Part I

DEA Models, Methods and Interrelations

Annals of Operations Research 66(1996)3–45

Chapter 1

Introduction: Extensions and new developments in DEA

W.W. Cooper[a], R.G. Thompson[b] and R.M. Thrall[c]

[a]*Graduate School of Business, The University of Texas at Austin, Austin, TX 78712-1174, USA*

[b]*College of Business, University of Houston, Houston, TX 77204-6282, USA*

[c]*Jones School of Administration and Computational and Applied Mathematics, Rice University*

The extensions, new developments and new interpretations for DEA covered in this paper include: (1) new measures of efficiency, (2) new models and (3) new ways of implementing established models with new results and interpretations presented that include treatments of "congestion", "returns-to-scale" and "mix" and "technical" inefficiencies and measures of efficiency that can be used to reflect all pertinent properties. Previously used models, such as those used to identify "allocative inefficiencies", are extended by means of "assurance region" approaches which are less demanding in their information requirements and underlying assumptions. New opportunities for research are identified in each section of this chapter. Sources of further developments and possible sources for further help are also suggested with references supplied to other papers that appear in this volume and which are summarily described in this introductory chapter.

Keywords: Technical inefficiency, mix inefficiency, returns to scale, congestion, mathematical programming.

1 Background

The great number and variety of applications of DEA (Data Envelopment Analysis) in recent years has been accompanied by important new developments in concepts and methodology. See the extensive bibliography by Seiford (1994) cited in our references. Originally designed to evaluate DMUs (Decision Making Units) such as schools and hospitals which use multiple inputs to produce multiple outputs with no readily identified "bottom line", DEA has since been accorded a variety of formulations and used for many other types of entities. The original applications were to U.S. institutions, but this is no longer true and centers of research are now located in many different parts of the world which have been the source of many new ideas as well as

new applications. Introductions to DEA are now available in a variety of sources – see, for instance, the opening chapters in the recently published book by Charnes et al. (1994) – so we here emphasize new developments and interpretations and try to suggest further extensions for use and research in DEA.

2 The TDT measure of efficiency[1)]

We start with the following measure of efficiency which has recently been introduced into the literature of DEA by Thompson, and Thrall (1994b), viz.,

$$\underset{u,v}{\text{maximize}} \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \Big/ \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}},$$

where, for any (u, v),

$$\frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} = \max \left\{ \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \,\middle|\, j = 1, \ldots, n \right\}, \tag{1}$$

so $\sum_{r=1}^{s} u_r y_{rk} / \sum_{i=1}^{m} v_i x_{ik}$ is maximal for this latter set of $j = 1, \ldots, n$ ratios. We refer to (1) as the "TDT measure" of relative efficiency. Note that the vectors (u, v) with variables $u_r, v_i \geq 0$ as components are selected to maximize the ratio in the objective of (1). All of the ratios in (1) are synthesized from the x_{ij} and y_{rj}, which are constants that represent the observed values of the $i = 1, \ldots, m$ inputs used and $r = 1, \ldots, s$ outputs produced by *each* of $j = 1, \ldots, n$ DMUs. DMU_o, as represented in the objective of (1), is the DMU to be evaluated from these data by maximizing its score relative to the denominator formed from ratios for the entire collection of DMU_j, $j = 1, \ldots n$ – with DMU_o included in the second as well as the first of the above expressions.

We can succinctly represent the *ratio of ratios*[2)] in the objective of (1) by

$$\frac{y_o}{x_o} \Big/ \frac{y_k}{x_k}, \tag{2}$$

where y_o and y_k represent "virtual outputs" and x_o and x_k represent "virtual inputs". Because y_k/x_k is maximal over the set $k = 1, \ldots, n$, which includes $k = o$, we have $y_o/x_o \leq y_k/x_k$. The above ratios therefore have a maximum value of unity and this is achievable if and only if DMU_o's performance is *not* bettered by some *other* DMU.

Several features of DEA are brought together compactly in (1). First, DMU_o's efficiency is evaluated *relative* to the maximal (best) value attained by the $j = 1, \ldots, n$,

[1)] The material in this section is adapted from Cooper and Tone (1995).

[2)] This can also be interpreted as a productivity measure and used to measure changes of productivity over time in the manner suggested by Färe and Grosskopf (1994).

DMUs (including DMU_o) with which it is being compared. The comparison is effected by assigning the same u, v vectors to each output and each input of every DMU. Because the optimal choice maximizes the efficiency score of DMU_o, no other u, v values can improve its relative value.

A priori assignment of weights is not required and relations between inputs and outputs need not be prescribed explicitly. The relative form of the efficiency evaluations effected by (1) for any DMU_o should also make the following properties clear: An increase in any input for DMU_o cannot improve its efficiency score unless increases in this same input also occur in other DMUs. Similarly, an increase in any of DMU_o's outputs cannot worsen its efficiency score unless increases in this same output occur in other DMUs. (See chapter 7 in this volume by Thompson et al. for a method that can be used to determine when switchovers from efficient to inefficient status will first occur when *all data are varied simultaneously*. See also Thompson et al. (1994a) where this method was introduced.)

What has just been said can be formalized as follows:

Definition 1: **Efficiency** (TDT or "ratio definition")
DMU_o is to be considered ratio efficient if and only if it achieves a value of unity in (1). Conversely, a value less than unity in (1) means that performances of other DMUs provide evidence that DMU_o is relatively inefficient.

We will shortly introduce developments which can be used to identify sources and estimate the amounts of inefficiency in *each* input and output for *every* DMU. Here, we note only that (1) is invariant to the units of measures used. That is, the value of (1) remains unaltered if we replace the y_{rj} and x_{ij} in (1) with new values $\hat{y}_{rj} = k_r y_{rj}$, $r = 1,\ldots,s$, and $\hat{x}_{ij} = c_i x_{ij}$, $i = 1,\ldots,m$, by applying arbitrary constant multipliers k_r, $c_i > 0$, to the units in which the outputs and inputs in every DMU_j are measured.[3] On the other hand, the optimal (u, v) values in (1) need not be unique. Different (u, v) values may yield the same maximal efficiency score.

This latter type of phenonemon is not unique to DEA. We might also identify the subject of alternate optima as a topic that has often been inadequately addressed not only in DEA but in other literatures which deal with subjects related to choices of weights in effecting decisions or evaluating their consequences. This type of problem discovery is not uncommon. New methods and new concepts, such as are involved in DEA, often help to identify lacunae or underemphasized topics in established literatures. An example is provided in "general equilibrium economics", where "rational expectations theory" coupled with uses of "game theory" has identified an under-researched topic in the form of the possible existence of multiple equilibria – some with properties that differ markedly from properties associated with other equilibria.

[3] With, of course, $k_r = 1$ or $c_i = 1$ when the units of measure on some outputs or inputs are not to be changed. See Charnes and Cooper (1985) for details and methods of proof.

3 The CCR ratio model

We should note that (1) may be interpreted and treated in various ways. For example, it may be treated by suitably extended versions of the statistical theory of extreme values. It may also be approached deterministically as a mathematical programming problem, and this is the way we will now proceed.

The following model, known as the "CCR ratio model", can help to clarify matters at this juncture:

$$
\begin{aligned}
\underset{u,v}{\text{maximize}} \quad & \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \\
\text{subject to} \quad & \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, \qquad j = 1, \ldots, n, \\
& \frac{u_r}{\sum_{i=1}^{m} v_i x_{io}} \geq \varepsilon, \qquad r = 1, \ldots, s, \\
& \frac{v_i}{\sum_{i=1}^{m} v_i x_{io}} \geq \varepsilon, \qquad i = 1, \ldots, m.
\end{aligned} \tag{3}
$$

The only new element in these expressions is ε, a positive "non-Archimedean" infinitesimal.[4] We elaborate on its mathematical properties later after noting that its use ensures that all u_r and $v_i > 0$, so all inputs and outputs are to be accorded "some" positive value. These values need not be specified explicitly but can be dealt with by computational processes like the ones described immediately after (7), below.[5] Here, we only note that these $\varepsilon > 0$ provide closure and bound the value of the objective from below. Since the other constraints bound the value of the objective in (3) from above, we can use "max" and "min" rather than "sup" and "inf" in our statement of the objective.

As is clear from (1), many other values could have been chosen, but the unity limit imposed on the first n constraints in (3) is intended to maintain contact with classical definitions of efficiency in engineering and science. In fact, as shown in Charnes et al. (1978), the formulation in (3) greatly generalizes the usual single-output-to-single-input ratio definitions of efficiency which are used in engineering and

[4] See chapter 5 for detailed discussion of dimensionality issues associated with the conditions on $\varepsilon > 0$.

[5] As noted in Arnold et al. (1995), the non-Archimedean elements involve *extensions* to the ordinary field of real numbers. Hence, "very small" real numbers approximations are not justified for mathematical use and are not, in general, satisfactory. A more general approach, as given in Charnes et al. (1991), does not require all $x_{ij}, y_{rj} > 0$. It also relaxes the requirement that all u_r and v_i must be positive and dispenses with any use of these non-Archimedean elements in DEA. See also Thompson et al. (1993) for a discussion of the importance of being able to treat zeros in both the data and the DEA solutions.

science. It also relates these engineering-science definitions and usages to definitions in economics – e.g., the Pareto–Koopmans definitions of efficiency given in Charnes et al. (1985) – which we formalize explicitly as follows.

Definition 2: **Efficiency** (Pareto–Koopmans)
The performance of DMU_o is to be considered fully (100%) efficient if and only if the performance of other DMUs does not provide evidence that some of the inputs or outputs of DMU_o could have been improved without worsening some of its other inputs or outputs.[6)]

We will shortly provide a transformation of (3) that makes it possible to identify the sources and estimate the amounts of inefficiency in each input and output for every DMU in a manner that requires only minimal assumptions for empirical studies. Here, we establish a relation to (1) by noting that a *necessary* condition for optimality in (3) is that at least one of the $j = 1,\ldots,n$ output-to-input ratios in the constraints must be at its upper bound of unity. We can therefore identify (3) with (1) by noting that the denominator in (1) has a value of unity in this case, and the efficiency evaluation for DMU_o simply reduces to whether the numerator in (1) is unity or less.

Maximizing y_o/x_o in (2) can be managerially interpreted in terms of achieving the greatest virtual output per unit virtual input. This provides a basis for extending DEA to evaluate returns to scale efficiencies, as we shall later see. Here, however, we simply note that this interpretation corresponds mathematically to finding values which can be associated with slopes of the supports that envelop the observations. Nothing need otherwise be said explicitly about the functions that govern the relations between inputs and outputs, and these relations are allowed to vary from one DMU to another.

