MEMORANDUM

No 16/2013

Measuring Efficiency and Effectiveness in the Public Sector

Finn R. Førsund

ISSN: 0809-8786

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.sv.uio.no/econ
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no
e-mail: frisch@frisch.uio.no

Last 10 Memoranda

No 15/13	Mads Greaker and Kristoffer Midttømme *Optimal Environmental Policy with Network Effects: Is Lock-in in Dirty Technologies Possible?*
No 14/13	Diderik Lund and Ragnar Nymoen *Comparative Statistics for Real Options on Oil: What Stylized Facts to Use?*
No 13/13	Nils-Henrik M. von der Fehr *Transparency in Electricity Markets*
No 12/13	Nils Chr. Framstad *When Can Environmental Profile and Emissions Reduction Be Optimized Independently of the Pollutant Level*
No 11/13	Nils Chr. Framstad and Jon Strand *Energy Intensive Infrastructure Investments with Retrofits in Continuous Time: Effects of Uncertainty on Energy Use and Carbon Emissions*
No 10/13	Øystein Kravdal *Reflections on the Search for Fertillity Effects on Happiness*
No 09/13	Erik Biørn and Hild-Marte Bjørnsen *What Motivates Farm Couples to Seek Off-farm Labour? A Logit Analysis of Job Transitions*
No 08/13	Erik Biørn *Identifying Age-Cohort-Time Effects, Their Curvature and Interactions from Polynomials: Examples Related to Sickness Absence*
No 07/13	Alessandro Corsi and Steinar Strøm *The Price Premium for Organic Wines: Estimating a Hedonic Farm-gate Price Equations*
No 06/13	Ingvild Almås and Åshild Auglænd Johnsen *The Cost of Living in China: Implications for Inequality and Poverty*

Previous issues of the memo-series are available in a PDF® format at:
http://www.sv.uio.no/econ/english/research/memorandum/

Measuring efficiency and effectiveness in the public sector*

by

Finn R Førsund

Department of Economics, University of Oslo, and

Norwegian Defence Research Establishment (FFI)

Memo 16/2013-v1

(This version July 2013)

Abstract: The distinction between the concepts outputs and outcomes can be made operational based on the consideration of the degree of control a public service producer has over its production activity. Resources are transformed into service outputs under the control of the organisation in question, while outcomes represent some higher social goals than outputs and are determined by the outputs and other exogenous variables, but the production of outcomes is outside the control of the organisation. The link to the calculation of savings potentials and efficiency measurement is provided based on introducing the concept of a benchmark frontier technology for the type of production in question. A new measure of overall preference effectiveness is introduced and its decomposition into output-oriented technical efficiency and output mix effectiveness is shown. The rather monumental task of providing the necessary information for calculating mix effectiveness is highlighted.

Keywords: Outputs; outcomes; factorially determined multioutput production; Farrell efficiency measures; savings potentials; overall preference effectiveness; output mix effectiveness

JEL classification: D24, H40

* This paper is an extension of Førsund (2012). The paper is written within the research programme "Kostnadseffektiv drift av Forsvaret (KOSTER III)" (Cost Efficiency in Defence) at the Norwegian Defence Research Establishment (FFI). I have benefitted from comments by Espen Berg-Knutsen, Kjell Arne Brekke, Erwin Diewert, Rolf Färe, Torbjørn Hanson, Sverre Kittelsen, Mika Kortelainen, Sverre Kvalvik, Wallace Oates, Jørn Rattsø and Kenneth Løvold Rødseth.

1. Introduction

The core of services produced in the public sector (municipalities and central government) consists of services that in many countries are not sold on markets. Examples of service providers are police, courts, public directorates, educational institutions, nursing homes, hospitals, and defence. These services may be pure public goods and therefore not suitable for provision through markets, like defence services, but typically the services are quasi-public goods that could in principle be provided by markets (e.g., education and health), but due to significant external effects, or distributional importance, it may politically be preferred to provide such services free of charge, or charge well below marginal costs, within the public sector in many countries.

The fact that services are not sold on markets to prices reflecting marginal costs immediately points to the difficulty of assessing if the resources consumed in such activities are used efficiently. There is no automatic check of revenues against costs in the accounts, only budget against expenditure. However, assessment of efficient use of resources if price information on services is lacking, may still be possible if information on the volume of services provided can be collected.

A public sector service provider employs resources of standard types, e.g., capital, labour, energy, materials and services, and transforms these resources into service outputs. We only account as outputs services that are made available to consumers, so intermediate services consumed by the service provider itself are not included. As a background for modelling public service production different types of activities should be recognized. Types may be classified into public producers providing services demanded by other producers or individual consumers, and producers that upheld law and regulations (Dixit, 2002). Another dimension is that the ultimate consumers may be present within the production activity, like students at universities, passengers on mass transit or patients in hospitals, or may be outside the production process like processing of tax returns, and other activities based on paper representation of the consumers like paying out public pensions and supports of various kinds – old-age pension, disability pension, and unemployment benefits – based on rights. We assume that the service provider has control over the transformation of resources into service

outputs. In principle the service outputs are measureable, although the proper treatment of quality may be difficult to capture as quantitative measures. A standard procedure is then to assume the same (unknown) quality level for all providers of the same type of service. This treatment of quality is, of course, not satisfactory. One approach may be to assume a multiplicative decomposition of the service into a quantity part, i.e. number of tax returns processed, and a quality part catching the accuracy of the work. It is reasonable to assume that better quality requires more resources (Chilingerian and Sherman, 1990; Solà and Prior, 2001), i.e., for the same number of tax returns processed more labour has to be used. There is a trade-off between numbers of tax returns processed and the quality of the work. The level of quality then has to be determined. Although this line of reasoning is very interesting to pursue, it will be outside the purpose of the present paper.

There is a well-established literature on measuring efficiency, but not on how to measure effectiveness. The purpose of the paper is to elaborate on the measurement of effectiveness and to show the connection to measurement of efficiency for service providers within the public sector. If we consider a production unit operating in competitive markets both on the input and output side efficiency concepts have been developed to characterise the efficiency of the transformation of inputs into outputs. Farrell (1957) introduced the concepts technical efficiency, allocative efficiency and overall efficiency using data on quantities on inputs, a single output and input prices. Allocative efficiency measures the cost efficiency of allocating inputs in such a way that costs for a given output level is minimised. If a set of products is produced effectiveness is used to characterise the mix of outputs. In the case of a profit-maximising unit the optimal mix is the mix that maximises profit. The well-known rule from production theory will be that the value of the marginal product of a factor for an output should be equal to the factor price, and as to substitution marginal rates of transformation should be equal to the ratio of output prices and marginal rates of substitution equal to ratio of input prices.

For a service provider in the public sector not participating in a competitive market with its outputs the measurement of effectiveness is not so straightforward. There may typically be two stages involved. In the first stage when standard inputs are transformed into *outputs* the classical concept of efficiency applies. At the second stage the service outputs affect individual consumer satisfaction either directly or through the formation of another type of outputs (than the services provided directly) in the form of public goods. Following an established terminology we can then talk about *outcomes* (Burkhead and Hennigan, 1978;

Bruijn, 2002; Schreyer, 2008). [The distinction between outputs and outcomes may originate in the health-economics literature (Schreyer, 2008) and political science and public administration literature (Ruggiero et al., 1995).] A distinction between efficiency and effectiveness found in the literature (see, among several papers, Fitz-Gibbon and Tymms, 2002) is that efficiency is a question of doing things (i.e. outputs) right, and effectiveness is a question of doing the right things (Drucker (1977) may be the first coining this formulation), i.e., producing the outputs that contribute the most to realise the outcomes. We will in the following use classifications where the services produced by the public unit when inputs are transformed into products are termed outputs and the impact of outputs on higher objectives are termed outcomes.

A typical situation is that the service provider is set up to serve more general social objectives than the actual services themselves reflect. Hospitals treat patients of various categories as a part of improving the general health of the public at large. Educational institutions provide education of various types serving a higher goal of contributing to the human capital formation. Labour offices provide training courses in various skills and do job searches for unemployed in order to reduce the rate of unemployment as the final goal. Branches of defence like army, air force and navy produce services to serve higher goals like preserving the peace and guarding the independence of a country. Such higher goals are the reasons for setting up the service-producing units in the first place, and the goals are usually expressed in statements of the intent of providing services. The societal value of providing services is expressed by the success of obtaining the higher goals, or the improvement in indices measuring such goals (see Hatry (1999) for more examples).

Ultimate goals may be lofty. In order to be operational the outcomes must in principle be measureable and be represented by indicators. The distinction between service outputs and outcomes may be fuzzy. In practical politics outcomes may degenerate to service outputs and vice versa. The ultimate goal for higher education may be increase in human capital, but outcomes may also be conceived as the number of candidates with different types of education. One ultimate objective of defence may be to keep the peace, but measurable outcomes may be the upkeep of national sovereignty, national crisis management, participation in international UN peace force operations, and similar more concrete activities, as stated in official Norwegian document concerning the defence sector (Norwegian defence Fact and figures 2010). Another problem with goals of the military is that in general either you have peace or not. Thus this outcome can only take two values. In order to value

effectiveness of outputs it is therefore necessary to develop indicators reflecting zero - one goals by trying to construct continuous indicators that make the higher goals operational. One approach is to construct scenarios for possible conflict situations and find expressions for how different levels and mix of outputs in terms of military capabilities fulfill the higher objectives (see Hanson (2012) for how this can be done for the Norwegian Home Guard).

