices regarding functional form makes the technique less attractive. The technique is likely to be regarded as a black box where the transparency is low, so that costs regarding implementing and communicating the model increase. A well specified functional form of the production function would yield high credibility, but is difficult to select. SFA is not corruptible in means of lobbying by agents, but the error term might in fact introduce uncontrollable risk on the agents. The reports of parameters that not conform to production theory might also lower the credibility of the technique significantly.

4.3.3 Proposed Technique for Estimating Technical Efficiency

There are written many articles comparing SFA and DEA in light of regular performance evaluation, but there are controversies about which is the best method. SFA is assumed to provide better estimations when measurement errors are present and when the functional form is known to some extent. Under a well-specified functional form SFA require less observations than the DEA to make a good approximation of the frontier, but is poor when it comes to multicollinearity. In an article by Mortimer (2002) the author compare the results from 41 articles reporting efficiency scores from both DEA and SFA and find medium to strong correlations. There are no obvious difference in which technique calculating the highest average score, but DEA normally identify more efficient units than SFA, while having a larger variance in efficiency scores. For a good paper summarizing comparisons between DEA and SFA, the reader is advised to read an article by Lin and Tseng (2005).

The fact that DEA identify more efficient DMUs than SFA is supported by the larger flexibility of DEA compared to SFA. In cases where DEA identify multiple efficient DMUs, the technique is not sufficient to identify the best employee. SFA are more likely to identify one unique efficiency score as the best unit, but low variance in efficiency scores might seem like a problem. This is particularly true when Bayesian MCMC is used for estimating parameters as the ranking of units might change.

DEA seems as a far more intuitive method than SFA with a significant higher transparency. This is likely to make DEA easier to accept by its users, as well as lowering the ex ante and ex post costs related to implementing and monitoring. DEA is also implicit conform with production theory of a monotonic and non-decreasing frontier, opposed to SFA. These final remarks are summarized in table 4.1.

	Atr.	Cred.	Balanced	Non-corruptible	Transaction cost
DEA	2	2	3	3	2
SFA	1	2	3	2	1

Table 4.1: The higher score, the better. Note that DEA is balanced in general only if assurance regions are used.

From table 4.1 one can see that DEA perform better than SFA on all criteria. DEA is therefore assumed to be the preferred technique for estimating technical efficiency in incentive systems.

4.4 Productivity Change Estimation

Productivity is normally described as the number of outputs over the number of inputs:

$$\text{productivity } = \frac{y}{x} \tag{4.18}$$

While being easy to calculate for one input and output, the calculation become more difficult as multiple inputs or outputs are introduced. In such cases we have to perform some kind of aggregation to obtain a ratio for productivity. While KPIs are thought of as partial productivity measures, I will in the following focus on total factor productivity and productivity change.

Productivity change from period s to t for a single-input single-input case is measured by the ratio:

$$\text{Productivity change } = \frac{y^t/y^s}{x^t/x^s} = \frac{y^t/x^t}{y^s/x^s} \tag{4.19}$$

This representation is intuitive and indicate that we can think of productivity change as either the output quantity index over the input quantity index or period-t productivity over period-s productivity. This can be represented for a multiple input, multiple output case using firm specific production mixes, $G(x, y)$:

$$\text{Productivity change } = F(x_t, y_t, x_s, y_s) = \frac{G(y^t, x^t)}{G(y^s, x^s)} \tag{4.20}$$

$G(x, y)$ being a homogeneous function of degree $+1$ in y and -1 in x.

In both equation (4.19) and (4.20) we must select a base period for s. Issues regarding the selection of a base period were briefly discussed in section 4.2.2, but should be given additional attention. When comparing the productivity between periods, we might either hold the base year fixed (fixed-based index) or compare each year with the previous year (chain index) and then combine the changes to measure change over a longer period. Ideally the change in productivity should be the same when measuring from e.g. period t to $t+2$ through $t+1$ and between t and $t+2$ as fixed base, and similarly between different units in the same period. If this is the case, the productivity index has the property of transitivity.

The property of transitivity is desirable as it ensure measurement consistency between units and time periods. In order to construct transitive indexes the technology should exhibit Hicks neutrality, input and output separability and simultaneously homotheticity. Note that a productivity index has the property of circularity if there are no productivity change when going from A to B to A, which is a sufficient but not necessary condition for transitivity (Peyrache 2010; for a detailed summary). If the conditions are not satisfied, the measurement of productivity might become biased and inconsistent, which one should strive to quantify. One, but computational expensive, method is to measure the productivity change from A to C through B and from A to C directly. The difference in productivity change would then quantify a possible inconsistency of the index. With a sufficient number of comparisons, this will enable us to estimate confidence intervals for the scores.

According to Battese et al. (2005), measuring TFP change from one period to another can be carried out in four ways:

- Malmquist TFP index
- Hicks-Moorsteen TFP index
- Profitability ratio (index numbers)
- Component-based approach

The component-based approach was treated in section 4.1.3, where TFP change was measured as a product of several individual factors. The remaining measures of productivity change will be treated in the order listed above.

4.4.1 Malmquist TFP index

The Malmquist approach as a TFP index was introduced by Caves et al. in two papers (982a, 982b). The technique measure productivity change of a DMU by comparing output and input vectors in period s and t based on either period-t technology or period-s technology. Alternatively, we can measure change relative to another DMU in the same or different period.

An output-orientated Malmquist TFP index based on period t technology

$$m_o^t(x_t, y_t, x_s, y_s) = \frac{d_o^t(x_t, y_t)}{d_o^t(x_s, y_s)} \tag{4.21}$$

An output-orientated Malmquist TFP index based on period s technology

$$m_o^s(x_t, y_t, x_s, y_s) = \frac{d_o^s(x_t, y_t)}{d_o^s(x_s, y_s)} \tag{4.22}$$

As we can see, the Malmquist approach take use of the distance function, here represented by the output-oriented distance function. The choice of reference technology is arbitrary when the production possibility curve exhibit CRS, but not necessarily for VRS. Färe et al. (1992) specified the Malmquist TFP index as the geometric average of the two indexes based on period t technology and s technology:

$$m_o(x_t, y_t, x_s, y_s) = \left[m_o^t(x_t, y_t, x_s, y_s) \times m_o^s(x_t, y_t, x_s, y_s)\right]^{0.5} \tag{4.23}$$

In the same paper they showed that this could equally be decomposed to

$$m_o(x_t, y_t, x_s, y_s) = \underbrace{\frac{d_o^t(x_t, y_t)}{d_o^s(x_s, y_s)}}_{\text{TEC}} \underbrace{\left[\frac{d_o^s(x_t, y_t)}{d_o^t(x_t, y_t)} \times \frac{d_o^s(x_s, y_s)}{d_o^t(x_s, y_s)}\right]^{1/2}}_{\text{TC}} \tag{4.24}$$

However, while the fixed base Malmquist index is transitive, the geometric average is not. Also note that the underlying production technology is assumed to be CRS to measure TFP change. This can easily be derived from equation (4.24), where $m_o(x_t, y_t, x_s, y_s) = 1$ when there are no technical change between the periods and the firm is technical efficient.