4 Linear programming equivalents

Reference to (3) shows that it is a nonlinear, non-convex programming problem, and hence is best used for conceptual clarification. To give concepts associated with (3) computationally implementable form, we introduce new variables defined as follows:[7)]

$$\begin{aligned} \mu_r &= t\mu_r \qquad && r = 1,\ldots,s, \\ v_i &= tv_i, && i = 1,\ldots,m, \\ 1 &= \sum_{i=1}^{m} v_i x_{io}, \quad \text{so} \quad t = 1 \Big/ \sum_{i=1}^{m} v_i x_{io} > 0. \end{aligned} \tag{4}$$

[6)] This might also be called the Pareto–Koopmans–Farrell definition since the reference to "evidence" supplied by the performance of *other* DMUs has its source in Farrell (1957).

[7)] An alternative approach as given in Charnes et al. (1991) does not require any change of variables to achieve these linear programming formulations.

These are the so-called "Charnes–Cooper transformations" from Charnes and Cooper (1962), which initiated the field of "fractional programming". We now use them to transform the problem in (3) to the first problem in the following dual pair of linear programming problems, with assurance (from fractional programming) that its optimal value will also be optimal for (3):

$$
\begin{aligned}
&\text{minimize} && \theta && -\varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)\\
&\text{subject to} && 0 = \theta x_{io} - \sum_{j=1}^{n} x_{ij}\lambda_j - s_i^-,\\
& && y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+;\\
&\text{maximize} && \sum_{r=1}^{s} \mu_r y_{ro}\\
&\text{subject to} && -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} \mu_r y_{rj} \le 0,\\
& && \sum_{i=1}^{m} v_i x_{io} = 1,\\
& && -v_i \le -\varepsilon,\\
& && -\mu_r \le -\varepsilon,
\end{aligned}
\tag{5}
$$

where the slacks s_i^- and s_r^+ as well as the λ_j are constrained to be non-negative.

As in (1),$i = 1,\ldots,m$ indexes the inputs, while $r = 1,\ldots,s$ indexes the outputs and $j = 1,\ldots,n$ indexes the DMUs. Also, as before, $j = o$ is used to identify the DMU to be evaluated by (a) placing it in the objective while also (b) leaving it in the constraints. Leaving the data for DMU_o in the constraints guarantees that solutions exist for both problems in (5) and, by the dual theorem of linear programming, it follows that they will have finite and equal optimal values. Further, because $\sum_{i=1}^{m} v_i x_{io} = 1$ with $v_i \ge 0$, all i, we will have $t > 0$, so we can move back and forth between (5) and (3). We therefore have the full power of available linear programming algorithms and computer codes to solve (5) or (3), as we wish. We also have its interpretative power available for use in DEA efficiency analyses and inferences.

Referring to (5) as "linear programming equivalents" of the CCR ratio models given in (3) and using * to denote an optimal value, the condition for full (100%) "DEA efficiency", as given in definition 2, now becomes:

$$
\sum_{r=1}^{s} \mu_r^* y_{ro} = 1 \tag{6}
$$

for the second problem in (5). However, interest usually attaches to identifying sources and amounts of inefficiency in each input and output of the DMU being evaluated. This is most easily done from the first problem, where the conditions for efficiency become

$$\begin{aligned} &\text{(i)} \quad \theta^* = 1, \\ &\text{(ii)} \quad \text{all optimum slack values are zero.} \end{aligned} \tag{7}$$

It is now to be noted that θ is to be preemptively minimized, after which the sum of the slacks in (5) is to be maximized.[8] In this way, the non-Archimedean element $\varepsilon > 0$ is given a computational form without any need to specify it explicitly. Its coefficients are then also easily identified with the non-zero slacks and assurance is provided that some DMU_o will not be mistakenly characterized as efficient because a solution is obtained with $\theta^* = 1$ and all slacks at zero while alternate solutions are present which associate non-zero slacks with $\theta^* = 1$.

It is also to be noted that the presence of non-zero slacks means that the measure of inefficiency resulting from (7) assumes the form $\theta^* - k^*\varepsilon$, where k^* = sum of slacks. Both θ^* and k^* are real numbers and hence are Archimedean,[9] whereas ε is a non-Archimedean infinitesimal so that $\theta^* - k^*\varepsilon$ is not a real number unless $k^* = 0$. The following ordering is applicable: If $\theta_1^* < \theta_2^*$, then $\theta_1^* - k_1\varepsilon < \theta_2^* - k_2\varepsilon$ for *all* values of k_1, k_2, while if $\hat{\theta}_1^* = \hat{\theta}_2^*$, then $\hat{\theta}_1^* - \hat{k}_1\varepsilon < \hat{\theta}_2^* - \hat{k}_2\varepsilon$ if and *only* if $\hat{k}_1 > \hat{k}_2 \geq 0$, while $\hat{\hat{\theta}}_1 - \hat{\hat{k}}_1\varepsilon = \hat{\hat{\theta}}_2 - \hat{\hat{k}}_2\varepsilon$ if and only if $\hat{\hat{\theta}}_1 = \hat{\hat{\theta}}_2$ *and* $\hat{\hat{k}}_1 = \hat{\hat{k}}_2$. See Arnold et al. (1994).

One way to ensure the achievement of a single *real-number* measure of inefficiency is to confine attention to "weak (or radial) efficiency" and thereby ignore the slacks. Also referred to as "Farrell" or "Farrell–Debreu" efficiency, this involves an assumption of "free disposal" and confines attention to θ^* as the measure of efficiency.[10] See Färe et al. (1985). Here, "free disposal" means that non-zero slack excesses may be disposed of without cost. See Koopmans (1951, pp. 40 and 70 and theorem 4.11) and Koopmans (1957, pp. 43 and 54). This is made mathematically explicit by replacing $\varepsilon > 0$ with a zero in the objective of the first problem in (5) or, equivalently, by omitting the slacks from explicit representation in the objective. However, when slacks are an important source of possible inefficiency (e.g., because inefficient *mixes* have been used), then alternative approaches and measures may need to be considered (like those given later on) which comprehend *all* inefficiencies (including slacks) in a single real number.

[8] Most DEA computer codes accomplish this in two stages, as follows: stage 1 obtains a value of $\min \theta = \theta^*$ with slacks all multiplied by 0 rather than $\varepsilon > 0$ in the objective of (5). This θ^* is then fixed in (5) so that it cannot be altered in a second stage, which is then directed to maximizing the sum of slacks. See Arnold et al. (1995).

[9] That is, they have the following property: Given any positive real number, n, it is always possible to find another real number x which satisfies $0 < x < n$. This is *not* possible with the non-Archimedean $\varepsilon > 0$.

[10] See Farrell (1957) and Debreu (1951).

We can observe that the non-Archimedean element $\varepsilon > 0$ is not present in the constraints for the first problem in (5). Hence, the values in the constraints involve only real numbers. Thus, when an optimal solution is available, we can secure new coordinates via

$$\theta^* x_{io} - s_i^{-*} = \sum_{j=1}^{n} x_{ij}\lambda_j^* = x_{io}^*$$

and (8)

$$y_{ro} + s_i^{+*} = \sum_{j=1}^{n} y_{rj}\lambda_j^* = y_{ro}^*.$$

First given in Charnes et al. (1978), these are known as "CCR projection operators", because they project the originally observed x_{io}, y_{ro} into x_{io}^*, y_{ro}^*, $i = 1,\ldots,m$; $r = 1,\ldots,s$, which form the coordinates of a point in the set of efficient production possibilities defined by the first problem in (5). These x_{io}^*, y_{ro}^* are coordinates of the point used to evaluate DMU_o with, of course, $x_{io}^* = x_{io}$ and $y_{ro}^* = y_{ro}$ when DMU_o is efficient. See (7).

We next note that if any $v_i^* = \varepsilon$ in the second problem in (5), then we cannot have (6) satisfied because no sum of positive multiples of ε can equal any positive real number. Hence, we then must have $\sum_{r=1}^{s} u_r^* y_{ro} < 1$. Via the duality theory of linear programming (which extends to general ordered fields)[11], we then also have

$$\theta^* - \varepsilon\left(\sum_{i=1}^{m} s_i^{-*} + \sum_{r=1}^{s} s_r^{+*}\right) = \sum_{r=1}^{s} \mu_r^* y_{ro} < 1.$$

Because θ is a real number, this equality requires equating the non-Archimedean $\mu_r^* y_{ro}$ on the right to terms involving non-zero slack on the left in a manner analogous to the way such relations are treated in complex variable analysis. See Arnold et al. (1994).

The use of non-Archimedean elements occurs in other parts of mathematical programming. One example is the so-called "big M", which is used in association with artificial variables and discarded after these variables are all eliminated in the first phase of a two-phase procedure *en route* to a solution of the originally stated problem. In DEA, however, the non-Archimedean element is retained and recent literature has shown that these elements bring properties into the solution that can be of interest in their own right. For instance, as shown by the "two-duals theorem" given in Arnold et al. (1994), these elements can be used to obtain solutions to more than one dual from an optimal tableau associated with a solution to a single primal problem. Whether and how this might be extended beyond DEA to other types of mathematical programming problems is a topic that invites further research attention.

[11] See Charnes and Cooper (1958).

5 Other models and the treatment of mix and technical inefficiencies

Adjoining the constraint $\sum_{j=1}^{n} \lambda_j = 1$ to (5) produces the following modification:

$$
\begin{aligned}
&\text{minimize} && \theta_o && -\varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\
&\text{subject to} && 0 = \theta_o x_{io} - \sum_{i=1}^{n} x_{ij}\lambda_j - s_i^-, \\
& && y_{ro} = \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \\
& && 1 = \sum_{j=1}^{n} \lambda_j;
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&\text{maximize} && \sum_{r=1}^{s} \mu_r y_{ro} + u_o \\
&\text{subject to} && -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} \mu_r y_{rj} + u_o \le 0, \\
& && \sum_{i=1}^{m} v_i x_{io} = 1, \\
& && -v_i \le -1\varepsilon, \\
& && -\mu_r \le -1\varepsilon;
\end{aligned}
$$

$$
0 \le \lambda_j, s_i^-, s_r^+ \quad \text{for} \quad j = 1, \ldots, n;\ i = 1, \ldots, m;\ r = 1, \ldots, s.
$$

We have explicitly assigned a unity coefficient to each $\varepsilon > 0$ in the last $m + s$ constraints which bound the values of the variables v_i and μ_r in the second problem in (9).[12] These unity coefficients serve to satisfy the dimensional requirements for these variables, which are stated in reciprocals of the inputs and outputs to which they refer. Hereafter, we shall simply regard these unity elements as also being present, as in the first problem in (9) where they serve as coefficients of the s_i^- and s_i^+ even though we have not identified these unity coefficients explicitly in the objective. See the "goal vectors" discussed by Thrall in chapter 5.

Although the modification in going from (5) to (9) is simple and straightforward, this new model (known as the "BCC model")[13] possesses very important properties which expand the range and uses of DEA as we shall see when we study "returns-to-scale efficiencies" later in this chapter.