The purpose of the paper is to explore the implications of the fundamental feature of producing service outputs in order to serve higher social goals. To distinguish between service outputs and outcomes turns out to be crucial for how to approach efficiency and effectiveness measurement. The quest for saving potentials is often the motivation for efficiency studies of the public sector. The paper defines saving potentials and link effectiveness measurement to such potentials. The information needed in order to measure priority or mix effectiveness of outputs is explicitly exposed and how to decompose effectiveness in providing outcomes is developed as a new contribution.

The plan of the paper is as follows. A literature review is presented in Section 2. In Section 3 the two types of production relationships of outputs and outcomes of a service provider is elaborated upon. In Section 4 the concepts of saving potentials and efficiency within service provision is introduced and in Section 5 the concept of effectiveness of outputs in the provision of outcomes is discussed and a decomposition of a Farrell-inspired type of overall effectiveness measure is presented. Section 6 concludes with emphasis on implications for information requirement for efficiency analyses.

2. Literature using outputs and outcomes

An interesting early paper that distinguishes between the services produced and the perception of the services by the consumers is Bradford et al. (1969). Service outputs provided by a public producer are classified as direct outputs (“D-output”), while “the thing or things of primary interest to the citizen-consumer” is termed “C-output” (p. 186). An example of D-outputs of the police is foot- and car patrols within a district and the C-output being the level of safety felt by inhabitants. However, efficiency or effectiveness concepts are not discussed.

Ruggiero (1996a); (1996b) and Duncombe et al. (1997) use the concepts of D-output and C-output in efficiency analyses. However, the distinction between efficiency and effectiveness is not pursued. It is the role of environmental variables as fixed variables together with discretionary inputs that is explored in a one-stage setting using data envelopment analysis (DEA) to calculate efficiency scores.

The basic idea of distinguishing between the outputs of a service provider and the outputs that consumers enjoy is found within transportation economics [Chu et al. (1992); Chiou and Chen (2006); Yu (2008), Yu and Lin (2008); Yu and Fan (2009)], and in studies of efficiency of public libraries [Hammond (2002); De Witte and Geys (2011; 2013)], the latter based on insights from administrative science. There is a distinction between potential service provision (outputs) and the actual services enjoyed by consumers (outcomes). Within transportation outcomes may be defined as the actual use of transportation capacities - bus, metro, railway, and aeroplane - measured by passenger miles or number of passengers transported and ton-miles of freight. Outputs are the *potential* for providing these services, i.e. number of transport units per time unit and seat miles and ton-miles. The objective of transport activities is to transport people and freight, but the transport companies cannot determine the volume (other than the upper limit set by the capacities). Analogously, in the case of public libraries the service output is the *potential* for lending out books, while the outcome is the actual lending. Neither in the transportation references nor in the library references is there any reference to Bradford et al. (1969), and the library papers have no references to the transportation literature, but in De Witte and Geys (2013) there is a list of type of service production with short definitions of service potentials and service delivered for service providers ranging from water utilities to public transport (Table 5, p. 601).

An implication of measuring outcomes this way is that the production of outcomes is uniquely tied to the agency providing the outputs. This may also be the case if the outcome of a teaching institution is measured by the quality of the education measured by the average score of the graduates from the institution, and if outcome of treatments at a hospital is measured by the number of patients that is cured, or the health improvements of its patients.

However, the way the outcome production will be modelled in this paper is more general than found in these strands of literature. A problem with the transportation literature is that service effectiveness is defined as the Farrell technical efficiency measure using outputs as standard inputs and outcomes as the outputs. A similar exercise is done in the library literature. But the

purpose here in measuring effectiveness is to characterise the choice of the output mix under the service provider's control. This is not done by comparing potential services and actual services. It is also the case that investing in capital and setting up capacities is only done based on calculations of demand. It is standard in production theory to regards capacities as capital inputs in a model using realised demand as outputs. Variable factors like labour are not employed for potential production, but for actual production.

In the health economics literature the health status after intervention at the patient level is often termed outcome. But to find the treatment that leads to the best post-intervention status is then a standard efficiency question of picking the optimal treatment. In Färe et al. (2008) a framework for assessing the efficiency of health care, based on DEA models, is introduced distinguishing between the use of standard inputs to produce medical interventions as service outputs and then the outcomes of the interventions as a function of the interventions. The outcomes are defined as the health status of a patient after the interventions and connected to Sen's idea about capabilities to enjoy commodities, i.e. the ability to enjoy the health outcomes. However, the distinction between efficiency and effectiveness is not pursued. A comparison between the pre-operation health status and quality of daily life activities and the post-operation health status and quality of daily life activities is made using a production set termed the capability set relating medical status and daily life activities based on the pre-intervention situation. The border of the production possibility set of outputs transformed into outcomes based on the pre-intervention situation is estimated serving as a benchmark for studying changes after the interventions using directional distance functions.

This is an interesting approach, but here we are more focused on the problem of prioritising between service outputs provided by a specific agency when the higher goals, or indices constructed to cover such goals, can only be influenced by the public service provider through the choice of the mix of its service outputs for given resources.

Solà and Prior (2001) distinguish between efficacy and effectiveness in a study of Catalan hospitals, using dictionary definitions of the terms. Thus, the former term is defined as achievements of targets, while the latter term is defined "as the degree at which production reaches the final targets" (p. 220). However, only efficiency and productivity measures based on DEA efficiency measures are actually computed.

In Medina-Borja and Triantis (2011) and Medina-Borja et al. (2006) a large-scale theoretical and empirical project of evaluating the performance of nonprofit human and social welfare

service organisations is undertaken. A distinction is made between service outputs and customer outcomes (Bradford et al. (1969) is referred to in Medina-Borja and Triantis, 2011). Effectiveness is used in the text to characterise outcome achievements. However, the efficiency for outcome achievement (Table 3, Medina-Borja and Triantis, 2011) is calculated by using a DEA model with service delivery as inputs and outcome achievements as outputs, e.g. there is no explicit question about prioritising between service outputs.

When providing pure public goods it may be the case that the public does not demand the service outputs provided by the agency, but demand the outcomes themselves. An example is the military. The public has preferences for the final outcomes, like keeping the peace, but do not demand troops, exercises, equipment, or the various activities at the service output level. However, there may also be types of public service outcomes that have individual demand, and then our formulation will coincide with C-outputs as defined in Bradford et al. (1969). Individuals demanding service outputs may then be transforming these services into individual outcomes based on a household production function approach a la Becker (1965) or Lancaster (1966). However, we will be thinking more in terms of outcomes as public goods. Estimation cannot then be based on observations of actions of individual consumers.

A typical feature of the relationship between service outputs and outcomes is that this transformation process is not controllable within a specific production activity. The way from output to outcome is a process happening to individuals consuming (or being exposed to) the service, and actions of individuals outside the control of the service provider influence the final outcome.

The concepts of efficiency and effectiveness are used somewhat differently in the literature. A definition of effectiveness found in Cooper and Ijiri (1983), is: “Ability to (a) state and (b) achieve objectives.” This is also stated in Charnes and Cooper (1985, p.71). The concept of outcome may correspond to objectives. However, in the DEA literature objectives have been stated as achieving target levels of outputs and effectiveness used for measuring distance between outputs and target for outputs (Golany and Tamir, 1995), and using effectiveness when imposing weights on outputs (Golany et al., 1993). The latter approach is also followed in Asmild et al. (2007) measuring what is there called effectiveness using more general weight restrictions. We will try to reserve efficiency for doing things right and use effectiveness to characterise doing the right things in a more explicit way.

3. The production relationships of an agency

Let us call a service-producing unit in the public sector for an agency. The multioutput nature of service production can be modelled in several ways, from very general formulations of a transformation function in multiple outputs and multiple inputs to more specialised formulations taking care of technical connections between outputs. Such formulations may involve independent and parallel activities for each output in chains of intermediate deliveries ending in the final service delivered to the consumers, or resources may be shifted around to produce any of the services. Concerning service production in the public sector the main input, at least in terms of current costs, will typically be labour of different qualities. Real capital may represent substantial investments in specialised buildings and machines like in hospitals, but in many cases capital is generic, like office buildings and computers.

Anyway, it seems reasonable to model a great deal of flexibility as regards the possibility of what mix of services to produce given the inputs. Therefore a standard transformation relation between agency outputs y_A and inputs x seems appropriate:

$$F(y_{A_1},..,y_{A_K},x_1,..,x_N;z_F)=0,\ \partial F/\partial y_{A_k}\geq 0,\partial F/\partial x_n\leq 0,\quad k=1,..,K,n=1,..,N \tag{1}$$

There are N types of resources $(x_1,..,x_N)$ and K types of services or outputs produced, (y_{A1},.., y_{AK}). Inputs can freely be allocated to any mix of outputs, implying a maximal *degree of assortment* of outputs (Frisch, 1965).The vector z_F represents variables that influence the relationship between inputs and outputs, but these variables are non-discretionary and will in our analysis be regarded as exogenous for an agency (symbolised by using a semi-colon in front of the vector). Such variables, also termed environmental variables in the literature, may occur for the type of services where the ultimate consumers are present in the production process like students in higher education and patients in hospitals, as mentioned above. Socio-economic background and inherent capabilities may be examples of exogenous variables in higher education.