Examples of use
The Malmquist index is usually coupled with DEA and has for this reason been applied to many of the same industries as DEA. Some examples are banking (Bukh et al. 1995) and agriculture (Coelli and Rao 2005). The Malmquist index has gained an increase in popularity the last years, and a search on Google Scholar with the words “Malmquist index” results in $4'050$ related articles.

Pros and cons
As the technique make use of distance functions, the Malmquist index relies on either a known or well-estimated technology to provide useful results. In cases where the technology is well-defined, the Malmquist index is easy to estimate and the decomposition might provide useful information about sources of TFP change.

The Malmquist index will on the other hand fail to accurately measure productivity change unless the underlying technology exhibits global crs (Grifell-Tatjé and Lovell 1995). Assuming crs by default when measuring productivity change seems little appealing and is the most negative aspect of the technique. The Malmquist index might also turned out to be infeasible when coupled with DEA infeasible LPs. This occur when an evaluated DMU is not a part of the reference set and cannot be radially expanded or contracted to the frontier (solving $d_o^t(x_s, y_s)$ or $d_o^s(x_t, y_t)$).

Use in incentive systems
Because the Malmquist in most applications rely on distance functions estimated by SFA or DEA, and those techniques are little exploited in reward systems, the Malmquist index itself is also little exploited. Two papers that might be related to incentive systems are Camanho and Dyson (2006), focusing on group performance

and Nghiem and Coelli (2002) investigating the effect of incentives on productivity. These papers focus on the other hand on the sources of productivity growth and not on the appraisal system itself.

As stated under pros and cons, the Malmquist index is dependent on a well-defined technology in order to provide useful results. In most cases this technology has to be estimated by either SFA or DEA, and the Malmquist index will not be regarded as transparent unless the procedure of defining the technology is transparent as well. This is true for satisfying the criteria of providing a balanced measurements and being non-corruptible as well. In cases where the technology is well-defined, the ease of calculating the index yields high attractivity and provide an understandable representation of productivity change.

The assumption of a global crs technology are likely to be regarded as an important limitation of the technique, unless a DEA CCR model is used for estimating the distance function or a crs technology is given. For vrs technologies, this limitation will violate the credibility criterion and the Malmquist index should not be applied.

4.4.2 Hicks-Moorsteen TFP index

The Hicks-Moorsteen TFP index (HMTFP) was developed by Diewert (1992) based on the work of Hicks (1961) and Moorsteen (1961). The TFP index measure growth in output over growth in input by using output and input quantity index numbers:

$$\text{HMTFP Index} = \frac{\text{Growth in output}}{\text{Growth in input}} = \frac{\text{Output quantity index}}{\text{Input quantity index}} \tag{4.25}$$

From the formula we can see that we have to select measures for growth in input and growth in output in order to calculate the HM TFP index. In his original paper, Diewert suggested using the Malmquist output quantity index divided by a Malmquist input quantity index. This would include both input and output distance functions, making the index simultaneously oriented. When the input and output growth is given this simple approach is easy to calculate, but it does not distinguish the source of productivity growth (TC or TEC). This index has also proved to be well-defined under general assumptions of vrs and strong disposability[8], but has scarcely been empirically applied.

In a recent paper by Epure et al. (2011) the authors adopt the HMTFP index proposed by Bjurek (1996), and tailor it for specific benchmarking perspectives: static (1), fixed base and unit (2), and dynamic TFP change (3). The static representation of HMTPF may be expressed as follows:

$$HMTFP_{st,t} = \frac{d_o^t(x_t^B, y_t)/d_o^t(x_t^B, y_t^B)}{d_i^t(x_t, y_t^B)/d_i^t(x_t^B, y_t^B)} \tag{4.26}$$

where t is period for analysis and (x_t, y_t) is production mix for analyzed DMU and (x_t^B, y_t^B) is production mix for a set benchmark. This measure is useful if the environment is stable, but the index does not include a time component.

If we want to track changes between periods, we may define a base year dynamic index:

$$HMTFP_{fb,k} = \frac{d_o^k(x_k^B, y_t)/d_o^k(x_k^B, y_k^B)}{d_i^k(x_t, y_k^B)/d_i^k(x_k^B, y_k^B)} \tag{4.27}$$

where k is base year and t is the year under analysis. From this representation we can see that the reference technology and benchmark are fixed, which makes it possible to track movements over time for a given DMU. However, we might argue against a fixed reference technology because the real world evolve through technical progress[9] and a static benchmark might become obsolete.

The dynamic HMTFP is similar the basic HMTFP index by Bjurek, where both years and reference units are dynamic:

$$HMTFP_s = \frac{d_o^s(x_s, y_t)/d_o^s(x_s, y_s)}{d_i^s(x_t, y_s)/d_i^s(x_s, y_s)} \tag{4.28}$$

values greater than 1 gives the DMUs improvement of TFP, while values below 1 indicate a TFP decrease.

When calculating the HMTFP index, the most applied technique is to couple with DEA and solve four linear programs (one for each distance function). This calculation is made easy through the DPIN software developed by O'Donnell (2010a), which also decomposes the index into sources of TFP change.

[8]No increase in inputs can lead to a reduction in outputs, and no reduction in outputs becomes infeasible without an increase in inputs.

[9]Expansion in the production possibilities set.

Examples of use
When searching for “Hicks-Moorsteen” on Google Scholar only 82 hits turned up. Most of the articles were focusing on the favorable characteristics of HM compared to Malmquist and only a few applications were found. Among these, one was measuring agricultural productivity (O’Donnell 2010b) while another measured productivity Change of UK airports (Barros and Managi 2008).

Pros and cons
Like the Malmquist index, the Hicks-Moorsteen index relies on either a known or well-estimated technology to provide useful results. This technology might exhibit crs as well as vrs. The work of Epure et al. has provided multiple references for productivity change, which increases the applicability of the HM index. The HM index is also a close representation of equation (4.19).

The downside to the HM index in light of an incentive system is the increased computational effort related to estimating four distance functions. These distance functions might also be infeasible when coupling HMTFP with DEA and in such cases one should either apply a DEA CCR-model, use SFA distance functions or adapt a continuous time index.

Use in incentive systems
As the literature on HM index is sparse, articles on its applicability in incentive systems seems to be non-existing.