[12] This means that these unity elements must be changed when changes are effected for the units in which the y_{ro} and x_{io} are measured. See the discussion for (14.1) ff. below.

[13] From Banker et al. (1984). See also Byrnes et al. (1984).

To further clarify what is being accommodated in DEA, we introduce yet another class known as "additive models", which we formulate as follows:[14]

$$
\begin{aligned}
&\text{maximize} && \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \\
&\text{subject to } x_{io} &&= \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \\
&\qquad y_{ro} &&= \sum_{j=1}^{n} y_{rj}\lambda_j \qquad - s_r^+, \\
&\qquad\quad 1 &&= \sum_{j=1}^{n} \lambda_j ; \qquad\qquad (10)\\
&\text{minimize} && \sum_{i=1}^{m} v_i x_{io} - \sum_{r=1}^{s} \mu_r y_{ro} + u_o \\
&\text{subject to} && \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} + u_o \geq 0, \\
&\qquad v_i && \geq 1, \\
&\qquad\quad \mu_r && \geq 1;
\end{aligned}
$$

$$0 \leq \lambda_j, s_i^-, s_r^+, \ j = 1, \ldots, n; \ i = 1, \ldots, m; \ r = 1, \ldots, s.$$

There is yet another version of the additive model that omits the condition $\sum_{j=1}^{n} \lambda_j = 1$ which we would use if we were studying relations with (5) rather than (9). Here, however, we need only note the following theorem, which is proved in Ahn et al. (1989).

Theorem 1
A DMU_o will be efficient under the additive model in (10) if and only if it is also efficient under the BCC model given in (9).

When a DMU is inefficient, however, its sources and amounts of inefficiencies may differ because of the different metrics employed for the efficiency evaluations in (9) and (10).

Other models, which have also been introduced in the DEA literature, will not be treated here.[15] These include the "multiplicative models" of Charnes et al. (1982,

[14] First published in Charnes et al. (1985).

[15] This does not mean that these models are unimportant or should be excluded from further consideration. It simply means that we cannot here treat them in the detail required to show how thay may be advantageously employed.

1983), the Free Disposal Hull (FDH) models of Tulkens (1993), and the "Russell Measure" (RM) models introduced by Färe et al. (1978, 1985). See also Deprins et al. (1984) and Russell (1988). The latter two classes of models are special cases of the models covered here. This is demonstrated for the FDH and RM models in Bardhan et al. (1996, Part II), while, as shown in Charnes et al. (1982, 1983), the multiplicative models transform to additive models when the observations are restated in logarithmic units.

Returning to (10) and (9), we omit the condition $\sum_{j=1}^{n} \lambda_j = 1$ in both of these models and thereby also delay our discussion of u_o in the dual until we examine returns to scale efficiencies in section 7. We can then start our present discussion with the following modification to the dual of (10):

$$\begin{aligned} \text{maximize } & \sum_{r=1}^{s} \mu_r y_{ro} - \sum_{i=1}^{m} \nu_i x_{io} \\ \text{subject to } & \sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} \nu_i x_{ij} \le 0, \\ & \mu_r \ge 1, \\ & \nu_i \ge 1, \end{aligned} \tag{11}$$

where $j = 1,\ldots,n$; $r = 1,\ldots,s$; $i = 1,\ldots,m$. Observing the form of constraints in (11), we see that we must have

$$\max \sum_{r=1}^{s} \mu_r y_{ro} - \sum_{i=1}^{m} \nu_i x_{io} = \sum_{r=1}^{s} \mu_r^* y_{ro} - \sum_{i=1}^{m} \nu_i^* x_{io} \le 0 \tag{12}$$

with equality holding if and only if the DMU_o being evaluated is efficient. Returning to (9), we next note that the ν_i and μ_r are expressed in units which are reciprocal to y_{ro} and x_{io}, so the same must be true for the slack coefficient in the objective for the primal in (9). See the constraints on the ν_i and μ_r in (9).

We now use the primals in (9) and (10) to show that *two* types of inefficiencies are considered in (7) – viz., (i) "mix inefficiencies" in the form of inefficient proportions and (ii) "technical inefficiencies" in the form of excessive uses of *all* resources without altering their proportions. For this purpose, notice that θ is first minimized in (9) without altering the input proportions. Hence, a value of $\theta^* < 1$ shows technical inefficiencies to be present in all inputs. The presence of non-zero slacks means that further reductions can be made. These further adjustments will necessarily alter the proportions used and hence they represent mix inefficiencies – because the preceding (preemptive) minimization of θ has exhausted all of the possibilities for removing excesses *without altering the proportions* in which the inputs were used.

Formally, we hereafter associate "technical inefficiencies" with θ and say that technical efficiency has been attained when $\theta^* = 1$ – in which case, DMU_o is at a point on the boundary (not necessarily efficient) of its production possibility set. The

further improvements associated with non-zero slacks will be referred to as "mix inefficiencies" because they involve altering the proportions in which inputs and outputs are used. The additive model in (10) does not distinguish between mix and technical inefficiencies. Although a supplementary analysis may be used to effect these distinctions *after* a solution has been achieved, it is probably better (in most cases) to utilize (9) or (5) when these distinctions are of interest. In addition to avoiding extra work, the latter choice avoids additional difficulties associated with the possible presence of alternate optima – problems which are avoided in characterizing purely technical inefficiency in (10) because the optimal $\theta = \theta^*$ is unique and invariant to the units in which the different inputs may be expressed.[16)]

This brings us to the topic of how these mix and technical inefficiencies may be combined in a single measure. We approach this by first utilizing the duality theory of linear programming and extending (12) to

$$\sum_{i=1}^{m} s_i^{-*} + \sum_{r=1}^{s} s_r^{+*} = \sum_{i=1}^{m} v_i^* x_{io} - \sum_{r=1}^{s} \mu_r^* y_{ro} \geq 0, \tag{13}$$

with the inequality replaced by equality if and only if all slacks are zero. In words, a DMU_o will be characterized as fully efficient if and only if all slacks are zero in an optimum solution to (10).

A valid criticism of the primal objective in (10) is that the amount of inefficiency and, indeed, the choice of non-zero slacks entering into an optimal solution can depend on the units in which the slacks are stated. See the discussion immediately following (9) above, and see chapter 5 for detailed discussions. Much recent work has gone into formulating measures which are better than the one provided in the objective of (10). See, e.g., Lovell and Pastor (1995), Banker and Cooper (1994) or Bardhan et al. (1996, Parts I and II) and Thrall (chapter 5). We cannot discuss this work in all detail. We return, instead, to (10) with its convexity condition represented by $\sum_{j=1}^{n} \lambda_j = 1$ and replace its objective with the following, which has been suggested by Cooper and Pastor (1995):

$$\max \frac{\sum_{i=1}^{m} \left(\frac{s_i^-}{R_i^-} \right) + \sum_{r=1}^{s} \left(\frac{s_r^+}{R_r^+} \right)}{m + s}, \tag{14.1}$$

where R_i^- and R_r^+ are Ranges defined by

$$\begin{aligned} R_i^- &= \bar{x}_{ij} - \underline{x}_{ij}, \quad i = 1, \ldots, m, \\ R_s^+ &= \bar{y}_{rj} - \underline{y}_{rj}, \quad r = 1, \ldots, s, \end{aligned} \tag{14.2}$$

[16)] See (35) below and the discussion following it for a model which effects this distinction *simultaneously* in inputs and outputs. See also the appendix for its relations to (5).

with $\bar{x}_{ij}, \underline{x}_{ij}$ and $\bar{y}_{rj}, \underline{y}_{rj}$ representing maximal and minimal observed values for each i or r taken over the $j = 1,\ldots, n$ DMUs for the input or output in question.

This objective has several desirable and easily interpreted properties. First, we have[17]

$$\begin{aligned} 0 \le s_i^- &= x_{io} - \sum_{j=1}^{n} x_{ij}\lambda_j \le \bar{x}_{ij} - \underline{x}_{ij}, \quad i = 1,\ldots,m, \\ 0 \le s_r^+ &= \sum_{j=1}^{n} y_{rj}\lambda_j - y_{rj} \le \bar{y}_{rj} - \underline{y}_{rj}, \quad r = 1,\ldots,s \end{aligned} \tag{15}$$

and, therefore,

$$0 \le \frac{\sum_{i=1}^{m}\left(\frac{s_i^-}{R_i^-}\right) + \sum_{r=1}^{s}\left(\frac{s_r^+}{R_r^+}\right)}{m+s} \le 1. \tag{16}$$

Each (s_i^-/R_i^-) and (s_r^+/R_r^+) measures the amount of inefficiency relative to the range of possible inefficiencies which the observations show to be possible for *each* input and output. Thus, (16) represents an average of the inefficiency proportions with zero achieved if and only if all slacks are zero and unity achieved if and only if equality is attained in every one of the expressions in (15). Moreover, if it is desired to emphasize "efficiency" rather than "inefficiency", one can replace (16) with

$$0 \le 1 - \frac{\sum_{i=1}^{m}\left(\frac{s_i^-}{R_i^-}\right) + \sum_{r=1}^{s}\left(\frac{s_r^+}{R_r^+}\right)}{m+s} \le 1. \tag{17}$$

Other measures of efficiency for other DEA models are also possible. See Cooper and Pastor (1995). In any case, for science purposes, the measures to be used need to satisfy a variety of criteria. Here, we note two such desiderata: The measures should be (i) complete and (ii) coordinate free. These properties can be described as follows:

(i) *Completeness.* The selected measure should reflect all of the pertinent aspects of the phenomena to be accounted for – which here take the form of technical and mix inefficiencies.

(ii) *Coordinate free.* The value of the selected measure should be invariant to choices of origin and the units in which the different inputs and outputs are measured.

We use the measures given in (16) and (17) for illustration. Because numerators and denominators are stated in the same units for each term in (16) and (17), it follows that these measures are "units invariant". Varying the units in which any output or

[17] It should be evident from these expressions that the slacks will be zero whenever the range is zero, as follows from the condition $\sum_{j=1}^{n}\lambda_j = 1$, so these constraints may simply be omitted ab initio and zero values assigned to these slacks in the objective.

input is measured will not affect the solution when (14.1) is used in the objective for the additive model in (10). For example, "miles" may be replaced by "kilometers". (The "row rule" of linear programming given in Charnes and Cooper (1961, p. 29) completes the analysis.)