In the external stage, linking service outputs to the effects on the objectives measured by the outcomes, another type of production relations may be more appropriate. The outcomes cannot be controlled by the agency; it can only observe (in principle) the outcomes due to its production of services (given the values of the exogenous variables). It then seems appropriate to use the special multi-output relationships of a type Frisch (1965) termed *factorially determined multioutput production*. It should be stressed that introducing a production function at this stage is more a conceptual and abstract idea than a description of production relationships that can be considered as well-defined in a technical sense (cf. the statement in Burkhead and Hennigan (1978, p. 35) that "in the public sector there is almost no production function that can be conceptualized with clarity").

There are M final outcomes y_O that are functions of the agency service outputs y_A:

$$y_{O_m} = g_m(y_{A_1},..,y_{A_K};z_{g_m}),\ \frac{\partial g_m}{\partial y_{A_k}} \lesseqgtr 0\ ,m=1,..,M,\ k=1,...,K \tag{2}$$

The outcomes are separable in the sense that each outcome can be expressed as an outcome-specific function with the same set of service outputs as arguments. [The z_{g_m} variables will be explained below.] In Frisch (1965) this is termed *product separation*. It is a special kind of separation in the sense that the degree of assortment is zero, meaning that for given outputs all the outcomes are determined. However, when varying the outputs different proportions of outcomes may be realised. Each outcome has its unique set of isoquants in the common output space. A consequence of the formulation is that the marginal productivity of an output may be positive for one or more outputs, but may be negative for one or more other outputs. It is a question of how the isoquants maps for each output activity in output space are located with respect to each other (see Frisch (1965), p. 272).

Objectives may be of a more general type than the indicators that represent outcomes for how objectives are influenced. In a setting based on indicators it may be the case that outcomes become agency-specific.

In addition to the controllable service outputs y_A we have also opened the possibility of other variables z_{g_m} (interpreted as a vector) influencing the outcomes. The general level of health in the population does not only depend on service output from hospitals, but also on individual characteristics such as smoking, obesity and other lifestyle variables. The formation of human capital does not only depend on the number of exams taken, but also on the quality of students

concerning development after leaving educational institutions. Notice that the exogenous variables z_F and z_{g_m} in the two production functions are not necessarily the same. [In Ruggiero (1996b, footnote 9, p. 501) it is stated that exogenous or environmental variables enter both stages referring to D-outputs and C-outputs introduced in Bradford et al. (1969).] If the general objective is to reduce the occurrence of criminality in the population other factors beside the services provided by the police will influence this. The defence objectives of keeping peace and independence of a country are highly influences by actions of other countries.

We would expect most of the partial productivities in (2) to be positive or zero for a single agency. An example of a negative productivity for an outcome may be the impact on keeping the peace for a country of participating in military actions in other countries. The latter activity may create reactions for instance involving terrorist attacks, as we have seen happening in USA and England.

When negative effects occur there is a conflict between service outputs in achieving objectives measured by outcomes. This occurrence may be relatively rare. The formulation (2) also accommodates the possibility that not all service outputs y_A influence all outcomes y_O (i.e. some partial derivatives may be zero).

The formulations (1) and (2) represent a single agency. However, it may be the case that the service outputs of several agencies influence the same objectives measured by outcomes y_O. It may also be the case that services from one agency having positive effect on the agency's own objectives have negative effects on objectives of other agencies. One example of conflict may be efficiency-related outcomes and outcomes based on distributional objectives.

Introducing S agencies (different types of service providers) the relationship between service outputs and outcomes may be generalised:

$$\begin{aligned} & y_{O_{ms}} = g_{ms}(y_{A_1},..,y_{A_S};z_{g_{ms}}),\ y_{As} = (y_{A_{1s}},..,y_{A_{Ks}}), s = 1,...,S \\ & \frac{\partial y_{O_{ms}}}{\partial y_{O_{ir}}} \lesseqgtr 0\ , m = 1,..,M,\ i = 1,...,K,\ r,s = 1,..,S \end{aligned} \tag{3}$$

where y_{As} is the vector of up to K services produced by agency s. Since we will not pursue this line of modelling a most simple representation is done. Instead of re-labelling the outcomes we keep the M types, but let y_{Oms} be a type of outcome that is specific to agency s that may be

different from the type of outcome y_{Omr}, where r is another agency (some y_{Oms} may be zero). The same is the case for labelling service outputs. The output y_{Aks} for agency s may be a different type than the output y_{Akr} from agency r. (Alternatively, we could say that the types K are given, but that an agency may not produce all types of K service outputs.) An agency has control over its own service outputs, but the outcomes that are the objectives of the agency may be influenced positively or negatively by the service outputs of other agencies. Other exogenous variables outside the agency's control are represented by the vector z_{ms}. These variables are in general specific for the outcome type of the agency, but it may be the case that the same exogenous factor influence different outcomes from different agencies. To keep the notation simple it is not attempted to cover such possibilities formally.

The two stages in defining service production, in the case of a single agency, are illustrated in Figure 1, including our modelling of the transformation processes. The notion of two stages

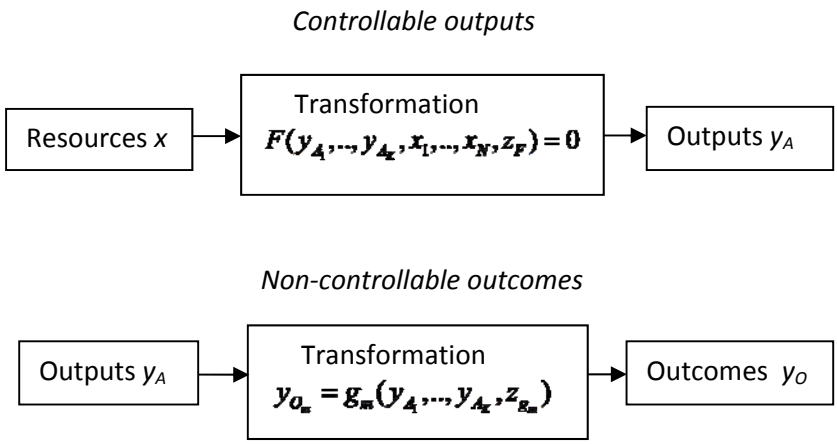

Figure 1. The two types of production

serves the purpose of making clear what kind of efficiency problems we deal with. While service outcomes are produced within a natural time unit there may typically be lags between providing service outputs and when changes in outcomes occur. Some service outcomes are stock variables while service outputs are typically flow concepts. To capture the links between service outputs, exogenous variables and outcomes, more or less complicated and involved dynamic relationships may have to be modelled.

Considering several agencies producing the same type of outputs the arguments in the outcome production functions may either be the sum of the outputs from all agencies

providing the same outputs, or just the outputs provided by the specific agency. In the former case we have

$$y_{O_m} = g_m(\sum_{j=1}^{n} y_{A_{1j}},..,\sum_{j=1}^{n} y_{A_{Kj}}; z_{g_m}),\ m = 1,..,M \qquad (2')$$

where n is the number of agencies. This will be the specification in the case of the outcomes having the character of a public good, like the outcomes for defence.

4. Efficiency of resource use in service output production

Our formulation of two stages implies that an agency cannot control directly how resource use influences outcomes. However, from a social point of view we are interested in efficient use of resources; after all the resources have alternative uses. Therefore we are interested in efficiency at the two stages; efficiency in the resource use of producing service outputs, and in producing the outcomes efficiently. However, the two efficiency considerations are somewhat different. We assume that the resources have well-defined prices q_n (n =1,.., N); inputs are bought in competitive markets. But typically service outputs are not sold in markets, and concerning outcomes they are more or less by definition not priced in any market. Therefore the question whether the right type of service is produced for the relevant outcome may need another type of approach than when studying efficiency in production service outputs. This is a question of effectiveness and will be treated in Section 5

Efficiency for the stage of service production of outputs can then in principle be measured in two ways; cost efficiency and technical efficiency. In the former case cost efficiency can be defined by the programme:

$$\begin{aligned} &\text{Min} \sum_{n=1}^{N} q_n x_n \\ &\text{subject to} \\ &F(y_A, x; z_F) = 0 \\ &y_{A_k} \geq y^o_{A_k} \end{aligned} \qquad (4)$$

where q_n is an input price. We will set up the Lagrangian function in the following way:

$$\begin{aligned} L &= -\sum_{n=1}^{N} q_n x_n \\ &-\gamma F(y_A, x; z_F) \\ &-\sum_{k=1}^{K} \mu_k (-y_{A_k} + y^o_{A_k}) \end{aligned} \tag{5}$$

The necessary first-order conditions are:

$$\begin{aligned} &\frac{\partial L}{\partial x_n} = -q_n - \gamma F'_{x_n} \le 0 \ (= 0 \text{ for } x_n > 0), n = 1, ..., N \\ &\frac{\partial L}{\partial y_{A_k}} = \mu_k - \gamma F'_{y_{Ak}} \le 0 \ (= 0 \text{ for } y_{A_k} > 0) \\ &\mu_k \ge 0 (= 0 \text{ for } y_{A_k} > y^o_{A_k}) \end{aligned} \tag{6}$$

The Lagrangian function is set up in such a way that the unrestricted shadow price γ on the production constraint is positive. Assuming that a resource is used and that the service output we look at is restricted to a positive value, rearranging the first-order conditions for resource x_n and service output y_{Ak} yields:

$$q_n = \mu_k \frac{-F'_{x_n}}{F'_{y_{Ak}}} \tag{7}$$

The ratio of derivatives on the right-hand side is the marginal productivity of resource type *n* in producing service type *k*, and the shadow price μ_k is the shadow price of the increase in the service *k* by employing one more unit of the resource *n*. The shadow price is the increase in costs of wanting one more unit of service output y_{Ak} to be produced. The output is valued at its marginal cost. The basic rule of optimal use of a resource becomes the familiar one: the unit cost should equal the value of the marginal productivity of the resource.