Like the Malquist index, the transparency of the HM index depends on the transparency of the procedure for defining the distance functions, likewise for regarding the technique as balanced and non-corruptible. As the distance functions are well-defined, the HM index provides an unbiased estimation of productivity change, with no known violations of the credibility criterion. In terms of attractivity, the formulation in equations (4.26) - (4.28) are slightly more complicated than equation (4.19).

4.4.3 Index Numbers

Index numbers are defined as real numbers and are the most commonly used instruments to measure changes in productivity and other economic variables. Index numbers are foremost used for measuring changes in total factor productivity, but also for handling panel data sets such as price data over time and space. In order to do so, we have to select a base period which becomes the standard for which all other periods are measured against. A DMU can also be used as a base while handling cross-sectional data. The general index number problem can be formulated as (Battese et al. 2005):

$$V_{st} = \frac{\sum_{m=1}^{M} p_{mt} y_{mt}}{\sum_{m=1}^{M} p_{ms} y_{ms}} \tag{4.29}$$

where

m : number of commodities

M : the commodity to consider

j : j-th period or DMU ($j = s, t$)

y_{mj} : quantity of commodity m , period j

p_{mj} : price of commodity m , period j

V_{st} : Value of the basket of quantities, (M), from periods s to t

The most well known indexes are the Laspeyers and Paasche index, Fisher index (geometric mean of Laspeyers and Paasche) and Tornquist index. The main difference between these indexes is which period the index use as base period, current or past. Apart from this, all of the mentioned techniques are multiplicative index numbers, and the reader is referred to Diewert (2005) for a detailed discussion on additive index numbers. In short, additive index numbers are defined in forms of differences and not ratios.

Examples of Use

Consumer price index (CPI) is perhaps the most widely used economic indicator. In the calculation of CPI, Laspeyers, Paasche and Fisher are perhaps the most common indexes. These indexes might yield different results and figure 4.9 display such differences for New Zealand consumer prices.

Pros and Cons

The indexes treated in this section are well-established in economical literature and in practice. They common techniques for estimating the magnitude of economic changes over time, but rely on both price and quantitative panel data and might be difficult when indexing non-financial metrics. As the indexes are straightforward to compute when data is obtainable, the most important aspect is measurement

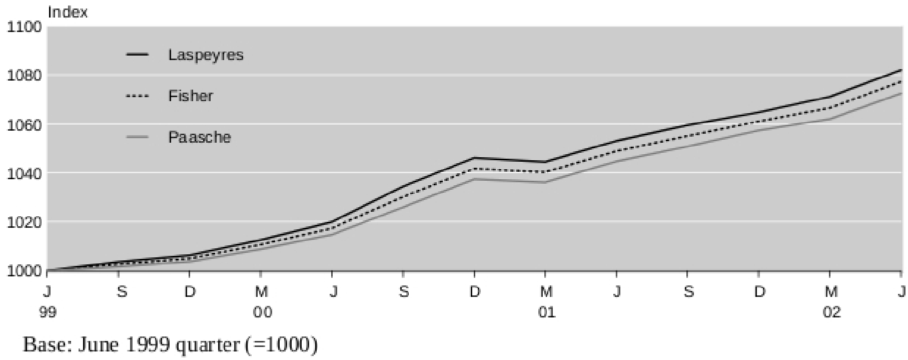

Figure 4.9: Alternative CPI indexes (Smedes 2005)

consistency in cross-sectional or panel data.

Use in Incentive Systems
While multiplicative index numbers are suitable for measuring economic change, they are poor for handling non-financial metrics. Index numbers share the same attributes as KPIs, except the ability to handle input- or output data without price information. One way to account for unavailable price information is to obtain estimates by subjective evaluation. This said, index numbers are most appropriate when DMUs are assumed to be profit maximizers.

4.4.4 Proposed Technique for Estimating Productivity Change

The Malmquist index in equation 4.23 and HM are identical if the technology exhibits global CRS and is inversely homothetic (Färe et al. 1996; for formal definition on homotheticity and proof). These necessary conditions seldom hold in practice and the indexes will in most cases yield different results. The Malmquist index seems to be far more applied than the HM index, both in theory an applied research. Some reason might be that until 2-3 years ago the Hicks-Moorsteen was not available in software and was not decomposed in TFP change. Recent papers have shed light over limitations of the Malmquist index as a TFP index, and state that the Malmquist index might be infeasible under certain assumptions while the Hicks-Moorsteen is not (Kerstens et al. 2010). O'Donnell (2010b) also questions the Malquists productivity index properties as a TFP index and state that it *"cannot in general be expressed as the ratio of an output quantity index to an input quantity index."* [p.4].

Apart from infeasibility issues the Malmquist index does not handle VRS, which makes it less applicable than the Hicks-Moorsteen index. The Hicks-Moorsten index is therefore the preferred productivity index under these circumstances. Index numbers are not considered to be appropriate for overall performance evaluation but might be exploited as possible parameters for financial performance.

Chapter 5

Proposed Evaluation Scheme

5.1 Summary of Reviewed Techniques

In the previous chapters, common practice for performance evaluation in incentive systems in addition to measures of productivity change and efficiency are treated. The techniques possess different attributes and have different strengths and weaknesses. Although a technique is favorable in one situation it is not necessarily favorable in another situation. In this section, common practice will be compared and evaluated against techniques for technical efficiency and productivity.

Most of the reviewed techniques may be divided into deterministic or stochastic and parametric or non-parametric methods. The deterministic methods were KPIs, Balanced Scorecard (composite measures), index numbers and DEA, while SFA was the only stochastic technique. Deterministic methods are suitable when measurement errors are either systematic (equally biased for all DMUs) or negligible. Measurement errors will on the other hand impose uncontrollable risk on the agent, which should be adjusted for. SFA might be more suited for handling noisy data as it contains a stochastic error term, but the end user has little control over how this term affect uncontrollable risk. In fact, uncontrollable risk might be generated in the model, especially if data of high quality is used. Balanced Scorecard and SFA, and in some cases KPIs and index numbers, are parametric methods and the principal have to be considerate regarding how the calculation is carried out. DEA is the only non-parametric method reviewed in this thesis and yield greater flexibility. Hicks-Moorsteen and Malmquist might belong to different classes depending on which technique for estimating technical efficiency is used.

Apart from SFA, none of the techniques adjusts for uncontrollable risk or heterogeneity in their original form and should be coupled with regression analysis in order to detect differences in operational environment. Productivity change measured by HMTFP might on the other hand adjust for heterogeneity if market fluctuations and heterogeneity is systematic.

Subjective evaluation does not fall into any of these categories, and might be thought of as a unique technique. This technique is crucial when mostly subjective data is used, and let the principal adjust for heterogeneity and observed uncontrollable risk. The credibility of such evaluations is on the other hand highly dependent on the managers ability to stay unaffected by the "halo effect" and deteriorated personal preferences. As this is difficult in practice, one should stress to implement objective measures and rather use subjective evaluation to evaluate distortion caused by imperfect objective performance measures.