As shown by Ali and Seiford (1990), the primal model on the left in (10) is "translation invariant". This means that the solution set to the primal of (10) is not altered when arbitrary constants are added to the x_{ij} or y_{rj}, for all $j = 1,\ldots,n$, in any of the constraints. This property of translation invariance is carried over into the objectives represented by (16) and (17), as demonstrated for their numerators by writing

$$\begin{aligned} s_i^- &= (x_{io} + d_i) - \sum_{j=1}^{n}(x_{ij} + d_i)\lambda_j = x_{io} - \sum_{j=1}^{n} x_{ij}\lambda_j, \qquad i = 1,\ldots,m, \\ s_r^+ &= \sum_{j=1}^{n}(y_{rj} + d_r)\lambda_j - (y_{ro} + d_r) = \sum_{j=1}^{n} y_{rj}\lambda_j - y_{ro}, \qquad r = 1,\ldots,s, \end{aligned} \tag{18.1}$$

because $\sum_{j=1}^{n} y_{rj}\lambda_j = 1$. Similarly, for the denominators,

$$\begin{aligned} R_i^- &= \bar{x}_{ij} - \underline{x}_{ij} = \bar{x}_{ij} + d_i - (x_{ij} + d_i), \qquad i = 1,\ldots,m, \\ R_s^+ &= \bar{y}_{rj} - \underline{y}_{rj} = \bar{y}_{rj} + d_r - (\underline{y}_{rj} + d_r), \qquad i = 1,\ldots,m. \end{aligned} \tag{19}$$

Hence these measures are "coordinate free".

In some applications, the measures used have natural zeros as, for instance, "number of acres used" or "number of miles traveled". Other applications use measures which have no natural origin as in "the number of degree days" that enter into evaluation of Air Force activities at different bases, or the Coopersmith (psychological) scores used to evaluate effects on "self esteem" for underprivileged children in different educational programs. The coordinate-free properties of (16) and (17) make them suitable for use in such cases. They can also be used to handle mixtures of input and output measures and the coordinate-free properties make it possible to treat negative values (such as loss versus profit outputs) in a straightforward manner.

Evidently, the measures in (16) and (17) are also "complete" in that the non-zero slacks represent *all* technical and mix inefficiencies identified by these additive models and these inefficiencies are the pertinent phenomena for evaluating efficiency as given in definition 2. This is clearly a desirable property and clear warnings should generally be given when this property is absent in the efficiency measures used to evaluate "total performance".

These measures also have other desirable properties. For instance, they are strictly monotonic in each input and output. That is, an increase in any input or decrease in any output with all other variables held constant increases the value of (16).[18] See

[18] Provided the limits prescribed in the denominators are not breached by the corresponding input or output variations-

Lovell et al. (1995). Further, the limit of zero in (16) is reached only when all slacks are zero and the limit of unity is reached for (16) when DMU_o, with inputs $\bar{x}_{ij}$ and outputs $\underline{y}_{rj}$, is evaluated by one or more DMU_j with inputs $\underline{x}_{ij}$ and outputs $\bar{y}_{rj}$.

Other desirable properties of this and other measures are discussed in Cooper and Pastor (1996) and Aida et al. (1997). There are also limitations to be considered. For instance (in keeping with what we said earlier about the additive model), this measure fails to distinguish between mix and technical inefficiencies and may thus fail to reflect alterations in mix which occur when several inputs or outputs are varied simultaneously. Also, as Thrall notes in chapter 4, further research on the duals is warranted since their properties will not, in general, be translation invariant even when this is true for the primal.

6 Congestion

Congestion refers to situations in which some inputs are used in amounts that interfere with output productions. An excess of miners bumping into each other in an underground mine is an example where a reduction in the number of miners can result in an increase in the amount mined. Extensions to multiple output-multiple input situations can be formally defined for use in DEA as follows:

Congestion: Evidence of congestion is present when *reductions* in one or more inputs can be associated with *increases* in one or more outputs or, proceeding in reverse, when *increases* in one or more inputs can be associated with *decreases* in one or more outputs.

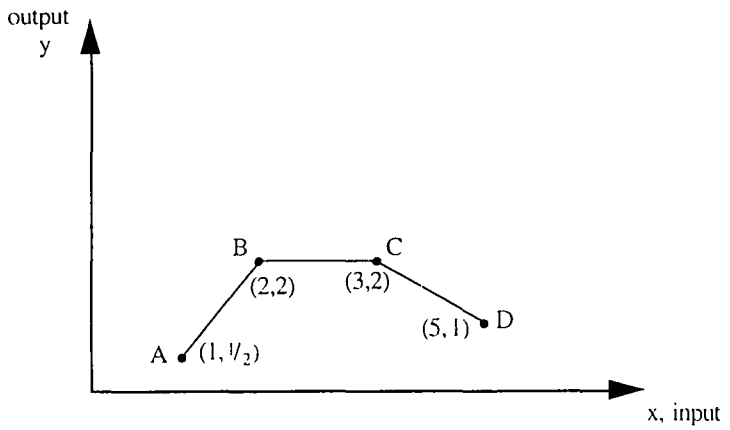


Figure 1. Congestion.

Figure 1 is illustrative of the single output-single input case. Here, the first number in the parentheses indicates an input amount and the second an output amount

associated with each point represented by a black dot for each of the DMUs labeled A to D. Evidently, a reduction from 5 units to 3 in the input used by D could bring it into coincidence with C. This provides evidence of congestion in which a *decrease* of two units of input is associated with a one unit *increase* in output obtained by moving from D to C.

We are still not done. C is not fully efficient since a movement from C to B results in a further reduction of input by one unit *without* any decrease in output. The input reduction in moving from C to B is not associated with an output *increase*. It thus represents an ordinary "technical" or "mix inefficiency"[19] of the kind previously discussed in which no *worsening* of output occurs with this input reduction. See definition 2. Stated differently, our evaluation of D contains a congestion component in its inefficiencies. Identification of this component is informational in the sense that an ordinary DEA analysis would correctly identify the need for further reducing C's input by 1 unit when effecting an evaluation of D's performance. See the earlier discussion following (7) and the association of $\varepsilon > 0$ with slack maximization.

In order to identify these congestion elements (and amounts), we first use the following model:

$$\begin{aligned}
\text{maximize} \quad & \phi & & + \varepsilon \left(\sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \right) \\
\text{subject to} \quad & 0 = \phi y_{ro} - \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+, \\
& x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-, \\
& 1 = \sum_{j=1}^{n} \lambda_j, \\
& 0 \leq \lambda_j, s_r^+, s_i^- \quad \text{for} \quad j = 1, \ldots, n;\ r = 1, \ldots, s;\ i = 1, \ldots, m.
\end{aligned} \tag{20}$$

As can be seen, (20) represents a modification of (9) in which the latter's input orientation is replaced by an orientation which is directed to maximizing the outputs.

We next proceed to a second-stage optimization using a new model which we develop as follows. First we solve (20) to obtain optimal values of ϕ^* and s_r^{+*}, s_i^{-*} for $r = 1, \ldots, s$ and $i = 1, \ldots, m$. Then we apply the CCR projection operators by modifying (8) in an obvious way to allow for moving from an input to an output orientation. This replaces the coordinates y_{ro}, x_{io} for DMU_o with new coordinates, y_{ro}^* and x_{io}^*, which we use to formulate the following new problem:

[19] The two are indistinguishable in the single output-single input case.

$$
\begin{aligned}
&\text{maximize} && \sum_{i=1}^{m} s_i^- \\
\text{subject to}\quad & y_{ro}^* = (\phi^* y_{ro} + s_r^{+**}) = \sum_{j=1}^{n} y_{rj}\lambda_j, && r = 1,\ldots,s, \\
& x_{io}^* = (x_{io} - s_i^{-*}) = \sum_{j=1}^{n} x_{ij}\lambda_j - s_i^-, \\
& 1 = \sum_{j=1}^{n} \lambda_j, \\
& s_i^{-*} \geq s_i^-, && i = 1,\ldots,m,
\end{aligned}
\tag{21}
$$

where all variables are constrained to be non-negative.[20]

As noted for (8), these y_{ro}^*, x_{io}^* are coordinates of the point on the efficiency frontier that was used to evaluate DMU_o. Here, in (21), we have expressed the first s constraints as equations by not allowing for slacks and hence we have eliminated possible reductions in the y_{ro}^*. We then use (21) to determine the largest possible sum of input slack values which are consistent with these y_{ro}^*. Denoting these values by s_i^{-**}, we then use

$$
0 \leq \hat{s}_i^- = s_i^{-*} - s_i^{-**}, \quad i = 1,\ldots,m, \tag{22}
$$

to obtain a value for $\hat{s}_i^-$ as the amount of congestion in the ith input.

For illustration, we insert the data for D from figure 1 into (20) to obtain

$$
\begin{aligned}
&\text{maximize} && \phi + \varepsilon(s^+ + s^-) \\
&\text{subject to} && 0 = \phi - \tfrac{1}{2}\lambda_A - 2\lambda_B - 2\lambda_C - 1\lambda_D + s^+, \\
& && 5 = 1\lambda_A + 2\lambda_B + 3\lambda_C + 5\lambda_D + s^-, \\
& && 1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D,
\end{aligned}
\tag{23}
$$

with all variables also constrained to be non-negative. The optimum is given by $\lambda_B^* = 1$ and all other $\lambda^* = 0$ to obtain $\phi^* = 2$, $s^{+*} = 0$ and $s^{-*} = 3$. Since $\lambda_B^* = 1$, we find B with coordinates (2, 2) serving as the evaluator of D in figure 1 – because this is where the slack maximization associated with $\varepsilon > 0$ has positioned the solutions.

To locate the congestion in $s^{-*} = 3$ and estimate its amount, we use our solution from (23) and substitute in (21) to form the following new problem:

[20] The last $i = 1,\ldots,m$ constraints can be treated by the "bounded variables routine" which are incorporated in many linear programming codes. See Charnes and Cooper (1961, pp. 561–562) or Goldfarb and Todd (1989, pp. 109–114). Note also that DMUs identified as efficient by (20) will have all $s_i^{-*} = 0$ and hence may be omitted from further consideration by virtue of the last $i = 1,\ldots,m$ constraints in (21).

$$
\begin{aligned}
&\text{maximize} \quad s^- \\
&\text{subject to} \quad 2 = \tfrac{1}{2}\lambda_A + 2\lambda_B + 2\lambda_C + 1\lambda_D, \\
&\qquad\qquad 2 = 1\lambda_A + 2\lambda_B + 3\lambda_C + 5\lambda_D - s^-, \\
&\qquad\qquad 1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D, \\
&\qquad\qquad 3 \geq s^-, \\
&\qquad\qquad 0 \leq \lambda_A, \lambda_B, \lambda_C, \lambda_D, s^-.
\end{aligned}
\tag{24}
$$

This produces $\lambda_C^* = 1$ with $s^{-**} = 1$, and (correctly) identifies C as the DMU to be used in obtaining $\hat{s}^- = s^{-*} - s^{-**} = 3 - 1 = 2$. Hence, $\hat{s}^- = 2$ is the "congesting" amount of input associated with the output reduction in going from C to D in figure 1.