If we assume unique interior solutions for all resources and all services we have that the solution of problem (4) can be expressed by a cost function for the agency

$$C = c(y_{A_1}, .., y_{A_K}, q_1, ..., q_N; z_F) \tag{8}$$

when variations in the given output levels are considered. This cost function is a standard textbook cost function with well-known properties.

Saving potentials

Existence of inefficiency in the provision of the service outputs must mean in our model that the optimal solution to problem (4) is not achieved by the agency in question with a transformation function $F(.)$.

The interest of public organisations like ministries responsible for public service providers in having efficiency studies carried out is invariably to calculate the scope for cost savings. The potential cost savings for the activity of an agency based on the solution to problem (4) as reference or benchmark is

$$Eff_x^{pot} = (\sum_{n=1}^{N} q_n x_n^o - \sum_{n=1}^{N} q_n x_n^*) \,\Big|\, y_{A_k}, z_F \text{ given}, k = 1,..,K \tag{9}$$

Here x_n^o is the observed use of resource n and x_n^* the optimal use of the resource following from the solution to the optimisation problem (4). The efficiency potential for resource saving Eff_x^{pot} is simply the total saving in money for all types of resources.

The measure of cost efficiency is directly related to the saving potential stated in (9). Cost efficiency C^{eff} is defined, conditional on the given output vector y_A and environmental variables z_F, as:

$$C^{eff} = \frac{\sum_{n=1}^{N} q_n x_n^*}{\sum_{n=1}^{N} q_n x_n^o} \,\Big|\, y, z_F \text{ given} \tag{10}$$

This efficiency measure is between zero and one. [Farrell (1957) called this a measure of Overall Efficiency. We will return to this below.]

The optimisation problem (4) is based on given volumes of service production. Therefore there is no endogenous prioritising between the services. This is only possible if the services are priced, either by being sold on market or the prices are derived, as value coefficients, from some preferences over the services, as will be explained in Section 5.

Reasons for inefficiency

The general observation based on the definition of saving potential (9) and cost efficiency (10) is that inefficiency implies that the transformation relation (1) is not realised. The relationship $F(y_A, x; z_F) = 0$ is used as a yardstick or benchmark for the most efficient way to combine inputs to produce given levels of outputs when solving problem (4). Concerning the technology we have to be precise as to the nature of the production function. There are two possibilities: the agency may be evaluated based on the production technology that is actually available to the unit, or based on a benchmark for the most efficient way of producing the type of services in question considering all agencies producing the services. The latter notion of technology is naturally the most interesting one.

In problem (4) the transformation relation $F(y_A, x; z_F) = 0$ can be defined as a frontier production relation, i.e., the notion of efficient/inefficient operations is introduced in the form of a production possibility set *P*:

$$P = \{(y_A, x; z_F) \,|\, x \text{ can produce } y_A \,|\, z_F \text{ given}\} \tag{11}$$

A general property of the production possibility set is that if it is possible to produce a given level of outputs using inputs x', it is also possible to produce the same outputs using inputs x'', where $x' < x''$. Similarly, if a vector y_A of outputs is produced using a given input vector *x*, then it is also possible to produce fewer outputs with the same input vector. We can connect the efficient border of the production possibility set with the transformation relation $F(y_A, x; z_F) = 0$, and denote inefficient points within the production possibility set by $F(y_A, x; z_F) < 0$ (Hanoch (1970).

Inefficiency is defined to be present at our observation if $F(y_A^0, x^0; z_F) < 0$ where $y_A{}^0$, x^0 are the observed vectors of outputs and inputs of an agency. When assessing the efficiency potential using (9) or (10) it is assumed that $F(y_A^0, x^*; z_F) = 0$, where $y_A{}^0$ is the given vector of outputs and x^* the optimal solution to problem (4). The transformation relationship $F(y_A, x; z_F) = 0$ is commonly called the frontier production relationship in the efficiency literature.

However, to point out that measured inefficiency must be due to the frontier technology, or the best technology, not being realised is not an explanation of why inefficiency occurs.

One reason for a positive savings potential may be irrational behaviour; resources are simply wasted. However, this explanation is not so satisfactorily, at least not if something is planned to be done to harvest the savings potential. In order to understand inefficiency in such a way that strategies for improving efficiency can be formulated, a first approach should be to investigate whether inefficiency can be rationalised (for an overview of approaches in the efficiency literature, see Førsund, 2010).

A rational reason for an apparent savings potential is that the objectives of the agency is not one of cost minimisation as in (4). In Leibenstein (1966) the concept of X-efficiency was introduced to explain the occurrence of slacks of different kinds. His main explanations were that lack of competition could motivate management to create slacks (in terms of excessive input usage), and that lack of incentives could lead to workers putting in less effort.

The Leibenstein approach to modelling the inside of the black box has been followed up in the literature developing behavioural theories of the firm. Aspects of agency theory that have been developed are the relation between managers and owners, the managerial effort, and the contractual arrangement determining the effort (Haskel and Sanchis, 1995, p. 301). In Haskel and Sanchis (1995); (2000), worker effort is introduced as a likely determinant of X-inefficiency. Thus, a high X-inefficiency is equated with a low level of effort. Introducing a utility function of workers, $U(w,e)$ ($U'_w > 0, U'_e < 0$) there is a trade-off between effort (e) and wages (w). Workers bargain for wages and slack related to effort levels. Effort levels below the maximal are then the cause of measured inefficiency (Haskel and Sanchis 2000).

Bogetoft and Hougaard (2003) take the idea that slack may occur because it also provides benefit a step further in formalisation by introducing a utility function, $W(\pi, s)$, for the organisation with profit (π) and slack (s) as arguments, both with positive marginal utility. A production unit has a trade-off between on-the-job consumption of slack and off-the-job consumption of profit. In this way it is possible to rationalise inefficiency; there is no need for resorting to e.g. bounded rationality or incomplete contracts.

So both at the worker level and at the management level inefficiencies may be rational within the black box as the workers may seek contracts involving effort and wages simultaneously, and effort level and inefficiency may be negatively related, and managers may seek contracts with owners involving profit and slacks (representing various types of working environment valued by managers), where profit and slacks are both positively valued.

When introducing the notion of a benchmark or frontier technology above it was assumed that such a technology is known to the organisations. But this may not be the case. The organisation may, in fact, not have complete information about the production function. This may be of special relevance for service providers where labour is the major input. A crucial factor is then often the organisation of the service production. It is difficult to formalise the impact of different types of organisations or, indeed, the role of management in running service-type operations.

Going back to Farrell (1957) a reason for inefficiencies is that the inputs are too heterogeneous between organisations regarding quality of labour and capital. For service providers dealing directly with clients (e.g., patients, students) differences in qualitative attributes of the clients may wrongly be classified as inefficiency. A part of labour is management. An inefficient use of resources is often attributed to management and organisation of the activities within the agency. However, in economics it is not common to specify management as an input at all and, as pointed out above, organisation is a variable that is difficult to operationalise in a quantitative framework (Lewin and Minton, 1986).

Connecting efficiency and saving potentials

The fundamental idea of Farrell (1957) was to measure technical efficiency by introducing a scaling factor for either inputs or outputs that would project an observation onto the frontier function. Combining observations of outputs y^o and inputs x^o with the frontier function and introducing a common scaling factor for inputs, we get:

$$F(y_A^0, x^P; z_F) = 0 \text{ where } x^P = E_1 x^0 \tag{12}$$

The scalar measure E_1 is the input-oriented technical efficiency measure introduced in Farrell (1957). We have that $E_1 \in (0,1]$. All the inputs are contracted with the same factor. Since we have that x^P represents a frontier point the efficiency measure must be less than one if $x^0 > x^P$ in all components (assuming a smooth frontier). Furthermore, assuming inputs to be essential factors in the sense that $F(y_A, 0; z_F) \Rightarrow y_A = 0$, the efficiency measure is bounded from below by zero. The value of one characterises an efficient operation. Before stating the cost-minimising problem (4) it was stated that there are two ways of measuring efficiency;

cost efficiency and technical efficiency. Equation (12) provides a way of calculating a technical efficiency measure.