	Atr.	Cred.	Balanced	Non-corruptible	Transaction cost
Subjective	2	2*	3	2*	1
KPI	3	2*	1	1	3
BCS, CM	3	2*	3	2*	2
DEA	2	2	3	3	2
SFA	1	2	3	2	1

*Table 5.1: The higher score, the better. *: Relies on the principal*

Of the common techniques, composite measures such as BCS is the preferred method for assessing overall performance when a pure objective evaluation is stressed. The fixed weighting of parameters makes the technique more corruptible than DEA, which is the preferred method for performance evaluation. Note however that this is true if the principal manage to communicate the method to the agents, where it is made simple and understandable. Nevertheless, none of the techniques seem perfect on all criteria and should be complemented with subjective evaluations.

5.2 Proposed Scheme for Performance Evaluation

In the past chapters, DEA and HMTFP index have been considered to possess the most preferred properties for implementation in incentive systems. In this section performance evaluation scheme is proposed, in addition to providing some additional insight on DEA and HMTFP.

Even though no technique is perfect, the most optimal way to set up an evaluation scheme seems to be:

1. Select appropriate parameters
The quality and appropriateness of data used in the calculations are just as important as the techniques themselves. As for the approach to select appropriate measures I will suggest the following steps:

(a) Identify possible parameters

(b) Analyze parameters

(c) Aggregate and prepare data for DEA

Identify possible parameters
For a guide on identifying possible parameters, the reader should revisit section 2.1. At this step one should focus on identifying as many possible parameters as possible and use next step to evaluate their appropriateness. Brainstorming and looking into parameters used by comparative businesses facilitates the process.

Analyze parameters
Identified parameters should satisfy criteria in section 2.3.1. Analyze if the parameters might induce employees to dysfunctional responses and if they capture value adding actions. When coupling DEA with HM, both panel and cross sectional data are required, so that we should pay attention to measurement consistency through units and time. Measurement errors might lead to outliers in DEA that affect the performance evaluation of other DMUs. Pay attention to aspects that might result in heterogeneity.

Aggregate and prepare data for DEA
In an article "Preparing your data for DEA" Sarkis (2007) treat important issues related to data characteristics for DEA calculation: Reduction of dataset, imbalance in data magnitudes, negative numbers and zero values, and missing data. As the discriminating power of DEA decrease with the number of parameters the

number of DMUs, n, should at least be $n \geq max\{m * s, 3(m + s)\}$. One way is to aggregate measures[1] or eliminate correlated data, but the reader should also look into principal component analysis. In some DEA software, round-off errors might occur if we have a large imbalance in the dataset. In this case, mean-normalizing the parameters will solve the problem (divide each observation of a parameter by the sample mean). The basic DEA model does not cope with negative or zero values, but multiple solutions to such occurrences are summarized by Sarkis. In case of missing values one should preferably select other parameters or seek good estimates.

2. Estimate efficiency scores (DEA)

As discussed earlier an output-oriented or input-oriented DEA-BCC model with super-efficiency should be implemented. Assurance regions should be added as additional constraints in the multiplier form, eliminating unwanted efficient production mixes. The assurance regions will ensure that effort is spread out on all parameters, providing a balanced measurement. Assurance regions will also prevent extreme specializations for the super efficiency model. The estimation procedure limit subjective evaluations to a minimum, and yields a credible calculation with a low transaction cost.

As the super-efficiency model might result in an infeasible model, we should adapt the work of Cook et al. (2008). They propose a new model where super-efficiency is calculated as the minimum movement in both input- and output direction needed to reach the frontier generated by the remaining DMUs. This new proposal is modeled for both input- and output-orientation, and provides similar results as the original model when it yields a feasible result.

There are developed many softwares which automatically calculate the efficiency scores when the parameters are given. Many of these let the user choose between a variety of extension and are quite intuitive. In order to gain full access, most of these programs must be paid for, especially the most sophisticated versions. See Hollingsworth (2004) for a systematic review of available programs.

3. Adjust for heterogeneity and uncontrollable risk

In extension of the original DEA-literature, four well-established methods that adjust for heterogeneity are suggested. Each of these methods have different applicability, where two of the methods adjust the reference set (Banker and Morey

[1]Consider Hicks and Leontief conditions for aggregation of goods or services.

1986, Charnes et al. 1981). This adjustment lets us determine which DMUs a unit should be measured against. E.g. post offices operating in rural areas are likely to have an uneven flow of customers compared to central offices. By manipulating the reference sets, post offices in rural areas are compared against each other, while central offices are compared against all DMUs. In the paper by Banker and Morey, they proposed to include favorable or unfavorable variables as variables that cannot be reduced or increased (non-discretionary variables). The first two approaches are good if we do not have data for environmental factors or that we have a clear understanding about which reference set a DMU belongs to.

In all other cases, we should adjust for uncontrollable risk and heterogeneity by regressing environmental factors against efficiency scores in 2. In this analysis, the signs of the coefficients of exogenous variables indicate the direction of influence, but p-values should be reported in addition to a subjective evaluation to detect possible arbitrariness. The efficiency scores are then adjusted for significant impacts, resulting in less biased scores. This method can account for many variables and is easy to calculate.

By adjusting for heterogeneity under peer evaluation the estimation becomes more just as differences in operational environments are accounted for. By adjusting the scores for uncontrollable risk, employees becomes more willing to spend effort on risky tasks and the principal may adopt a higher degree of output-based pay.

4. Calculate productivity change

In order to calculate the HM-index, one might apply the DPIN software. This program uses DEA to calculate and decompose the productivity change, but cannot handle assurance regions or super-efficiency. Super-efficiency does not alter the shape of the efficient frontier, but assurance regions put restrictions on efficient production mixes. For technological consistency, we should include the additional restrictions from the DEA-model to the DEA-HM calculation. For each DMU we have to solve four linear programs, which might turn out to be infeasible because $DMU_k(x_t, y_{t+1})$ and $DMU_k(x_{t+1}, y_t)$ does not exist in the reference set I for period t (same reason as for infeasibility in the super-efficiency model).

When infeasibility occurs, the productivity index should be discarded and we might adapt the DEA window analysis instead. By adapting a two year rolling window we can measure technical efficiency change from one period to the other, but this technique will not yield a transitive chain index between periods. If we on the other hand keep adding periods to the reference set and re-estimate the scores for all periods, we get a transitive index. This is an easy but computational expensive technique, which provides good estimates when no technology change has occurred between the periods.