Returning to (24), as developed from figure 1, we see that $\lambda_D = 1$ is not an admissible solution. To admit $\lambda_D = 1$, we would first have to *reduce* the output value from $y^* = 2$ to $y = 1$ in the first constraint. Thus, as our definition of congestion requires, we have this input *increase* as the congesting amount associated with an output *reduction*.

Before concluding the present analysis, we might note that other paths are also available. Färe, Grosskopf and Lovell (1985, pp. 68–77), for instance, provide an alternate approach which utilizes the concept of "allocative efficiency" to arrive at estimates of congestion. As discussed in section 9, the concept of allocative efficiency assumes knowledge of the exact value of prices and it also assumes that these prices remain fixed over the periods that are pertinent to the analysis. The approach we have outlined above, however, requires no such knowledge or assumptions, so it can be employed in circumstances besides those in which the FGL approach is applicable.[21]

The example we have used for illustration is confined to the case of one output and one input. Extensions to multiple outputs and inputs provide additional possibilities for attention. For instance, an initial analysis along the above lines might be supplemented by reducing one or more of the outputs to associate subsets of congesting inputs with these reductions. The process could then be continued until all of the congesting inputs have been identified. Here, however, we simply stop with the initial analysis and assign the congesting amounts of each input obtained from (22) to all of the reductions in outputs with which they might be associated.

Additional complications arise when alternate optimum possibilities are present. Here, too, we cut our analysis short and simply remark that in some cases this may be regarded as providing an additional opportunity for choosing among solutions which allow for different mixes available by following different paths for rectification. Managements have sometimes welcomed such possibilities.[22] Scientists, however,

[21] Rolf Färe has called our attention to an earlier article which introduced the concept of "congestion" and deals only with considerations of technical efficiency. See Färe and Grosskopf (1983).

[22] In fact, this was the case at the Gulf Oil Company in the very first industrial applications of linear programming. See Charnes et al. (1954). See also Brockett et al. (1997).

generally prefer to have solutions which are unique when the same methods of analysis are applied to the same data. In such cases, one may try to attain an optimum which is unique. See the discussion of "strong complementary slackness" as given by Thrall in chapter 5. See also the interior-point algorithm which can be used for obtaining solutions involving "geometric centers" in González-Lima et al. in chapter 6.

7 Returns to scale: Qualitative characterizations [23]

DEA formulations for returns to scale have undergone a rapid evolution. In Banker et al. (1984), uniqueness in solutions was assumed. See also Färe et al. (1985). This assumption was subsequently dropped in an article by Banker and Thrall (1992), which also extended the earlier treatment of returns to scale in a variety of ways. Still further extensions and relaxation of assumptions are undertaken here which build on this earlier work.

We start with theorems from Banker and Thrall, which we develop as follows. Writing (X_o, Y_o) for input and output vectors, respectively, with components x_{io}, y_{ro} as given in (9), we utilize the dual in (9) and state the following theorem.

Theorem 2 (Banker and Thrall)

(i) Increasing returns to scale prevail at (X_o, Y_o) if and only if $u_o^* > 0$ for *all* optimal solutions.

(ii) Decreasing returns to scale prevail at (X_o, Y_o) if and only if $u_o^* < 0$ for *all* optimal solutions. (25)

(iii) Constant returns-to-scale prevail at (X_o, Y_o) if $u_o^* = 0$ in *any* optimal solution.

Returns to scale generally has an unambiguous meaning only if (X_o, Y_o) is on the efficiency frontier, and Banker and Thrall assumed this in the model they provide for bounding the scale elasticties.[24] Subsequently, Banker et al. (1995) removed the need for making this assumption by replacing the model given by Banker and Thrall (1992, p. 81) with a modification which we develop as follows. Suppose we have achieved an optimum to (9) with $u_o^* < 0$. We can avoid having to check *all* alternate optima, as required for (ii) in (25), by solving for max $u_o = u_o^{**}$ in

[23] Materials in this section have been adapted from Banker, Bardhan and Cooper (1996) and Banker, Chang and Cooper (1996).

[24] Banker and Thrall assume only "weak (or radial) efficiency", but implicitly extend this by allowing the possibility of infinite and zero returns to scale, as noted in the discussions following (41) below. Technical and scale inefficiencies are then not distinguishable in these regions of the frontier, and either characterization may be used to interpret and evaluate performances.

$$
\begin{aligned}
&\text{maximize } u_o \\
\text{subject to } &-\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} \mu_r y_{rj} + u_o \le 0, \quad j = 1, \ldots n;\ j \neq o, \\
&-\sum_{i=1}^{m} v_i x_{io}^* + \sum_{r=1}^{s} \mu_r y_{ro}^* + u_o \le 0, \quad j = o, \\
&\sum_{i=1}^{m} v_i x_{io}^* = 1, \\
&\sum_{r=1}^{s} \mu_r y_{ro}^* + u_o = 1, \\
&v_i \ge 0, \\
&\mu_r \ge 0, \\
&u_o \le 0,
\end{aligned}
\tag{26}
$$

where x_{io}^* and y_{ro}^* are obtained by applying (8) to the solution obtained from (9).

Attention is now directed to the last constraint which requires $u_o \le 0$ so that, referring to (25), we see that constant returns to scale will prevail for DMU_o if and only if we can attain a solution to (26) with $u_o^{**} = 0$, while decreasing returns to scale will prevail if and only if an optimum is achieved with $u_o^{**} < 0$. Thus it is not necessary to examine *all* alternate optima. The formulation given in (26) confines solutions to those which are critical and the use of x_{io}^*, y_{ro}^* obviates the need for *assuming* that one is on the efficiency frontier.

Next, turning to the case when we have a solution to (9) with $u_o^* > 0$, we can replace $u_o \le 0$ in the last constraint of (26) with $u_o \ge 0$ and reorient the objective to min u_o. Again, returns to scale are constant, if we achieve a solution with a new $u_o^{**} = 0$. Otherwise, all alternate optima have the sign originally obtained from (9), and all pertinent possibilities are covered as required for (i) in (25).

This does not end the possible approaches that can be used for returns-to-scale characterizations in DEA. As noted in Banker et al. (1984), the model (5) may also be used to obtain returns-to-scale characterizations by reference to whether $\sum_{j=1}^{n} \lambda_j^* \ge 1$ or $\sum_{j=1}^{n} \lambda_j^* \le 1$ at an optimum. Banker and Thrall also extend this result to allow for the presence of alternative optimum possibilities from a use of (5), as follows.

Theorem 3 (Banker and Thrall)

$$
\begin{aligned}
&\text{If } \textstyle\sum_{j=1}^{n} \lambda_j^* = 1 \text{ in } \textit{any} \text{ alternate optimum then constant returns to scale prevail.} \\
&\text{If } \textstyle\sum_{j=1}^{n} \lambda_j^* > 1 \text{ for } \textit{all} \text{ alternate optima then decreasing returns to scale prevail.} \\
&\text{If } \textstyle\sum_{j=1}^{n} \lambda_j^* < 1 \text{ for } \textit{all} \text{ alternate optima then increasing returns to scale prevail.}
\end{aligned}
\tag{27}
$$

As developed by Banker and Thrall, this theorem also assumes that the point to be evaluated is on the frontier but, again, the need for this assumption was subse-

quently eliminated by Banker et al. (1996) in a formulation which we develop as follows. Suppose an optimum solution has been obtained for (5) with $\sum_{j=1}^{n} \lambda_j^* < 1$. To check on alternate optima possibilities we then replace (5) with

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{n} \hat{\lambda}_j + \varepsilon \left(\sum_{i=1}^{m} \hat{s}_i^- + \sum_{r=1}^{s} \hat{s}_r^+ \right) \\ \text{subject to} \quad \theta_o^* X_o &= \sum_{j=1}^{n} X_j \hat{\lambda}_j + \hat{s}^-, \\ Y_o &= \sum_{j=1}^{n} Y_j \hat{\lambda}_j - \hat{s}^+, \\ 1 &\geq \sum_{j=1}^{n} \hat{\lambda}_j, \end{aligned} \tag{28}$$

where X_j, Y_j, X_o and Y_o are vectors of the observed values for inputs and outputs[25] and the components of the slack vectors $\hat{s}^-$ and $\hat{s}^+$, as well as the components of the vector $\hat{\lambda}$, are constrained to be non-negative. Here, θ^* is the optimal θ obtained from (5). The optimal solution to (28) yields values $\hat{\lambda}_j^*$ for which $\sum_{j=1}^{n} \hat{\lambda}_j^*$ is maximal, so the following theorem is immediate.

Theorem 4 (Banker, Chang and Cooper)
Given the existence of an optimal solution with $\sum_{j=1}^{n} \lambda_j^* < 1$ in (5), the returns to scale at (X_o, Y_o) are constant if and only if $\sum_{j=1}^{n} \hat{\lambda}_j^* = 1$ and returns to scale are increasing if and only if $\sum_{j=1}^{n} \hat{\lambda}_j^* < 1$ in (28).

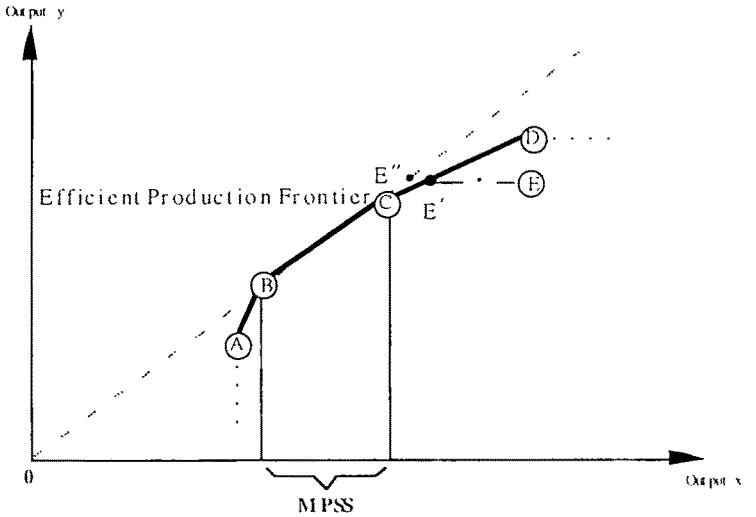
We are here restricting attention to solutions of (28) with $\sum_{j=1}^{n} \hat{\lambda}_j \leq 1$, but the examples we provide show how to treat situations in which θ^* is associated with solutions of (5) that have values $\sum_{j=1}^{n} \lambda_j^* > 1$. To develop what is involved, we use figure 2 with coordinate values listed on the bottom as follows:

$$\text{A} = (1,1),\ \text{B} = (\tfrac{3}{2}, 2),\ \text{C} = (3,4),\ \text{D} = (4,5),\ \text{E} = (4, \tfrac{9}{2}), \tag{29}$$

where the first parenthesized value is an input amount and the second an output amount. Using A, which is BCC but not CCR efficient in figure 2, we substitute from (29) into (5) and write[26]

[25] That is, these observation vectors are *not* adjusted to lie on the efficiency frontier because this is accomplished automatically by the $\hat{\lambda}_j^*$ in (28). See Banker et al. (1995).