The measure of cost efficiency (10) gives the overall cost-reduction factor. Using the technical efficiency measure E_1 defined in (12) the cost savings of projecting the observation radially to the frontier is:

$$\sum_{n=1}^{N} q_n (x_n^o - x_n^P) = (1 - E_1) \sum_{n=1}^{N} q_n x_n^o \tag{13}$$

Farrell (1957) developed a connection between cost efficiency and technical efficiency as mentioned above. The costs of inputs x^P at the frontier is $\sum_{n=1}^{N} q_n x_n^P$. Farrell's measures of efficiency using costs are:

Technical efficiency (input-oriented): $$E_1 = \frac{\sum_{n=1}^{N} q_n x_n^P}{\sum_{n=1}^{N} q_n x_n^o} \tag{14a}$$

Allocative efficiency (or input-price efficiency): $$AE = \frac{\sum_{n=1}^{N} q_n x_n^*}{\sum_{n=1}^{N} q_n x_n^P} \tag{14b}$$

Overall efficiency: $$OE = \frac{\sum_{n=1}^{N} q_n x_n^*}{\sum_{n=1}^{N} q_n x_n^o} = \underbrace{\frac{\sum_{n=1}^{N} q_n x_n^P}{\sum_{n=1}^{N} q_n x_n^o}}_{E_1} \bullet \underbrace{\frac{\sum_{n=1}^{N} q_n x_n^*}{\sum_{n=1}^{N} q_n x_n^P}}_{AE} \tag{14c}$$

The overall efficiency measure is the same as the cost efficiency measure (10), and from the second expression in (14c) we see that it decomposes multiplicatively into the technical efficiency E_1 and the allocative efficiency measure; $OE = E_1 \bullet AE$.

An output-oriented technical efficiency measure, E_2, can be introduced in a similar way as in (12) using a common scaling of outputs projecting the observation to the frontier:

$$F(y_A^P, x^0; z_F) = 0 \text{ where } y_A^P = \frac{y_A^0}{E_2} \quad (15)$$

[The notation E_1 and E_2 was introduced in Farrell (1957), but there in lowercase letters; uppercase letters were introduced in Farrell and Fieldhouse (1962).] To keep the efficiency measure between zero and one it is defined as the inverse of the scaling factor. But notice that the possibility of calculating such a measure does not mean that we have got around the problem of not having prices or valuation coefficients for the services. We still do not have a measure of priority efficiency for outputs. Such measures require valuation of outputs. The measure E_2 is a scalar between the values zero and one, where the value of one characterises an efficient operation. The measure is not a priority measure for the services, but simply expands each service with the same factor to achieve efficiency.

Estimating the frontier function

The crucial information needed both for calculating the cost efficiency measure and the measures of technical efficiency is to establish the transformation relation representing the efficient technology. Farrell (1957) introduced estimating a piecewise linear frontier, and also discussed estimating parametric frontiers. Linear programming was used to estimate a piecewise linear frontier in Farrell and Fieldhouse (1962) assuming constant returns to scale and a single output. Charnes et al. (1978) generalised to multiple outputs and Afriat (1972) and Banker et al. (1984) generalised to variable returns to scale. Comprehensive treatments of estimating piecewise linear frontier functions can be found in Cooper et al. (2000) and Fried et al. (2009). Seminal paper developing the parametric frontier approach are Aigner and Chu (1968), Afriat (1972) (Afriat (1972) pioneered both the non-parametric approach with variable returns to scale and estimation of parametric frontiers), Aigner et al. (1978) and Meeusen and Broeck (1978). (See Førsund and Sarafoglou (2002) for an historical account.) A comprehensive treatment of parametric frontier function estimation can be found in Kumbhakar and Lovell (2000).

Estimating frontier functions and efficiency measures is at the heart of an expanding research field of efficiency measurement. For cross-section estimation a sufficient number of units are crucial. The estimation procedures are based on the assumption that there are a sufficient number of agencies producing the same type of services.

5. Effectiveness in the provision of outcomes

As mentioned in Section 4 the question whether the right type of service is produced for the relevant outcome may need another type of approach than when studying efficiency in producing service outputs. In the transport and library literature mentioned previously effectiveness is calculated in the same way as efficiency in the production of outputs, just using the production function for outcomes, and the measurement of efficiency and effectiveness is not linked in the way we want to do it here. Both stages portrayed in Figure 1 must be treated simultaneously when we want to measure effectiveness.

When output prices do not reflect consumers' evaluation the introduction of a preference function is necessary in principle in order to be able to prioritise between the outcomes, and thereby enabling a prioritising between outputs (cf. Burkhead and Hennigan (1978, p. 37) stating: "The ultimate objective function – that which is to be maximized – should be described as a social state: are citizens better or worse off as a result of a particular government service delivery?"). We have assumed that the measurable outcomes y_O are related to the ultimate objectives of providing public services. The preference function $W(y_{O_1},..,y_{O_M})(W'_{y_{O_m}} > 0 \forall m)$ is based on this links between the ultimate objectives and the measurable outcomes. [We are looking at a single agency. If a number of agencies producing the same outputs is considered the choice between the specifications (2) and (2') will influence the modelling.] We can then give optimal social planning conditions for priority effectiveness of outputs by assuming a given budget, *B,* for the resources *x* and maximise the value of the preference function:

$$
\begin{aligned}
&Max\ W(y_{O_1},..,y_{O_M}) \\
&\text{subject to} \\
&\sum_{n=1}^{N} q_n x_n \le B \qquad (16) \\
&y_{O_m} = g_m(y_A; z_{g_m}),\ m = 1,..,M \\
&F(y_A, x; z_F) \le 0
\end{aligned}
$$

The variables y_A, y_O, x, z_{g_m} and z_F are interpreted as vectors. It seems reasonable to enter the relations between outcomes and outputs using equalities, because the production relations are autonomous in the sense that the transformation process is not under the control of any agency. For services consumed by individuals the transformation of service outputs to outcomes takes place within the consumers themselves (c.f. household production functions) and can be expressed as an aggregate for relevant groups of consumers. Exogenous variables of type z_{g_m} influencing the process can also act at an individual level, e.g. the state of health of a person treated by the health system may depend on whether the person smokes, and also other lifestyle factors, including exposure to air pollution. So effectiveness is not connected to the production relations $g_m(.)$ not being realised. However, in the case of transforming inputs into outputs it is opened up for the possibility that the benchmark frontier function $F(y_A, x; z_F)$ may not be realised by an agency.

The Lagrangian function for problem (16), inserting the outcome production functions into the preference function for simplification, is

$$\begin{aligned} &L = W(g_1(y_A; z_{g_1}),.., g_M(y_A; z_{g_M})) \\ &-\beta(\sum_{n=1}^{N} q_n x_n - B) \\ &-\gamma F(y_A, x; z_F) \end{aligned} \tag{17}$$

The necessary first-order conditions are

$$\begin{aligned} &\frac{\partial L}{\partial x_n} = -\beta q_n - \gamma F'_{x_n}(y_A, x; z_F) \le 0 (= 0 \text{ for } x_n > 0), n = 1,..., N \\ &\frac{\partial L}{\partial y_{A_k}} = \sum_{m=1}^{M} W'_m g'_{mk}(y_A; z_{g_m}) - \gamma F'_{y_k}(y_A, x; z_F) \le 0 (= 0 \text{ for } y_{A_k} > 0), k = 1,..., K \\ &\beta \ge 0 (= 0 \text{ for } \sum_{n=1}^{N} q_n x_n < B) \\ &\gamma \ge 0 (= 0 \text{ for } F(y_A, x; z_F) < 0) \end{aligned} \tag{18}$$

Assuming an interior solution, to realise the maximal value of the preference function a full utilisation of both the budget and being on the production frontier are necessary. Allocative efficiency of the inputs is implied by the optimality conditions in (18),

$F'_{x_n} / F'_{x_r} = q_n / q_r \, (n, r = 1, ..., N)$ i.e. the marginal rate of substitution between inputs is equal to the factor price ratio. The marginal resource cost is measured as the alternative cost of using resource x_n using the shadow price γ on the transformation function constraint to convert the expression into a measure of preference function units per unit of input. This alternative cost should be equal to the factor price adjusted by the shadow price on the budget, measuring the impact on the preference function of a marginal increase in the budget. So the cost in terms of preference function units of using resource x_n - by crowding out other resources keeping the budget - is set equal to the similarly measured cost in production. The second first-order condition in (18) tells us that the alternative cost of producing a unit more of the service y_{Ak} is set equal to the value created in terms of outcomes, where the value is measured by the marginal impacts on the preference function.

Eliminating the Lagrangian parameter γ for the transformation function yields

$$\sum_{m=1}^{M} W'_m g'_{mk}(y_A; z_{g_m}) \frac{-F'_{x_n}}{F'_{y_{Ak}}} = \beta q_n \, , k = 1, ..., K \, , n = 1, .., N \qquad (19a)$$

The second ratio term on the left-hand side, $(-F'_{x_n} / F'_{y_{Ak}})$ is the marginal productivity of resource x_n in producing service y_{Ak}. The first term is the evaluation of the outcomes generated at the margin by the service y_{Ak}. Using the Frisch system of factorially determined multi-outcome production we have to sum over all the outcomes that are influenced by the marginal change in the service y_{Ak}. The measuring unit on the left-hand side is therefore preference-function units per input unit.