5. Verify results by expert evaluation
Apart from selecting the appropriate parameters and determine the assurance region, all steps above might be implemented as an automatic algorithm. This would yield a pure deterministic calculation, which relies on high quality data free from measurement errors. As measurement errors might exist and the deterministic calculation does not evaluate whether employees has altered the test scores, the final results should be supplemented with an expert evaluation. The objective of this evaluation is to validate the model's fulfillment of criteria in section 2.3.2.

5.2.1 A Few Remarks

At first glance, the proposed evaluation scheme might seem complex and is perhaps not as straight forward as e.g. Balanced Scorecard. As noted by Sexton et al. in the North Carolina buss case who implemented DEA for incentive purposes: *"Success depend on our ability to communicate the methodology"* (pg. 89). While this is true for implementing any performance evaluation scheme, it is even more important when introducing an unfamiliar technique. If the scheme is not accepted by its end users, it will not constitute a successful incentive system even though the proposed scheme is thought to be superior to common evaluation schemes. To ease the acceptance by an organization, an intuitive software facilitating all steps should be developed. An user interface should let the principal determine the assurance regions and give guidance on the choice between orientations and returns-to-scale. The output from such a software should be a ranking of units and their performance score based on either technical efficiency or productivity change, in addition to a report on how environmental factors might affect the scores together with a sensitivity analysis. The report on environmental factors may be used to adjust the final scores, while the sensitivity analysis may give insight on how measurement errors affects the ranking. To increase the employees willingness to participate in the incentive system the strengths of the evaluation scheme compared to common schemes should emphasize on the reduced bias and inaccuracy. By providing an intuitive explanation of the methodology, the scheme will be more transparent thus increasing user acceptance. E.g. *"By measuring your performance on multiple tasks, you will be evaluated against other employees and your own previous performance. Your performance on each task are automatically combined into an overall assessment by a computer program, which will make you look as good as possible compared to your peers. Differences in the operational environment amongst employees will also be accounted for, thus providing a fair and reliable assessment of everyone's performance. Top management will ensure that the calculation is carried out correctly and adjust performances for factors that may not be captured by objective measures."*

Chapter 6

Conclusion and Future Work

6.1 Concluding Remarks

This thesis was introduced by noting that biased and inaccurate performance evaluation reduces the productivity of an organization. The objective of capturing value adding actions by a performance evaluation scheme is challenged by the necessity of balancing multiple measures, while being attractive and credible and mitigate strategic responses. In the literature review of agency theory, sources to strategic behavior in incentive systems were identified, focusing on dysfunctional behavior and risk. On the basis of these findings, a performance evaluation technique should be attractive, credible and non-corruptible, while balancing multiple measures at a low transaction cost.

Common techniques for performance evaluation were argued to violate at least one of the criteria, although they scored high on others. While subjective evaluation and composite measures are good at balancing multiple measures, KPIs are not. KPIs have on the other hand low transaction cost, but are highly corruptible as they are unsuitable for handling the multitasking problem. In fact, all common techniques were evaluated to be corruptible to some extent. While KPIs fail to mitigate dysfunctional responses due to the multitasking problem, subjective evaluations give rise to unproductive rent seeking as workers seek to influence the evaluation process.

Technical efficiency and productivity have scarcely been applied in incentive systems, and an explicit evaluation of their applicability for evaluating groups or individuals seems to be non-existing.

On the conceptual level, technical efficiency and productivity might cope with strategic responses through relative performance evaluation in time and units. Relative performance evaluation is more likely to eliminate unwanted Nash-equilibrium and prevent rent seeking, so that the effectiveness of an incentive system is optimized. While productivity change in most cases gives a strong incentive to improve, empirical estimated technical efficiency might provide weak incentives for the efficient agents. The choice of base period might also affect the responses of agents, thus giving rise to moral hazard due to superior knowledge about systematic market fluctuations.

While both DEA and SFA are common techniques for an empirical estimation of the production frontier, SFA seems less applicable in incentive systems. The determination of an explicit functional form yield low attractivity and the stochastic component might induce uncontrollable risk on the agent. Its ability to handle stochastic data is out-weighted by its rather low credibility as the technique have reported to yield infeasible results and contradict production theory in applied research. DEA is regarded to be a more suitable technique and receive high scores on all criteria. As proposed in section 5.2, a DEA-BCC super efficiency model with assurance regions seems to be the optimal evaluation technique in general. The automatic weighting of parameters reduce transaction cost and reduce lobbying which might arise in composite measures, while assurance regions eliminate unwanted efficient production mixes. The super efficiency model prevent weak incentives for efficient agents but might result in infeasible models, in such cases normal efficiency scores should be estimated.

While evaluating techniques for measuring productivity change, the Hicks-Moorsteen was identified as the most promising technique. The Malmquist index fail under technologies exhibiting globally vrs, and the geometric average does not yield transitive productivity scores. The less widespread HM index is more promising as it cope with technologies exhibiting vrs, but might be infeasible when coupled with DEA and is twice as computational expensive as the Malmquist index.

Productivity measures might cope with environmental factors if heterogeneity is systematic, but uncontrollable risk and heterogeneity should be accounted for explicitly. By regressing the efficiency scores from DEA against environmental factors and adjust for significant impacts, risk is reduced. By adjusting the reference set, we might also choose which units to be compared and further reduce heterogeneity between units.

From the observations above, this conclusion will be ended by noting that the potential for technical efficiency and productivity in incentive systems seem promising from a theoretical point of view. The limited occurrence in reward systems seems to be caused by weak links between TE and productivity and agency theory on a group- or individual level rather than any lack of suitability. Bias and inaccuracy in performance measures are reduced through objective calculations while subjective evaluations are minimized. Nevertheless, the results from these techniques should be verified by expert evaluation as this will mitigate incentive distortion caused by imperfect objective measures. The success of the proposed technique is also highly dependent on the evaluator's ability to communicate the methodology to its employees.

6.2 Further Work

The argumentation for why technical efficiency and productivity are suitable methods for performance evaluation in incentive systems is anchored in theory. Because empirical studies of the applicability of technical efficiency and productivity is lacking, the techniques should be investigated by applied research. The main focus of such studies should be to explore the attractivity and transparency through in-depth interviews, as communicability and user acceptance are believed to be the main threats.

Other techniques for estimating technical efficiency should also be evaluated, such as the StoNED method (Stochastic Non-parametric Envelopment of Data, `http://www.nomepre.net/stoned/`). This technique seeks to combine the virtues of both DEA and SFA, but has not received any attention in this thesis as it is under development.

Measures for productivity change in continuous time should also be explored, as infeasibility will not occur for such indexes. Although some work has been done on developing continuous time DEA models, no software exist for implementing the methodology. Further work should therefore also be directed against a development of commercial software for empirical applications of continuous time models.