[26] Here, we omit the slacks (with coefficient $e > 0$) from the objective since this is handles at the next stage, as in (28).



A=(1,1), B=$(\frac{3}{2}, 2)$, C=(3,4), D=(4,5), E=$(4, \frac{9}{2})$

Figure 2. Most productive scale size.

$$\begin{aligned} &\text{minimize } \theta \\ &\text{subject to } 1\theta \geq 1\lambda_A + \tfrac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E, \\ &\qquad 1 \leq 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \tfrac{9}{2}\lambda_E, \\ &\qquad 0 \leq \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E. \end{aligned} \tag{30}$$

This problem has min $\theta^* = 3/4$ with alternate optima represented either by $\lambda_B^* = 1/2$ or by $\lambda_C^* = 1/4$ and all other $\lambda^* = 0$. For each of these optima, we have $\sum_{j=1}^n \lambda_j^* < 1$, so we utilize (28) and write

$$\begin{aligned} &\text{maximize} \quad \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E + \varepsilon(\hat{s}^- + \hat{s}^+) \\ &\text{subject to } \tfrac{3}{4} = 1\hat{\lambda}_A + \tfrac{3}{2}\hat{\lambda}_B + 3\hat{\lambda}_C + 4\hat{\lambda}_D + 4\hat{\lambda}_E + \hat{s}^-, \\ &\qquad 1 = 1\hat{\lambda}_A + 2\hat{\lambda}_B + 4\hat{\lambda}_C + 5\hat{\lambda}_D + \tfrac{9}{2}\hat{\lambda}_E \qquad\qquad \hat{s}^+, \\ &\qquad 1 \geq \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E, \end{aligned} \tag{31}$$

so that $\sum_{j=1}^n \hat{\lambda}_j^* \equiv \hat{\lambda}_A^* + \hat{\lambda}_B^* + \hat{\lambda}_C^* + \hat{\lambda}_D^* + \hat{\lambda}_E^*$ with all $\hat{\lambda}$ non-negative. Because both optimal solutions $\hat{\lambda}_B^* = 1/2$ and $\hat{\lambda}_C^* = 1/4$ and all other variables zero give $\sum_{j=1}^n \hat{\lambda}_j^* < 1$, it follows from theorem 4 that increasing returns to scale prevails at A.

We next turn to E in (29) as a point which is not on either (i) the BCC efficiency frontier represented by the solid lines in figure 2 or on (ii) the CCR efficiency frontier represented by the broken line from the origin. Substituting the coordinates for E in the CCR model (5), we obtain

$$\begin{aligned}
&\text{minimize} \quad \theta \\
&\text{subject to} \quad 4\theta \geq 1\lambda_A + \tfrac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E, \\
&\qquad\qquad \tfrac{9}{2} \leq 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \tfrac{9}{2}\lambda_E, \\
&\qquad\qquad 0 \leq \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E.
\end{aligned} \tag{32}$$

Again, we have alternate optima with, now, $\theta^* = 27/32$ for either $\lambda_B^* = 9/4$ or $\lambda_C^* = 9/8$ and all other $\lambda^* = 0$. Hence, in both cases we have $\sum_{j=1}^{n} \lambda_j^* > 1$. Proceeding in an obvious way, we next reorient the last constraint and the objective in (28) to obtain

$$\begin{aligned}
&\text{minimize} \quad (\hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E) - \varepsilon(\hat{s}^- + \hat{s}^+) \\
&\text{subject to} \quad \tfrac{27}{8} = 1\hat{\lambda}_A + \tfrac{3}{2}\hat{\lambda}_B + 3\hat{\lambda}_C + 4\hat{\lambda}_D + 4\hat{\lambda}_E + \hat{s}^-, \\
&\qquad\qquad \tfrac{9}{2} = 1\hat{\lambda}_A + 2\hat{\lambda}_B + 4\hat{\lambda}_C + 5\hat{\lambda}_D + \tfrac{9}{2}\hat{\lambda}_E - \hat{s}^+, \\
&\qquad\qquad 1 \leq \hat{\lambda}_A + \hat{\lambda}_B + \hat{\lambda}_C + \hat{\lambda}_D + \hat{\lambda}_E, \\
&\qquad\qquad 0 \leq \hat{\lambda}_A, \hat{\lambda}_B, \hat{\lambda}_C, \hat{\lambda}_D, \hat{\lambda}_E.
\end{aligned} \tag{33}$$

This has its optimum at $\lambda_B^* = 9/4$ or $\hat{\lambda}_C^* = 9/8$ with all other variables equal to zero and so, in conformance with theorem 3, as given in (27), we associate E with decreasing returns to scale.

There is confusion in the literature on the returns-to-scale characterizations obtained from theorems 2 and 3 and the BCC and the CCR models with which they are associated. Hence, we proceed a bit further as follows.

As noted earlier, returns to scale generally has an unambiguous meaning only for points on the efficiency frontier. When the BCC model is used on the data in figure 2, the primal model projects E into E′ with coordinates (7/2, 9/2) on the segment of the line $y = 1 + x$ which connects C to D on the BCC efficiency frontier. This result identifies E as having an inefficiency in the amount of 1/2 unit in its input. This is a technical inefficiency, as previously noted, in our discussion of (9). Turning to the dual for E formed from the BCC model as given in (9), we obtain a value of $u_o^* = -1/4$. This negative value of u_o^* suggests that returns to scale are either decreasing or constant at E′ = (28/8, 9/2), the point to which E is projected in order to obtain access to (26). Substitution in the latter model yields a value of $u_o^* = -2/7$, which is also negative, thereby also identifying E′ with the decreasing returns to scale that prevail on this portion of the efficiency frontier.

Next, we turn to the conditions specified in theorem 3 – opposite (27) – which are identified with the CCR model (5). Here, we find the projection is to a new point E″ = (27/8, 9/2) which is on the line $y = 4/3x$ corresponding to the broken line from the origin that coincides with the segment BC from B to C in figure 2. This ray from

the origin constitutes the efficiency frontier for the CCR model (5) which, when used in the manner we have previously indicated, simultaneously evaluates the technical, mix and returns-to-scale performances of E. In fact, as can be seen from the solution to (33), this evaluation is effected by either $\hat{\lambda}_B^* = 9/4$ or $\hat{\lambda}_B^* = 9/8$ – which are variables associated with vectors in a "constant returns-to-scale region" that we will shortly associate with "most productive scale size" (MPSS) for the BCC model. The additional 1/8 unit input reduction effected in going from E′ to E″ is needed to adjust to the efficient mix that prevails in this MPSS region which the CCR model is using to evaluate E.

The situation is general, as we will see. The CCR model simultaneously evaluates scale as well as mix and technical inefficiencies, while the BCC model separates out the scale inefficiencies for evaluation in the dual. Further, the scale evaluations in the BCC model are conducted "locally" by reference to DMUs like C and D in figure 2, whereas the CCR model effects its evaluations "globally" by reference to segments like BC.

In order to extend this analysis, we need to specify what is to be meant by returns to scale in multiple input-multiple output situations. For this, we utilize the concept of "Most Productive Scale Size" (MPSS) introduced by Banker (1984). To see what this means in multiple output-multiple input situations, consider the proportions represented by the scalars β, $\alpha \geq 0$ in

$$(X_o\alpha, Y_o\beta), \tag{34}$$

with X_o and Y_o representing input and output vectors, respectively. We can continue to move toward a possibly better (i.e., more productive) returns-to-scale situation as long as $\max \beta/\alpha \neq 1$. In other words, we are not at a point which is MPSS when either (a) *all* outputs can be increased in proportions that are at least as great as the corresponding proportional increases in *all* inputs needed to bring them about, or (b) all inputs can be decreased in proportions that are at least as great as the proportions in the accompanying reductions in all outputs. We will be at MPSS only when this is no longer possible. We will then have $\beta/\alpha = 1$ or $\alpha = \beta$, so returns to scale are constant at MPSS.

Recourse to prices, costs (or similar weights) would generally be required to determine a "best" or "most economical" scale size. Here, however, we are using the concept of MPSS in a way that avoids the need for such additional information by allowing all inputs and outputs to vary simultaneously in the proportions prescribed by α an β in (34). Hence, MPSS allows us to continue to confine attention to technical and mix inefficiencies, as before, while allowing for other possible choices after scale size possibilities have been identified and evaluated in our DEA analyses.

The interpretation we have just provided for (34) refers to returns to scale locally, as is customary. However, this does not exhaust the uses that can be made of Banker's MPSS. For instance, we can now replace our preceding local interpretation of (34) by

one which is oriented globally and we can do this in a way that enables us to relate the two approaches in theorems 2 and 3 to each other and thereby provide further insight into how the models in (5) and (9) enter into scale size (and other) evaluations. For these purposes, we introduce the following formulations:[27]

$$\begin{aligned} &\text{maximize } \beta/\alpha \\ &\text{subject to } \beta Y_o \le \sum_{j=1}^{n} Y_j \lambda_j, \\ &\qquad \alpha X_o \ge \sum_{j=1}^{n} X_j \lambda_j, \\ &\qquad 1 = \sum_{j=1}^{n} \lambda_j, \\ &\qquad 0 \le \beta, \alpha \text{ and } \lambda_j, \ j = 1, \ldots, n. \end{aligned} \tag{35}$$

As already noted, we are moving to a global interpretation of (34). We are also altering the characterization so that these α and β values now yield new vectors $\hat{X}_o = \alpha X_o$ and $\hat{Y}_o = \beta Y_o$ which we can associate with points which are MPSS as in the following theorem.

Theorem 5
When incorporated in (35), a *necessary* condition for DMU_o, with output and input vectors Y_o and X_o, to be MPSS is max $\beta/\alpha = 1$, in which case returns to scale will be constant.

This theorem enables us to use MPSS to bring our global interpretation into contact with the local returns-to-scale interpretations we previously supplied for (34) and follows readily from the fact that $\beta = \alpha = 1$ with $\lambda_j = \lambda_o = 1$ is a solution of (35), so that always max $\beta/\alpha = \beta^*/\alpha^* \ge 1$. See the appendix.

We illustrate with A = (1, 1) in figure 2, which we insert in (35) to obtain

$$\begin{aligned} &\text{maximize } \beta/\alpha \\ &\text{subject to } 1\beta \le 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \tfrac{9}{2}\lambda_E, \\ &\qquad 1\alpha \ge 1\lambda_A + \tfrac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E, \\ &\qquad 1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E, \\ &\qquad 0 \le \lambda_A, \ldots, \lambda_E. \end{aligned}$$

[27] B. Golany has called our attention to the earlier paper by Golany and Yu (1994), in which this same formulation appears. However, their use of it is very different from ours.