A unique solution to problem (16) implies that a resource x_n is used in such a way to produce a service y_{Ak} that the preference function over outcomes y_O is maximised. The condition (19a) tells us that an optimal use of a resource x_n is characterised by the cost of a unit of the resource, weighted with the shadow price on the budget, being equal to the benefit it creates in terms of an evaluation of the final outcomes y_O through the production of a service y_{Ak}. The shadow price on the budget constraint expresses the increase in the value of the preference function of a unit increase in the total budget. The shadow price β translate from the money unit of the budget B to the units of the preference function. To get an expression for (19a) closer to the standard expression in production theory we can deflate the evaluation of the marginal changes in outcomes with the budget shadow price to get

$$\frac{1}{\beta}\sum_{m=1}^{M} W'_m g'_{mk}(y_A; z_{g_m}) \frac{-F'_{x_n}}{F'_{y_{Ak}}} = q_n, k = 1,...,K, n = 1,..,N \tag{19b}$$

The measuring unit on the left-hand side is now money per unit of resource *n*. The condition tells us that the monetised value created by employing a unit of a resource x_n to produce service y_{Ak} is equal to the unit resource price.

The questions of how to prioritise between services and between outcomes are answered by the simultaneous solution of the endogenous variables in problem (16). The process cannot be separated into two stages of prioritising service outputs and outcomes separately. In Figure 1 two stages are portrayed, but it should be quite clear from the analysis above that these stages are directly interconnected.

The preferences are over the outcomes y_O, but to clarify the implications of priority efficiency for outputs y_A we need to see the implications of preferences for outcomes for the outputs. Considering changes in two outputs y_{Ak} and y_{Al} total differentiation of the preference function yields:

$$\begin{aligned} &\sum_{m=1}^{M} W'_m g'_{mk}(y_A; z_{g_m}) dy_{Ak} + \sum_{m=1}^{M} W'_m g'_{ml}(y_A; z_{g_m}) dy_{Al} = 0 \Rightarrow \\ &\frac{dy_{Al}}{dy_{Ak}} = -\frac{\sum_{m=1}^{M} W'_m g'_{mk}(y_A; z_{g_m})}{\sum_{m=1}^{M} W'_m g'_{ml}(y_A; z_{g_m})}, l, k = 1,...,K \end{aligned} \tag{20}$$

We will call this ratio the *marginal preference rate of substitution* between outputs y_{Ak} and y_{Al}. It combines the preferences for outcomes with the properties of the outputs as arguments in the system of production functions (2) for outcomes. The marginal productivity of an output is weighted with the marginal preference impact for each of the outcomes affected. The measuring unit for the total expression is then in preference-function units per unit of output. The values are conditional on the values of the exogenous variables z_{g_m} (and the budget *B*).

Considering the same pair (*k*, *l*) of outputs, using the second condition in (18) for each output in the pair, the condition for priority efficiency is:

$$\frac{\sum_{m=1}^{M} W'_m g'_{mk}(y_A; z_{g_m})}{\sum_{m=1}^{M} W'_m g'_{ml}(y_A; z_{g_m})} = \frac{F'_{y_{Ak}}(y_A, x; z_F)}{F'_{y_{Al}}(y_A, x; z_F)},\ k, l = 1, ..., K \tag{21}$$

The ratio on the right-hand side of the equality sign is the marginal rate of transformation between the two outputs in the production function (1). The condition for priority- or output mix effectiveness is that the marginal rate of transformation between two outputs y_{Ak} and y_{Al} is equal to the ratio of marginal preference rate of substitution.

Potentials for improvement

It is clear from (18) that a point on the frontier function for resources and outputs must be realised for a maximum of the welfare function to be obtained. Comparing an observation (x^0, $y_A{}^0$, z^0) with an optimal solution of problem (16) the potential increase in the value of the preference function by implementing an optimal solution is:

$$Eff_W^{pot} = [W(y^*_{O_1}, .., y^*_{O_M}) - W(y^0_{O_1}, .., y^0_{O_M})] \big| B \text{ given} \tag{22}$$

The observed input vector x^0 may be different in composition from the optimal input vector x^*, but the total budget is the same by assumption. For fixed input prices the budget B is a linear aggregation of the inputs to an input bundle. However, the marginal productivities of inputs in the function (1), $(-F'_{x_n} / F'_{y_{Ak}})$ will typically be different for the input vectors of the two mixes.

We can reformulate the potential welfare improvement by forming efficiency measures analogous to the Farrell efficiency measures in (14a-c). The overall efficiency of Farrell may here be termed *Overall Preference Effectiveness*, *OPE*, and being measured based on the variables of the relative savings potential (22). The improvement of the value of the preference function comes from two sources: realising the frontier production function by eliminating output inefficiency by proportionally increasing the outputs to $y_A^P = y_A^0 / E_2$ [see (15)], and by changing the output mix of the proportional frontier projection $y_A{}^P$ to the optimal mix of $y_A{}^*$ on the frontier:

$$\underbrace{OPE = \frac{W(y_{O_1}^0,..,y_{O_M}^0)}{W(y_{O_1}^*,..,y_{O_M}^*)}}_{\text{Overall preference effectiveness}} \Big| B = \underbrace{\frac{W(g(y_A^0;z_g)\big|x^0)}{W(g(y_A^P;z_g)\big|x^0)}}_{\text{Technical effficiency}} \bullet \underbrace{\frac{W(g(y_A^P;z_g)\big|x^0)}{W(g(y_A^*;z_g)\big|x^*)}}_{\text{Mix effectiveness}} \tag{23}$$

where $g(.)$ is the vector of M functions $g_m(.)$. It seems reasonable to use the observed inputs x^0, indicated by the notation $\big|x^0$ when the output point is moved to the output production frontier, but using the optimal mix x^* of inputs when changing the output mix to the optimal y_A^*. The first term on the rhs. reflects doing things right, and the second term doing the right things. Overall preference effectiveness not only assumes that efficiency in producing outputs is obtained, but also that effectiveness is achieved by providing the most potent mix of outputs.

The situation can be illustrated looking at a pair of outputs, y_{Ak} and y_{Al}, set out in Figure 2. We

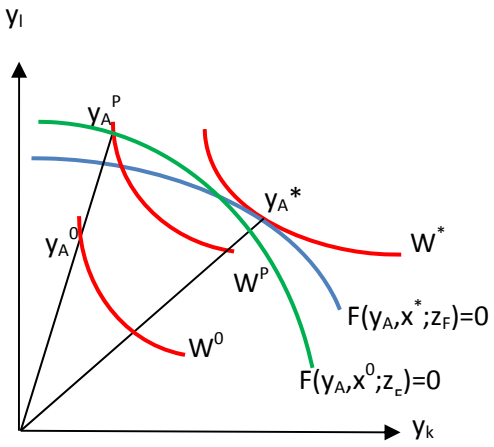

Figure 2. Effectiveness and priority efficiency

have an observation (y_A^0, x^0) and a given expenditure B on inputs. This budget is also kept in the optimisation problem (16) so we have $\sum_{n=1}^{N} q_n x_n^0 = \sum_{n=1}^{N} q_n x_n^* = B$. The transformation between outputs y_{Ak} and y_{Al} at the frontier is shown by the transformation curve labelled $F(y_A, x^0; z_F) = 0$ for the initial bundle x^0 of inputs. The curve labelled W^0 going through the point y_A^0 is not strictly speaking a contour curve of the preference function $W(.)$, but is the curve defined by the marginal preference rate of substitution between outputs y_{Ak} and y_{Al} in (20), in the case of two outputs, letting y_{Ak} and y_{Al} vary in such a way that the value of the preference

function along this curve is fixed at W^0, i.e. the induced changes in outcomes by varying the outputs must be such that the value of the preference function is fixed. A contour curve for the function $W(.)$ will be in the outcome space of y_O, while we are now operating in the output space of y_A.

Moving proportionally to the frontier to eliminate inefficiency in the production of outputs for the given inputs x^0 point y_A^P is realised. But the solution to the optimisation problem (16) for the outputs imply another mix than y_A^P, namely y_A^*. The curve labelled W^P, defined the same way as described above, implicitly determined by the properties of the preference function and the outcome production functions, passing through the frontier output point y_A^P, has a smaller value than the curve W^*, determined from the right-hand side of (21) (by keeping the value of the preference function fixed at W^* but varying y_{Ak} and y_{Al}) being tangential to the frontier with x^* as inputs at point y_A^*. Therefore we move from point y_A^P to point y_A^*. This is the realisation of priority effectiveness. Effectiveness is achieved by producing the optimal mix of outputs yielding the maximal value of the preference function for a given budget. The production possibility sets will differ for different mix of the inputs for a constant budget. Comparing the vectors x^0 and x^* some inputs will decrease, other increase to keep the budget constant. As stated above the marginal productivities of inputs in the function (1) will be different for the two mixes, as illustrated in Figure 2 by the two transformation curves labelled $F(y_A, x^0, z_F)$ and $F(y_A, x^*, z_F)$, respectively. In order to understand Fig. 2 it may help to introduce the concept of *cost-indirect output set* (Shephard, 1974). By construction this set will envelope all production possibility sets spanned by input vectors obeying the budget constraint in (16) (see Färe and Grosskopf, 1994), i.e., the coordinates (y^*, x^*) will always be on the cost-indirect frontier.

However, output efficiency, the first term on the right-hand side in (23) (after the second equality sign), is in general not identical to the output-oriented Farrell technical efficiency measure E_2 defined in (15) for the same data and frontier function. The measures will only coincide if both the preference function is homogeneous of degree 1 in the outcomes and the outcome production functions are homogenous of degree 1 in the outputs:

$$\frac{W(g(y_A^0;z_g)\big|x^0)}{W(g(y_A^P;z_g)\big|x^0)} = \frac{W(g(y_A^0;z_g)\big|x^0)}{W(g(\frac{1}{E_2}y_A^0;z_g)\big|x^0)} = \frac{W(g(y_A^0;z_g)\big|x^0)}{W(\frac{1}{E_2}g(y_A^0;z_g)\big|x^0)} = E_2 \qquad (24)$$

If this is the case the contour curves illustrated in Figure 2 will be radial projections of each other, and the spacing of the contour curves is constant in relative terms. In the transport and library literature mentioned previously what is termed service effectiveness is measured just by the first term on the right-hand side of (23) without using a preference function $W(.)$.