Appendix A

Examples of KPIs

	Typical indicators	
Level	Category	Examples
Global	Political	Number and intensity of conflicts
	Economic	Proportion of population below poverty line, Degree of income inequality
	Environmental	Degree and trends in types of pollution and strategic natural resources depletion
	Political	Quality of democracy
	Social	Literacy, Quality of education and Health treatment of minority groups
National	Economic	GDP, National deficits, Inflation, Degree of income inequality, Proportion of population below poverty line
	Environmental	Existing levels of pollution
	Financial	Return on Investment (ROI), Economic Value Added (EVA)
Organizational	Competitive	Market share trend, Ratio of new to total products, Level of core competencies
	People	Employment stability level and turnover, Employee empowerment

	Environmental	Environmental policies (IS0-14000)
Process	Quality	Per cent defect-free output, % Returns, Service time, Customer complaints
	Productivity	Employee-, Material-, Energy-productivity indicators
	Cycle time	order- processing tile, set-up change time
Work teams and individuals	Cohesiveness	Team spirit
	Employee satisfaction	Degree of multi-skill training, Suggestions for improvements
	Loyalty	Employee turnover

(Dervitsiotis 2004)

Appendix B

WinBUGS code

Example code for $N = 167$ DMUs with one input and 7 outputs. The functional form is Cobb-Douglas and the probability density function is half-normal.

Code:

```
model
{
for (i in 1:N) {
u[i] ~ dnorm.trunc0(0,lambda)I(0,1000)
eff[i] <- exp(- u[i])
}

for ( k in 1:K ) {
firm[k] <- data[k, p + 1]
mu[k] <- alpha - u[firm[k]] + inprod(beta[1:p], data[k, 1:p]) +
beta[p + 1] * data[k, 1] * data[k, 1]
y[k] ~ dnorm(mu[k], prec)
}

lambda0 <- 1/37.5
lambda ~ dgamma(1, lambda0)

alpha ~ dnorm(0.0, 1.0E-06)
for (i in 1:p+1) {
beta[i] ~ dnorm(0.0, 1.0E-06)
}
prec ~ dgamma(0.001, 0.001)
sigmasq <- 1 / prec
tot <-sum(eff[])/N
}
```

```
list(y=c(...),
data=structure(.Data=c(...), .Dim=c(167,9)),N=167,K=167,p=8)

list(u=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
lambda=10,alpha=-8,prec=1,beta=c(0,0,0,0,0,0,0,0,0))
```

Bibliography

Agrell, P., Bogetoft, P., and Tind, J. (2002). Incentive plans for productive efficiency, innovation and learning. *International Journal of Production Economics*, 78(1):1–11.

Agrell, P., Bogetoft, P., Tind, J., and for Industrial Economics, C. (2000). *Multi-period DEA incentive regulation in electricity distribution*. Centre for Industrial Economics.

Aigner, D. J. and Chu, S. (1968). On estimating the industry production function. *American Economic Review*, 58:826–839.

Aigner, D. J., Knox, C. A. L., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1):21–37.

Albright, M. and Levy, P. (1995). The effects of source credibility and performance rating discrepancy on reactions to multiple raters. *Journal of Applied Social Psychology*, 25(7):577–600.

Alexander, W., Haug, A., and Jaforullah, M. (2007). A two-stage double-bootstrap data envelopment analysis of efficiency differences of new zealand secondary schools. *Journal of Productivity Analysis*, pages 1–12.

Andersen, P. and Petersen, N. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management science*, 39(10):1261–1264.

Bain & Company. Management tools: Balanced scorecard. Accessed: 12.04.2011.

Baker, G. (2002). Distortion and risk in optimal incentive contracts. *Journal of human resources*, 37(4):728–751.

Baker, G., Gibbons, R., and Murphy, K. (1993). Subjective performance measures in optimal incentive contracts.

Baker, G., Jensen, M., and Murphy, K. (1988). Compensation and incentives: Practice vs. theory. *Journal of Finance*, 43(3):593–616.

Balk, B. (2001). Scale efficiency and productivity change. *Journal of Productivity Analysis*, 15(3):159–183.

Banker, R. and Morey, R. (1986). Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research*, 34(4):513–521.

Banker, R. D. (1980). A game theoretic approach to measuring efficiency. *European Journal of Operational Research*, 5(4):262–266.

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

Banker, R. D. and Gifford, J. (1988). A relative efficiency model for the evaluation of public health nurse productivity.

Barr, R., Seiford, L., and Siems, T. (1991). *An envelopment-analysis approach to measuring the managerial quality of banks*. Dept. of Computer Science and Engineering, Southern Methodist University.

Barros, C. and Managi, S. (2008). Productivity change of uk airports: 2000-2005. *Working Papers*.

Bates, J. (1979). Extrinsic reward and intrinsic motivation: A review with implications for the classroom. *Review of Educational Research*, 49(4):557.

Battese, G., Coelli, T., Rao, D., and O'Donnell, C. J. (2005). *An introduction to efficiency and productivity analysis*. Springer, 2 edition.

Berg, N. and Fast, N. (1975). The Lincoln Electric Company. *Harvard Business School Case*, pages 376–028.

Bjurek, H. (1996). The Malmquist total factor productivity index. *The Scandinavian Journal of Economics*, 98(2):303–313.

Bloom, M. and Milkovich, G. (1997). *The relationship between risk, incentive pay, and organizational performance*. Center for Advanced Human Resource Studies, Cornell University, ILR School.

Bogetoft, P. (1994). Incentive efficient production frontiers: an agency perspective on dea. *Management Science*, 40(8):959–968.

Bogetoft, P. (1995). Incentives and productivity measurements. *International Journal of Production Economics*, 39(1-2):67–81.

Bukh, P., Førsund, F., and Berg, S. (1995). *Banking efficiency in the Nordic countries: A four-country Malmquist index analysis*. Norges Bank.

Camanho, A. and Dyson, R. (2006). Data envelopment analysis and malmquist indices for measuring group performance. *Journal of productivity Analysis*, 26(1):35–49.

Caves, D., Christensen, L., and Diewert, W. (1982a). Multilateral comparisons of output, input, and productivity using superlative index numbers. *The economic journal*, 92(365):73–86.

Caves, D., Christensen, L., and Diewert, W. (1982b). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica: Journal of the Econometric Society*, 50(6):1393–1414.

Charnes, A., Clark, C., Cooper, W., and Golany, B. (1984). A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the us air forces. *Annals of Operations Research*, 2(1):95–112.

Charnes, A., Cooper, W., Golany, B., Seiford, L., and Stutz, J. (1985). Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics*, 30(1-2):91–107.

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

Charnes, A., Cooper, W., and Rhodes, E. (1981). Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. *Management Science*, 27(6):668–697.

Clark, P. B. and Wilson, J. Q. (1961). Incentive systems: A theory of organizations. *Administrative Science Quarterly*, 6(2):pp. 129–166.