An optimum solution is $\lambda_B^* = 1$, with $\beta^* = 2$, $\alpha^* = 3/2$, all other $\lambda^* = 0$, so that $\beta^*/\alpha^* = 4/3 > 1$ and MPSS is not achieved.

Turning to D = (4, 5), we again utilize (35) to obtain

$$\begin{aligned}
&\text{maximize } \beta/\alpha \\
&\text{subject to } 5\beta \le 1\lambda_A + 2\lambda_B + 4\lambda_C + 5\lambda_D + \tfrac{9}{2}\lambda_E, \\
&\qquad 4\alpha \ge 1\lambda_A + \tfrac{3}{2}\lambda_B + 3\lambda_C + 4\lambda_D + 4\lambda_E, \\
&\qquad 1 = \lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_E, \\
&\qquad 0 \le \lambda_A, \ldots, \lambda_E.
\end{aligned}$$

This has an optimum at $\lambda_B^* = 1$, with $\alpha^* = 3/8$ and $\beta^* = 2/5$, to give $\beta^*/\alpha^* = 16/15$ as the maximum value of this ratio. It also has an alternate optimum with $\lambda_C^* = 1$ and $\alpha^* = 3/4$, $\beta^* = 4/5$ so, again, $\beta^*/\alpha^* = 16/15$. Evidently, all convex combinations of these two solutions are also optimal with solutions $\beta^*/\alpha^* = 16/15$. These additional optima are derived from the solutions composed from $\lambda_B = 1$ and $\lambda_C = 1$ which locate end-points where returns to scale are constant in figure 2.

The same constant returns to scale interval is used to evaluate all of the DMUs in figure 2 so that, in accordance with the above theorem, we have max $\beta/\alpha = 1$ only in this interval. Formally, let max $\beta_j/\alpha_j = \beta_j^*/\alpha_j^*$ when DMU_j is the DMU_o to be evaluated in (35). We then have

$$\underset{j=1,\ldots,n}{\text{minimize}} \left\{ \frac{\beta_j^*}{\alpha_j^*} \right\} = \frac{\beta_k^*}{\alpha_k^*} = 1. \tag{36}$$

We can now augment the preceding theorem to the following, which is proved in the appendix.

Theorem 6

For DMU_k to be MPSS, *both* of the following conditions must be satisfied:

$$\begin{aligned}
&\text{(i)} \quad \beta_k^*/\alpha_k^* = 1. \\
&\text{(ii)} \quad \text{All slacks are zero.}
\end{aligned} \tag{37}$$

This is intended to mean that the slacks in *all* alternative optima are zero and, to align this with (7), we note that this is all taken care of if the *maximum* sum of slacks is zero.

We have omitted slacks from the objective in (35) because we want to focus on other relations between (5) and (9).[28] To study these relations, we start with the following

[28] We also note that (35) is a fractional programming problem, so we can use the Charnes–Cooper transformation to transform it to an ordinary linear programming problem, just as was done in moving from (3) to (4). See the appendix.

Theorem 7 (Ahn, Charnes and Cooper)
If a DMU_o is found to be efficient with the CCR model, then it will also be found to be efficient with the BCC model.

Now we remark that the converse of this theorem is not necessarily true. See figure 2. It therefore follows that all points of intersection on the efficiency frontier will be common to both the CCR and BCC models, and this makes it possible to use either model to evaluate returns to scale. We also have

Theorem 8
An active vector[29] in an optimal basis is necessarily efficient.

These active vectors must therefore lie in the intersection between the efficiency frontiers of both the BCC and CCR models. Note, therefore, that the CCR models simultaneously evaluate both returns to scale and technical (and mix) inefficiencies and, as previously noted, the active vectors which evaluate performance of all DMUs in the CCR model are to be found in this intersection.

Next we appeal to the following

Theorem 9 (Banker and Thrall)

(i) $u_o^* > 0$ for all optimal solutions to (9) if and only if $\sum_{j=1}^n \lambda_j^* > 1$ for all optimal solutions to (5).

(ii) $u_o^* < 0$ for all optimal solutions to (9) if and only if $\sum_{j=1}^n \lambda_j^* < 1$ for all optimal solutions to (5).

(iii) $u_o^* = 0$ for some optimal solution to (9) if and only if $\sum_{j=1}^n \lambda_j^* = 1$ in some optimal solution to (5).

Via this theorem, we can remove the possibility that alternate optima (when present) might lead to different returns-to-scale characterizations. This is accomplished by simply noting that we must have $\sum_{j=1}^n \hat{\lambda}_j^* = \sum_{j=1}^n \hat{\hat{\lambda}}_j^*$ for (28)[30], by the definition of an optimum, even when some of the $\hat{\lambda}_j^*$ and $\hat{\hat{\lambda}}_j^*$ differ in alternate optimum solutions to (28). Hence, two different returns to scale characterizations cannot occur from (27) and, via theorem 9, the same is therefore true for (25) when (26) is used.[31]

[29] Active vectors are those which have non-zero coefficients as members of an optimal basis. See chapter 5 by Thrall for detailed developments and discussions of properties of non-basic solutions in which vectors with non-zero coefficients may appear as part of an optimum.

[30] Or its modification when $\sum_{j=1}^n \lambda_j^* > 1$.

[31] Such differences can occur, however, when input (minimization) orientations are replaced by output (maximization) orientations in the DEA models used. See Golany and Yu (1994) for an attempt to exploit these differences.

Finally, we can relate (35) to (5) by virtue of the following

Theorem 10
Let $\hat{\lambda}_j^*, \beta^*, \alpha^*$ designate an optimal solution to (35). We then have

(1) Dividing by β^* and setting $\hat{\lambda}_j^* = \lambda_j^*/\beta^*$, we obtain $\sum_{j=1}^n \hat{\lambda}_j^* = 1/\beta^*$, with $\hat{\theta}^* = \alpha^*/\beta^*$ optimal for (5).

(2) Dividing by α^* and setting $\hat{\lambda}_j^* = \lambda_j^*/\alpha^*$, we obtain $\sum_{j=1}^n \hat{\lambda}_j^* = 1/\alpha^*$, with $\hat{\phi} = \beta^*/\alpha^*$ optimal for (10).

The converse of this theorem is also true. See the appendix for proofs. Given an optimal solution to (5), we can therefore obtain an optimal solution to (35) and we therefore have a way of moving back and forth between these models. Also, returning to the way technical and mix inefficiencies were distinguished in the discussion immediately following (12), we can note that (35) allows us to make this distinction in inputs and outputs simultaneously.

8 Returns to scale: quantitative measures

We now turn from qualitative characterizations to quantitative measures. One possibility is provided by (35). The value of max $\beta/\alpha = \beta^*/\alpha^*$ is invariant to the units in which inputs and outputs are measured, which is a property required for the elasticity measures used in economics. See below. It also allows the measures of technical inefficiency supplied by β^* and α^* to be applied, component by component, to each input and output. However, these measures must be supplemented to allow for mix inefficiencies, as in the following formulas:

$$\begin{aligned} \beta^* y_{ro} + s_r^{+*} &= y_{ro}^*, \qquad r = 1, \ldots, s, \\ \alpha^* x_{io} - s_i^{-*} &= x_{io}^*, \qquad i = 1, \ldots, m, \end{aligned} \tag{38}$$

which are more symmetric than the CCR projection formulas given earlier in (8) and provide a way of distinguishing between "technical" and "mix" inefficiencies in outputs and inputs simultaneously.[32] These y_{ro}^* and x_{io}^* are coordinates of a point on the efficiency frontier which is also MPSS, so the projections in (38) differ from those in (8), which are local. It follows that β^*/α^* also provides a measure of scale elasticity which relates the point being evaluated to a point which is MPSS.

Interest is usually directed to local measures of returns to scale for which we might take advantage of the different properties of the CCR and BCC models to develop a measure of the form

$$h^*/z^*, \tag{39}$$

[32] See the discussion following (12).

where $h^* = \theta^*$ is optimal for (5) and $z^* = \theta_o^*$ is optimal for (9). This is an approach which is elaborated on by Färe et al. (1985). See also Banker et al. (1984), as well as Banker et al. (1995) and Zhu and Shen (1995). However, this approach does not deal explicitly with the possibility of alternate optima, and an assumption of uniqueness associated with (38) and (39) can cause these measures to fall short of what is required in considering movement intended to take advantage of scale economics or remove scale diseconomies.

To show what is involved in such considerations, we utilize the development in Banker and Thrall (1992)[33] and return to the problem on the right in (9), from which we obtain

$$\sum_{r=1}^{s} \mu_r^* y_{ro} + u_o^* = 1 = \sum_{i=1}^{m} v_i^* x_{io} \tag{40}$$

on the assumption that efficiency has been achieved.[34] Then we introduce the following new variable:

$$\rho_o^* = \frac{\sum_{i=1}^{m} v_i^* x_{io}}{\sum_{r=1}^{s} \mu_r^* y_{ro}} = \frac{1}{\sum_{r=1}^{s} \mu_r^* y_{ro}} = \frac{1}{1 - u_o^*}, \tag{41}$$

which we may define as an elasticity. When the solutions are not unique, we can develop bounds via

$$\rho_o^+ = \min\left\{ \frac{1}{1 - u_o^+} \right\} \quad \text{and} \quad \rho_o^- = \max\left\{ \frac{1}{1 - u_o^-} \right\} \tag{42}$$

to obtain

$$\rho_o^- \leq \rho_o^* \leq \rho_o^+. \tag{43}$$

We obtain these u_o^+ and u_o^- values from (26) in the following manner. First we omit the constraint $u_o \leq 0$ and obtain max $u_o = u_o^+$. Then we continue with the thus modified problem and reorient the objective to obtain min $u_o = u_o^-$. Note that when this is done we can obtain solutions with $u_o^+ = -\infty$ which gives $\rho_o^- = 0$ as a lower bound in (43). We can also obtain $u_o^+ = 1$ which gives $\rho_o^+ = \infty$ as an upper bound.

We elucidate by returning to A in figure 2 and obtain $\mu^* = 0$ by substitution in (26) *after* eliminating the condition $u_o \geq 0$. This give $u_o^+ = 1 - \mu^* = 1$ and produces $\rho_o^+ = 1/(1 - u_o^+) = \infty$ as an upper bound. Then replacing the objective in (26) by min u_o in its modified version, we obtain $\mu^* = 1/2$ so $u_o^- = 1 - \mu^* = 1/2$ and $\rho_o^- = 1/(1 - u_o^-) = 2$ places a lower bound on the returns to scale at A. That is, we have

$$2 = \rho_o^- \leq \rho_o^* \leq \rho_o^+ = \infty \tag{44}$$

[33] See also Banker et al. (1984).