The marginal preference weights in (19a,b) are variable. If it is assumed that the marginal weights are constants, this is equivalent to the preference function $W(.)$ being linearised; $W'_m(y_{O_1},..,y_{O_M}) = w_m \ , m = 1,..,M$. Such constant valuation coefficients may play the role of prices of the outcomes. But notice that such prices relate to implicit prices of outputs in a complicated way involving the production functions (2).

The question of how to construct preference functions for public sector outcomes is a research field in itself. There is a literature focusing on how to construct scalar-valued objective functions for macro-economic decision models. Pioneers were the first joint Nobel Prize winners Frisch and Tinbergen (see the account of the ideas of Frisch of establishing preference functions by interviewing decision-makers in Bjekholt and Strøm, 2002).

An example of linerarising a preference function over outcomes is found in Lauer et al. (2004) based on works of WHO of performance ranking of health systems of 191 member countries. Five outcome variables for the health sector of a country are used; level of population health, inequalities in health, level of responsiveness, inequalities in responsiveness and fairness in financial contributions. The establishment of fixed weights was based on responses to a survey of over 1000 health experts.

In the literature there are examples of just a single outcome (Bradford et al. (1969) mention safety level for the police sector and average scores for schools). Then there is no preference function to be maximised, just the index for the single outcome. However, the problem of prioritising between the outputs remains. The contour curve in Fig. 2 will be an isoquant of the single outcome production function of type (2). In that case the *OPE* measure reduces to an outcome effectiveness measure, *OE*:

$$\underbrace{OE = \frac{y_O^0}{y_O^*}}_{\text{Outcome effectiveness}} \quad \Big| B = \underbrace{\frac{g(y_A^0; z_g)\big|x^0}{g(y_A^P; z_g)\big|x^0}}_{\text{Technical effficiency}} \bullet \underbrace{\frac{g(y_A^P; z_g)\big|x^0}{g(y_A^*; z_g)\big|x^*}}_{\text{Mix effectiveness}} \tag{23'}$$

where the $g(.)$ function is now the production function for a single outcome.

Calculating priority effectiveness is not so simple as calculating cost- or technical efficiency for use of resources to produce outputs. The informational requirement is quite formidable. We must be able to define outcomes in the first place. Then we have to know not only the preference function over outcomes (in the case of more than one outcome), but also how outcomes are influences by service outputs and other exogenous variables. This last task is quite another exercise than determining the transformation function involving resources and service outputs. Diewert (2011), when addressing methods for measuring prices of nonmarket goods, states that the most desirable method is some form of purchaser valuation. A general equilibrium approach for the economy embedding public service outputs is suggested as a way of obtaining user based evaluations. However, he comments that the information required to implement such an approach is "just too great" (p. 181). Thus, the method is declared theoretically sound, but not practical.

Using cost information

If establishing a preference function for outcomes is not possible, the question of how to prioritise among services arises. In practice using information on the cost of providing the service outputs together with knowledge about the relationships between service outputs and outcomes is often used. Although the services y_A are not sold on markets marginal costs of producing a given amount may be calculated based on knowing the transformation function $F(y_A, x; z_F)$ and the corresponding cost function (8). Assuming marginal costs to be constant, cost coefficients c_k may be used as prices (we leave open whether cost coefficients reflect both technically efficient production and allocative efficiency of inputs. when formulating an optimisation problem giving the priority rules for outputs):

$$\begin{aligned} &Min \sum_{k=1}^{K} c_k y_{A_k} \\ &\text{subject to} \\ &y_{O_m} = g_m(y_A; z_{g_m}) \text{ given} \end{aligned} \qquad (25)$$

The Lagrangian function for the problem is

$$L = -\sum_{k=1}^{K} c_k y_{A_k} - \sum_{m=1}^{M} \gamma_m (-y_{O_m} + g_m(y_A; z_{g_m})) \tag{26}$$

The first-order conditions are:

$$\frac{\partial L}{\partial y_{A_k}} = -c_k + \sum_{m=1}^{M} \gamma_m g'_{mk}(y_A; z_{g_m}) \leq 0 (= 0 \text{ for } y_{A_k} > 0), k = 1, ..., K$$
$$\gamma_m \geq 0 (= 0 \text{ for } y_{O_m} > g_m(y_A; z_{g_m})), m = 1, .., M \tag{27}$$

Costs should be allocated on the services such that the cost coefficient for service y_{Ak} is equal to the value of the marginal increases in the outcomes, evaluated at the shadow prices on the outcome constraints. The shadow price for outcome y_{Om} expresses the increase in total costs of providing one more unit of outcome y_{Om}. For an interior solution (i.e. the service y_{Ak} will be produced) equation (27) tells us that optimal prioritising of services is characterised by the unit cost of a service being equal to the total marginal "value" created by increases in the outcomes, where values are actually the marginal costs of increases in all outcomes. Equation (27) represents a simplification of equation (19a), where all necessary information is assumed to be available, focussing only on the costs of producing outputs in order to satisfy objectives for outcomes. Comparing (27) and (19b) we see that welfare weights, resource price and productivity of a resource in producing service outputs have been eliminated by the use of cost coefficients.

Combining the optimality conditions (27) for a pair of outputs y_{Ak} and y_{Al}, assuming interior solutions, yields, analogous to (21):

$$\frac{\sum_{m=1}^{M} \gamma_m g'_{mk}(y_A; z_{g_m})}{\sum_{m=1}^{M} \gamma_m g'_{ml}(y_A; z_{g_m})} = \frac{c_k}{c_l}, \; l, k = 1, ..., K \tag{28}$$

Considering a pair of outputs the ratio of the marginal impacts of outputs on outcomes weighted with the common shadow prices on outcomes is equal to the marginal cost ratio. A stylised illustration is presented in Figure 3 for two outputs y_{Ak} and y_{Al}. The curve labelled "y_O given" is a contour curve of the production function (2) in the case of a single outcome. The

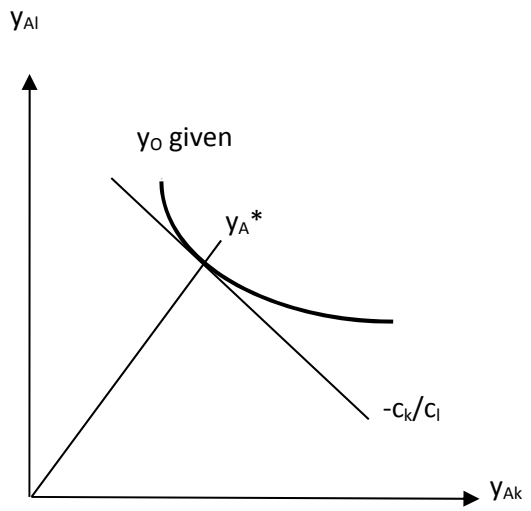

Figure 3. Cost efficiency of providing outcomes

optimal ratio between the outputs is given by y_A*. However, to be able to prioritise between outcomes there must an explicit preference function available. The weighting with the shadow price of outcomes in (27) does not represent preferences, but only reflects the resource cost side via outputs. It is not a technical question how to do such a prioritising, but a political question. Lacking a preference function in outcomes the best a bureaucrat can do is to work out the marginal cost schedules for providing services and then prioritise between services based on considerations of minimising total costs of producing the services given levels of goals for outcomes. This situation may be realistic for the production of many types of public services. Diewert (2011) declares using cost of production as the second best method if output prices are not available.

We have looked at an agency in isolation in the sense that no effects of services on other outcomes than the ones the agency is interested in are included in the analysis. But it is straightforward in principle to include such external-type of effects using the specification in Equation (3).

5. Conclusions

The concepts of efficiency and effectiveness are often used in the literature dealing with efficiency. We have tried to make the distinction between these concepts operational by using the terms outputs and outcomes based on the consideration of the degree of control a public service producer has over its production activity. The apparatus of production theory works best when dealing with resources transformed into service outputs under the control of the organisation in question. Outcomes in this paper represent some higher social goals than outputs and are determined by the outputs and other exogenous variables, but these latter and the outcome production processes will typically be outside the control of the organisation.

The link to the efficiency measurement is provided based on introducing the concept of a benchmark frontier technology for the type of production in question. The measurement of savings potential and their relations with cost efficiency and Farrell's measures of technical efficiency are provided. Technical efficiency measurement can be done without having prices of outputs and inputs, but cost efficiency calculations require input prices.