Coelli, T. (1996). *Assessing the performance of Australian universities using data envelopment analysis*. University of New England, Dept. of Econometrics.

Coelli, T. and Rao, D. (2005). Total factor productivity growth in agriculture: A malmquist index analysis of 93 countries. *Agricultural Economics*, 32:115–134.

Cook, W., Liang, L., Zha, Y., and Zhu, J. (2008). A modified super-efficiency dea model for infeasibility. *Journal of the Operational Research Society*, 60(2):276–281.

Cook, W. and Seiford, L. (2009). Data envelopment analysis (dea)-thirty years on. *European Journal of Operational Research*, 192(1):1–17.

Courty, P., Do Han Kim, G., and Marschke, G. (2008). Curbing Cream-Skimming: Evidence on Enrolment Incentives. *IZA Discussion Paper*, (3909).

Courty, P. and Marschke, G. (2004). An empirical investigation of gaming responses to explicit performance incentives. *Journal of Labor Economics*, 22(1):pp. 23–56.

Cox, R. and Ahrens, D. (2003). Managements perception of key performance indicators for construction. *Journal of Construction Engineering and Management*, 129:142.

Dalkey, N. and Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management science*, pages 458–467.

Debreu, G. (1951). The coefficient of resource allocation. *Econometrica*, 19(3):273–292.

Dervitsiotis, K. (2004). The design of performance measurement systems for management learning. *Total Quality Management & Business Excellence*, 15(4):457–473.

Diewert, W. (1992). Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis*, 3(3):211–248.

Diewert, W. (2005). Index number theory using differences instead of ratios. *The American Journal of Economics and Sociology*, 64(1):311–360.

Dyson, R. and Thanassoulis, E. (1988). Reducing weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, 39(6):563–576.

Eopold Simar, L. and Wilson, P. (2002). Non-parametric tests of returns to scale. *European Journal of Operational Research*, 139:115–132.

Epure, M., Kerstens, K., and Prior, D. (2011). Technology-based total factor productivity and benchmarking: New proposals and an application. *Omega*, In Press, Corrected Proof.

Färe, R., Grosskopf, S., Lindgren, B., and Roos, P. (1992). Productivity changes in swedish pharamacies 1980–1989: A non-parametric malmquist approach. *Journal of Productivity Analysis*, 3:85–101.

Färe, R., Grosskopf, S., and Roos, P. (1996). On two definitions of productivity. *Economics Letters*, 53(3):269–274.

Färe, R. and Primont, D. (1995). *Multi-output production and duality : theory and applications*. Kluwer Academic.

Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3):253–290.

Feng, G. and Serletis, A. (2009). Efficiency and productivity of the us banking industry, 1998-2005: evidence from the fourier cost function satisfying global regularity conditions. *Journal of Applied Econometrics*, 24(1):105–138.

Ferrier, G. and Lovell, C. (1990). Measuring cost efficiency in banking: econometric and linear programming evidence. *Journal of Econometrics*, 46(1-2):229–245.

Folan, P. and Browne, J. (2005). A review of performance measurement: towards performance management. *Computers in Industry*, 56(7):663–680.

Folger, R. and Konovsky, M. (1989). Effects of procedural and distributive justice on reactions to pay raise decisions. *The Academy of Management Journal*, 32(1):115–130.

Gibbons, R. and Murphy, K. (1990). Relative performance evaluation for chief executive officers. *Industrial and Labor Relations Review*, 43(3):30–51.

Gibbs, M., Merchant, K., Van der Stede, W., and Vargus, M. (2009). Performance measure properties and incentive system design. *Industrial Relations: A Journal of Economy and Society*, 48(2):237–264.

Greene, W. H. (1990). A gamma-distributed stochastic frontier model. *Journal of Econometrics*, 46(1-2):141–163.

Grifell-Tatjé, E. and Lovell, C. (1995). A note on the malmquist productivity index. *Economics letters*, 47(2):169–175.

Gundersen, G. B. (2010). Et litteraturstudie innen identifisering av beste praksis - med anvendelse på 167 norske postkontorer. Unpublished paper related to the authors specilization in operation research at NTNU, Norway.

Habib, M. and Ljungqvist, A. (2005). Firm value and managerial incentives: a stochastic frontier approach. *Journal of Business*, 78:2053–2094.

Hicks, J. R. (1961). *Measurement of Capital in Relation to the Measurement of Other Economic Aggregates*, chapter Measurement of Capital in Relation to the Measurement of Other Economic Aggregates. Macmillian.

Hollingsworth, B. (2004). Non parametric efficiency measurement. *The Economic Journal*, 114(496):F307–F311.

Holmstrom, B. and Milgrom, P. (1991a). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7(special issue):24.

Holmstrom, B. and Milgrom, P. (1991b). Regulating trade among agents. *The New institutional economics: a collection of articles from the Journal of institutional and theoretical economics*, 146:335.

Huang, Y. and McLaughlin, C. (1989). Relative efficiency in rural primary health care: an application of data envelopment analysis. *Health Services Research*, 24(2):143.

Ittner, C. and Larcker, D. (2000). Non-financial performance measures: What works and what doesn't. *Financial Times Mastering Management Series, London.*

Ittner, C., Larcker, D., and Meyer, M. (2003). Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. *Accounting Review*, 78(3):725–758.

Jamasb, T. and Pollitt, M. (2001). Benchmarking and regulation of electricity distribution and transmission utilities: Lessons from international experience. *Benchmarking and Regulation of Electricity Distribution and Transmission Utilities: Lessons from International Experience.*

Jensen, M. (1983). Organization theory and methodology. *Accounting Review*, 58(2):319–339.

Jensen, M. and Meckling, W. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360.

Kaplan, R. and David, P. (1992). The balanced scorecard: Measures that drive performance. *Harvard Business Review*, pages 71–79.

Kaplan, S. E., Petersen, M. J., and Samuels, J. A. (2007). Effects of subordinate likeability and balanced scorecard format on performance-related judgments. *Advances in Accounting*, 23:85–111.

Keegan, D., Eiler, R., and Jones, C. (1989). Are your performance measures obsolete? *Management accounting*, 70(12):45–50.

Kerstens, K., Hachem, B. A. M., and de Woestyne, I. V. (2010). Malmquist and hicks-moortseen productivity indices: An empirical comparison focusing on infeasibilities.

Kim, Y. and Schmidt, P. (2000). A review and empirical comparison of bayesian and classical approaches to inference on efficiency levels in stochastic frontier models with panel data. *Journal of Productivity Analysis*, 14:91–118.

Knittel, C. (2002). Alternative regulatory methods and firm efficiency: stochastic frontier evidence from the us electricity industry. *Review of Economics and Statistics*, 84(3):530–540.

Koop, G. (1994). Recent progress in applied bayesian econometrics. *Journal of Economic Surveys*, 8(1):1–34.

Koop, G. and Steel, M. F. (2007). Bayesian analysis of stochastic frontier models. In *A Companion to Theoretical Econometrics*, pages 520–537. Blackwell Publishing Ltd.

Koopmans, T. (1951). Analysis of production as an efficient combination of activities. *Activity analysis of production and allocation*, 36:27–56.

Kornbluth, J. S. H. (1991). Analysing policy effectiveness using cone restricted data envelopment analysis. *The Journal of the Operational Research Society*, 42(12):1097–1104.

Kraft, E. and Tirtiroglu, D. (1998). Bank efficiency in croatia: A stochastic-frontier analysis. *Journal of Comparative Economics*, 26(2):282–300.

Kuenzle, M. (2005). Cost efficiency in network industries: Application of stochastic frontier analysis.

Lau, L. J. (1986). Chapter 26 functional forms in econometric model building. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume 3, pages 1515–1566. Elsevier.

Lee, D. and Ahn, J. (2005). Rewarding knowledge sharing under measurement inaccuracy. *Knowledge management research & practice*, 3(4):229–243.

Lewis, R. (1955). Measuring, Reporting and Appraising Results of Operations with Reference to Goals, Plans and Budgets. *Planning, Managing and Measuring the Business: A case study of management planning and control at General Electric Company.*

Lin, L. and Tseng, L. (2005). Application of DEA and SFA on the measurement of operating efficiencies for 27 international container ports. In *Proceedings of the Eastern Asia Society for Transportation Studies*, volume 5, pages 592–607.

Malmquist, S. (1953). Index numbers and indifference surfaces. *Trabajos de Estadística y de Investigación Operativa*, 4(2):209–242.

Marschke, G. and Courty, P. (2003). Dynamics of performance measurement systems. *Oxford Review of Economic Policy.*

Matsushima, H. (2010). Role of relative and absolute performance evaluations in intergroup competition. *Japanese Economic Review.*

Meeusen, W. and van Den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. 18(2):435–444.

Moorsteen, R. (1961). On measuring productive potential and relative efficiency. *The Quarterly Journal of Economics*, pages 451–467.

Mortimer, D. (2002). Competing methods for efficiency measurement: A systematic review of direct dea vs sfa. *DFA Comparisons, Centre for Health Program Evaluation Working Paper.*

Nankervis, A. and Compton, R. (2006). Performance management: Theory in practice? *Asia Pacific Journal of Human Resources*, 44(1):83.

Neely, A. (2002). *Business performance measurement: theory and practice.* Cambridge Univ Pr.

Nghiem, H. and Coelli, T. (2002). The effect of incentive reforms upon productivity: evidence from the vietnamese rice industry. *Journal of Development Studies*, 39(1):74–93.

O'Donnell, C. (2010a). Dpin version 1.0: a program for decomposing productivity index numbers. *Centre for Efficiency and Productivity Analysis Working Papers WP01/2010. University of Queensland, Queensland.*

O'Donnell, C. (2010b). Measuring and decomposing agricultural productivity and profitability change. *Australian Journal of Agricultural and Resource Economics*, 54(4):527–560.

Parkin, D. and Hollingsworth, B. (1997). Measuring production efficiency of acute hospitals in scotland, 1991-1994: Validity issues in data envelopment analysis. *Applied Economics*, 29(11):1425–1434.

Peyrache, A. (2010). Multilateral productivity comparisons and homotheticity. *CEPA Working Papers Series.*

Pinto, J. K. and Slevin, D. P. (1987). Critical factors in successful project implementation. *IEEE Transactions of Engineering Management*, EM34(1):22–27.

Prendergast, C. and Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, 37(2-3):355–365.

Rasmusen, E. (2007). *Games and information: An introduction to game theory.* Wiley-Blackwell, 4 edition.

Richmond, J. (1974). Estimating the efficiency of production. *International Economic Review*, 15(2):515–521.

Ridgway, V. (1956). Dysfunctional consequences of performance measurements. *Administrative Science Quarterly*, 1(2):240–247.

Rosko, M. (2001). Cost efficiency of us hospitals: a stochastic frontier approach. *Health Economics*, 10(6):539–551.

Russell, R. (1998). Distance functions in consumer and producer theory. *Index numbers: essays in honor of Sten Malmquist*, page 7.

Sarkis, J. (2007). Preparing your data for dea. *Modeling data irregularities and structural complexities in Data Envelopment Analysis*, pages 305–320.

Seiford, L. and Zhu, J. (1999). An investigation of returns to scale in data envelopment analysis. *Omega*, 27(1):1–11.

Sexton, T. R., Sally, S., and Taggart, R. J. E. (1994). Improving pupil transportation in north carolina. *Interfaces*, 24(1).

Sheldon, T. (1998). Promoting health care quality: what role performance indicators? *Quality in Health Care*, 7:45–50.

Shephard, R. (1953). Cost and production functions. *Princeton University Press.*

Simar, L. and Wilson, P. (2010). Estimation and inference in cross-sectional, stochastic frontier models. *Econometric Reviews*, 24:62–98.

Simon, H. (1976). Administrative behavior. *New York: Free Press*, 3.

Smedes, M. (2005). Measurement issues in the new zealand consumers price index.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):3–23.

Stevens, P. (2005). A stochastic frontier analysis of english and welsh universities. *Education Economics*, 13(4):355–374.

Stevenson, R. E. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics*, 13(1):57–66.

Stiglitz, J. (1987). The design of labor contracts: The economics of incentives and risk sharing. *Incentives, Cooperation, and Risk Sharing: Economic and Psychological Perspectives on Employment Contracts*, pages 47–68.

Stricker, L., Jacobs, P., and Kogan, N. (1974). Trait interrelations in implicit personality theories and questionnaire data. *Journal of Personality and Social Psychology*, 30(2):198–207.

Thorndike, E. (1920). A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29.

Tremblay, M., Sire, B., and Balkin, D. (2000). The role of organizational justice in pay and employee benefit satisfaction, and its effects on work attitudes. *Group & Organization Management*, 25(3):269.

van den Broeck, J., Koop, G., Osiewalski, J., and Steel, M. F. (1994). Stochastic frontier models : A bayesian perspective. *Journal of Econometrics*, 61(2):273–303.

Warburton, A. (1983). Quasiconcave vector maximization: Connectedness of the sets of Pareto-optimal and weak Pareto-optimal alternatives. *Journal of optimization theory and applications*, 40(4):537–557.

Williamson, O. E. (1979). Transaction-Cost Economics: The Governance of Contractual Relations. *The University of Chicago Press*, 22:233–261.

Williamson, O. E. (1985). The economic institutionsof capitalism: Firms, markets, relational contracting. *New York: Free Press.*