The relationship between outcomes and outputs and variables not under the control of the service provider, is cast within a framework based on Frisch's scheme of factorially determined multioutput production with outputs and non-discretionary variables as inputs. In order to be able to measure effectiveness in the choice of outputs, i.e., calculate a measure of output mix effectiveness; we must have some kind of evaluation of the outcomes. Introducing a preference function over outcomes optimality conditions for providing an effective output mix for a given resource budget are derived. It is shown that the measure for overall preference effectiveness can be multiplicatively decomposed into the technical output efficiency of realising a frontier technology for the transformation of resources to outputs, and the mix effectiveness of reallocating the use of resources so the optimal mix of outputs is produced. The rather monumental task of providing the necessary information for calculating mix effectiveness is highlighted. A preference function over outcomes must be established, if the organisation in question produces outputs influencing more than one outcome, and also the production relations between outcomes on one hand and outputs and exogenous variables on the other. As far as we know this approach has not been attempted in the literature. An

additional complication is that to capture the links between service outputs, exogenous variables and outcomes, quite complicated dynamic relationships involving time lags may have to be modelled.

It is therefore understandable that empirical applications of measuring efficiency and saving potentials within the public sector have been limited to transformation of resources into outputs within a process controlled by the service provider.

References

Afriat S (1972): "Efficiency estimation of production functions", *International Economic Review* 13(3), 568-598

Aigner DJ and Chu SF (1968). On estimating the industry production function. *American Economic Review* 58, 226-239

Aigner DJ, Lovell CAK and Schmidt P (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics* 6(1), 21-37

Asmild M, Paradi, JC, Reese, DN, and Tam, F (2007). Measuring overall efficiency and effectiveness using DEA. *European Journal of Operational Research* 178, 305-321

Banker RD, Charnes A, and Cooper WW (1984). Some models for estimating technical and scale inefficiency in data envelopment analysis. *Management Science* 30 (9), 1078-1092

Becker GS (1965). A theory of the allocation of time. *Economic Journal* 75, 493-517

Bjerkholt O and Strøm S (2002). Decision models and preferences: the pioneering contributions of Ragnar Frisch. In Tangian AS and Gruber J (Eds), *Constructing and applying objective functions*. Lecture Notes in Economics and Mathematical Systems 510, pp.17-36, Berlin – Heidelberg – New York: Springer-Verlag

Bogetoft P and Hougaard JL (2003). Rational inefficiencies. *Journal of Productivity Analysis* 20, 243-271

Bradford DF, Malt RA and Oates WE (1969). The rising cost of local public services: some evidence and reflections. *National Tax Journal* 22(2), 185-202

Bruijn H de (2002). Performance measurement in the public sector: strategies to cope with the risks of performance measurement. *The International Journal of Public Sector Management* 15(7), 578-594

Burkhead J and Hennigan P. (1978). Productivity analysis: a search for definitions and order. *Public Administration Review* 38, 34-40

Charnes A and Cooper WW (1985). Preface to topics in data envelopment analysis. *Annals of Operations Research* 2, 59-94

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

Chilingerian JA and Sherman, HD (1990). Managing physician efficiency and effectiveness in providing hospital services. *Health Services Management Research* 3(1), 3-15

Chiou Y-C, Chen Y-H (2006). Route-based performance evaluation of Taiwanese domestic airlines using data envelopment analysis. *Transportation Research Part E*, 42, 116-27

Chu X, Fielding GJ and Lamar BW (1992). Measuring transit performance using data envelopment analysis. *Transportation Research Part A*, 26(3), 223-230

Cooper WW and Ijiri Y (1983). *Kohler's dictionary for accountants*, sixth edition, Englewood Cliffs, N. J.: Prentice-Hall

Cooper WW, Seiford LM and Tone K (2000): *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*, Dordrecht/ Boston/London: Kluwer Academic Publishers

De Witte K and Geys B (2011). Evaluating efficient public good provision: theory and evidence from a generalised conditional efficiency model for public libraries. *Journal of Urban economics* 69, 319-27

De Witte K and Geys B (2013). Citizen coproduction and efficient public good provision: theory and evidence from local public libraries. *European Journal of Operations Research* 224, 592-602

Diewert WE (2011). Measuring productivity in the public sector: some conceptual problems. *Journal of Productivity Analysis* 36,177–191

Dixit A (2002). Incentives and organizations in the public sector. An interpretative review. *The Journal of Human Resources* 37(4), 696-727

Drucker P (1977). *An introductory view of management*. New York: Harper College Press.

Duncombe W, Miner J and Ruggiero J (1997). Empirical evaluation of bureaucratic models of inefficiency. *Public Choice* 93, 1-18

Farrell MJ (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, *Series A (General)* 120 (III), 253-281

Farrell MJ and Fieldhouse M (1962). Estimating efficient productions functions under increasing returns to scale. *Journal of the Royal Statistical Society*, *Series A (General)* 125 (2), 252-267

Färe R and Grosskopf S (1994). *Cost and revenue constrained production*. Bilkent University Lecture Series. Berlin: Springer-Verlag

Färe R, Grosskopf S, Lundström M and Roos P (2008). Evaluating health care efficiency. In J.L.T. Blank and V. Valdmanis (eds), *Evaluating hospital policy and performance: contributions from hospital policy and productivity research, Advances in Health Economics and Health Services Research* 18, 209-228

Fitz-Gibbon CT and Tymms P (2002). Technical and ethical issues in indicator systems: doing things right and doing wrong things. *Education Policy Analysis Archives 10(6)*, 1-26 [Retrieved 1/9/2006 from http://epaa.asu.edu.epaa/v10n6/]

Fried HO, Lovell CAK, Schmidt SS (Eds). *The measurement of productive efficiency and productivity growth.* Oxford: Oxford University Press

Frisch R (1965). *Theory of production.* Dordrecht: D. Reidel

Førsund FR (2010). Dynamic efficiency measurement. *Indian Economic Review* 45(2), 123-157.

Førsund FR (2012). Measuring Efficiency in the Public Sector. Memorandum No. 09/2012 from Department of Economics, University of Oslo

Førsund FR and Sarafoglou N (2002). On the origins of Data Envelopment Analysis. *Journal of Productivity Analysis* 17, 23-40.

Golany B and Tamir E (1995). Evaluating efficiency-effectiveness-equality trade-offs: a data envelopment analysis approach, *Management Science* 41(7), 1172-1184

Golany B, Phillips FY, Rousseau JJ (1993). Models for improved effectiveness based on DEA efficiency results, *IIE Transactions*, 25(6), 2-10

Hanoch G (1970). Homotheticity in Joint Production. *Journal of Economic Theory* 2, 423-426

Hanson T (2012). Efficiency and productivity in the operational units of the armed forces. Memorandum No. 07/2012 from Department of Economics, University of Oslo

Haskel J and Sanchis A (1995). Privatisation and X-inefficiency: a bargaining approach. *The Journal of Industrial Economics* 43 (3), 301-321

Haskel J and Sanchis A (2000). A bargaining model of Farrell inefficiency. *International Journal of Industrial Organization* 18, 539-556

Hatry HP (1999). *Performance measurement: getting results*. Washington D.C.: The Urban Institute Press (2nd edition 2006)

Kumbhakar S and Lovell CAK (2000). *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press

Lancaster KJ (1966). A new approach to consumer theory. *Journal of Political Economy* 74(2), 132-157

Lauer JA, Lovell CAK, Murray CJL and Evans DB (2004). World health system performance revisited: the impact of varying the relative importance of health system goals. *BMC Health Services Research*

Leibenstein H (1966). Allocative efficiency vs. "X-efficiency". *American Economic Review* 56(3), 392-415

Lewin AY and Minton JW (1986). Determining organizational effectiveness: another look, and an agenda for research. *Management Science* 32, 514-537

Medina-Borja A. and Triantis K (2011). Modeling social services performance: a four-stage DEA approach to evaluate fundraising efficiency, capacity building, service quality, and effectiveness in the nonprofit sector. *Annals of Operations Research*. Published online 30 June 2011, DOI 10.1007/s10479-011-0917-0

Medina-Borja A, Pasupathy KS and Triantis K (2006). Large-scale data envelopment analysis (DEA) implementation: a strategic performance management approach. *Journal of the Operational Research Society* 58(June), 1084-1098

Meeusen W and Broeck J van den (1978). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18(2), 435-444

Norwegian Defence Fact and Figures (2010). Oslo: Norwegian Ministry of Defence

Ruggiero J (1996a). On the measurement of technical efficiency in the public sector. *European Journal of Operational Research* 90, 553-565

Ruggiero J (1996b). Efficiency of educational production: an analysis of New York school districts. *Review of Economics and Statistics* 78(3), 499-509

Ruggiero J, Duncombe W and Miner J (1995). On the measurement and causes of technical inefficiency in local public services: with an application to public education. *Journal of Public Administration Research and Theory*, *J-PART* 5(4), 403-428

Schreyer P (2008). Output and outcome – measuring the production of non-market services. *Paper Prepared for the 30th General Conference of The International Association for Research in Income and Wealth,* Portoroz, Slovenia, August 24-30, 2008 (http://www.iariw.org/papers/2008/schreyer.pdf)

Shephard RW (1974). *Indirect production functions*. Meisenheim Am Glan: Verlag Anton Hain

Solà M and Prior D (2001). Measuring productivity and quality changes using data envelopments analysis: an application to Catalan hospitals. *Financial Accountability & Management* 17(3), 219-245

Yu M-M and Fan C-K (2009). Measuring the performance of multimode bus transit: A mixed structure network DEA model. *Transportation Research Part E* 45, 501-515

Yu M-M and Lin ETJ (2008). Efficiency and effectiveness in railway performance using a multi-activity network DEA model. *Omega* 36, 1005-1017