[34] Achievement of efficiency can always be arranged, as was noted in the developments leading to (26) and (28).

to bound the value of ρ_o^* at A. Then, turning to D = (4,5) in figure 2 we similarly obtain

$$0 = \rho_o^- \leq \rho_o^* \leq \rho_o^+ = 4/5. \tag{45}$$

To see what these results mean, we return to (34) and again interpret these α and β values in terms of returns to scale locally. Note, for instance, that a one unit reduction in input at point D is accompanied by a one unit reduction in output. The *proportionate* reductions are therefore $\alpha = 1/4$, $\beta = 1/5$, so $\beta/\alpha = 4/5$ shows that the proportionate reduction in input exceeds the proportionate reduction in output and therefore decreasing returns to scale would be experienced by movement in this direction from D. Turning next to A in figure 2, we find that a unit increase in output can be secured from a 1/2 unit increase in input by movement along this portion of the frontier from A to B. Hence, we have $\beta/\alpha = 2$ and returns to scale are increasing in this direction.

For this single output-single input example, we have $\alpha = \Delta x/x$ and $\beta = \Delta y/y$ so $x\alpha = \Delta x$ and $y\beta = \Delta y$, while $\beta/\alpha = (\Delta y/y)/(\Delta x/x) = (x/y)(\Delta y/\Delta x)$ from (34). This is expressed in a manner easily identified with the elasticity measure used in economics. The uniqueness usually assumed for this measure in economics may be missing because our functions, although continuous, are not analytic and hence (i) the limit of this expression need not exist, and (ii) its value will, in general, depend on the intended direction of movement. Evidently the ρ_o^+, ρ_o^- in (42) and (43) provide information that is needed when input increases or decreases are to be considered for returns-to-scale properties at all extreme points. Indeed, in figure 2 only the point E′ is associated with a *unique* value in which $\rho_o^+ = \rho_o^- = \rho_o^*$.

Banker's MPSS concept allows us to extend this formulation to the case of multiple outputs and inputs. Interpreting y and x as "virtual" outputs and inputs, as defined in (2), we characterize α and β as follows. For increasing returns to scale, the value of $\alpha = \Delta x/x$ means that all inputs are increased by *at most* this proportion and all outputs are increased in *at least* the proportion $\beta = \Delta y/y$, with $\beta/\alpha > 1$ and at least one output and one input achieves its bound of β or α. Turning to decreasing returns, the value $\beta = \Delta y/y$ represents the maximal proportions in which all outputs are decreased with decreases in all inputs of at least $\alpha = \Delta x/x$ and $\beta/\alpha < 1$ (the proportionate decreases in inputs exceed the proportionate decreases in outputs). MPSS is reached only when $\beta/\alpha = 1$ *and*, in (i), the case of increasing returns to scale, when a further proportionate increase in inputs is associated with a situation in which at least one output will exhibit a smaller than proportionate increase in its value while, for (ii), decreasing returns to scale, at least one input will exhibit a smaller than proportionate decrease than is exhibited by the outputs.

To continue, we examine the lower bound of zero in (45) and the infinite upper bound in (44). To help interpret these results, we note that the relations in (40) correspond to algebraic expressions for hyperplanes in n-dimensional spaces. In the two-dimensional space of figure 2, these hyperplanes become straight lines which we rotate around points like A and D by varying u_o^* until coincidence is achieved with an adjacent frontier segment.

These extremes, as represented in u_o^+ and u_o^-, are reflected in the ρ_o^+ and ρ_o^- in (42) which provide the above elasticity measures and, once again, we have DEA as a data-based technique. The lower bound $\rho_o^- = 0$ in (45), for example, means that no increase in output is evidenced by an increase in input starting at D. Hence, the relation associated with ρ_o^- in (45) means that the corresponding line (= hyperplane in (41)) is rotated until it coincides with the dotted line extension from D that is portrayed in figure 2. Similarly, the upper bound given by $\rho_o^+ = \infty$ in (44) means that returns to scale are infinite on the segment which coincides with the dotted line extending vertically downward from A. The data, as given, are thus interpreted to mean that even an infinitesimal decrease in input at A will reduce output to zero or, conversely, a reversal of this infinitesimal input decrease will result in an output jump from zero to unity.[35] Therefore, to finally align our analysis with other parts of this paper, we bring our non-Archimedean infinitesimal into play by writing our proportional input increment as $\Delta x/x = \varepsilon/x$ so that

$$\beta/\alpha = \frac{x}{y}\frac{\Delta y}{\Delta x} = \frac{x}{y}\frac{\Delta y}{\varepsilon} = M\frac{\Delta y}{y}x,$$

with $\varepsilon^{-1} = M$, the so-called "big M" of ordinary linear programming (which is also non-Archimedean) represented as the reciprocal of $\varepsilon > 0$ when infinite returns to scale occurs and, of course, we have

$$\frac{x}{y}\frac{\Delta y}{\Delta x} = \beta/\alpha = \frac{x}{y}\frac{0}{\Delta x} = 0$$

when zero returns to scale is exhibited.

These developments, which rely heavily on Banker and Thrall (1992), evidently supply guidance beyond what was previously available. This is not the end of what is needed, however, and more work is under way. Golany and Yu (1994), for example, proceed in a different manner to establish bounds which we briefly indicate as follows. The projections obtained from (5) will generally yield points that differ from the projections obtained when the output oriented model (20) is used. Using these different projections as bounds, they then try to squeeze them together until a point is reached where their scale elasticity values coincide. Other refinements could explore alternative directions to find one that yields the greatest advantage – with, perhaps, changes in the mix of inputs and outputs needed to secure them.

9 Allocative efficiencies and assurance region extensions

Technical or mix inefficiencies are present when some input or output may be improved without worsening any other input or output. Removal of returns-to-scale

[35] See footnote 16.

inefficiencies, on the other hand, involves movements on the efficiency frontier where input-output tradeoffs are necessary. To take advantage of increasing returns to scale, for example, it is necessary to augment inputs in order to achieve a more than proportional increase in outputs. Use of such an opportunity implies that the output augmentations are at least as valuable as the input increases needed to secure them. Conversely, the frontier movements needed to eliminate decreasing returns to scale imply that the output reductions are less valuable than the more than proportional decreases in inputs required. Note, however, that this can all be justified if inequality exists so that total returns exceed total costs. Information on exact values of costs and prices is not required.

The topic of "allocative efficiency" to which we now turn extends the required valuations a good deal beyond such inequality bounds. In particular, as we will see, the information (and results) depends on exact knowledge of the relevant valuations with inequalities replaced by equations which are to be satisfied by the choices to be made.

Figure 3 can help to relate what is being said to standard versions of micro-economic theory if we interpret the solid line from C to E as an isoquant.[36] That is, we interpret this line as representing all of the technically efficient input combinations of two inputs in amounts (x_1, x_2) that can produce a single output in amount y.

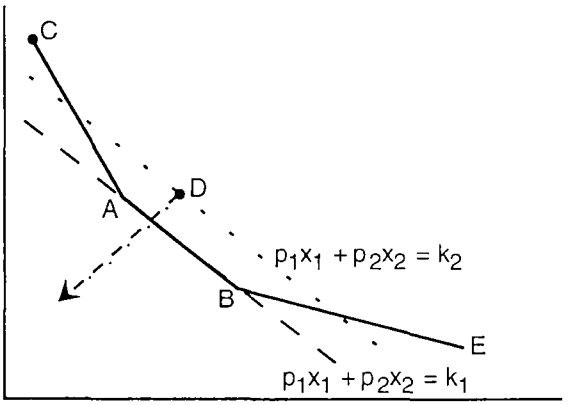


Figure 3. Allocative efficiency.

The amount of this scalar is obtained by passing a plane at some prescribed level y through the production surface associated with

$$y = f(x_1, x_2), \tag{46}$$

[36] We could replace this isoquant assumption with the more general concept of a "unitized frontier", as discussed in Cooper et al. (1995), but here we prefer to adhere to the concepts commonly used in micro-economic theory. See chapter 2 in E. Rhodes (1978) for a detailed development.

where $f(\cdot)$ is a "production function" so that, in accordance with the usual assumptions in micro-economics, the output level y is maximal for any input combination (x_1, x_2) that is used.

We now introduce a cost function,

$$p_1 x_1 + p_2 x_2 = k, \tag{47}$$

where the positive constants (p_1, p_2) represent the prices per unit of the corresponding inputs. Geometrically this expression corresponds to downward sloping lines like those associated with k_2 and k_1 in figure 3. If we seek to minimize the cost of producing y, we achieve the usual results from micro-economics

$$-\frac{dx_2}{dx_1} = \frac{\partial f/\partial x_1}{\partial f/\partial x_2} = \frac{p_2}{p_1}, \tag{48}$$

which represent necessary conditions for a minimum in the open portion of any segment such as A–B in figure 3 where these derivatives exist.

The p_1 and p_2 (unit prices) on the right in (48) can be regarded as components of the normal vector represented by the arrow in figure 3 that shows the direction in which optimization is undertaken. Evidently, cost is minimized at $k = k_1$ because the normal to the segment A–B with components $\partial f/\partial x_1$ and $\partial f/\partial x_2$ is equal to the normal of the cost line and because coincidence is achieved with the segment of the isoquant represented by A–B and no further movement in the direction of the normal is possible. In contrast to assumptions like "an absence of technical inefficiency", which is customary in micro-economics, we need to allow for additional possibilities but we can also use micro-economics to identify issues we need to consider.

First, we observe that we cannot guarantee that empirical data will necessarily conform to the assumptions required to justify characterizing the solid lines CABE as an isoquant in the sense of micro-economics. Hence we regard the conditions noted in (48) as *necessary* but not sufficient to guarantee that minimum cost has been attained for the production of y and that mix inefficiencies have thereby also been eliminated. Second, we observe that technical efficiency is *necessary* but not sufficient for cost minimization. Note, for instance, that D is not technically efficient but is nonetheless preferred to C even though the latter is technically efficient. Finally, we observe that we can continue to obtain improvements at D by moving this line in the direction indicated by the arrow until the segment A–B is achieved, where (in the sense of DEA) technical and mix efficiency are both satisfied. Thus, as stated, technical efficiency is a necessary but not a sufficient condition for "allocative efficiency".

Although allocative efficiency can help to clarify matters and provide conceptual guidance, it can be of limited value in actual applications because it (a) imposes severe data requirements, and (b) utilizes assumptions which may be difficult to justify. Exact knowledge of prices is often difficult or impossible to come by and this difficulty is compounded when one has to deal with entities like schools, hospitals or

air force units where inputs and/or outputs have no easily ascertained costs or prices.[37] In addition, prices can be (and often are) subject to variation in very short periods so that additional choices and assumptions are involved concerning their pertinence.

One route around the latter problem involves a recourse to averages. Another possibility is to introduce constraints with lower and upper bounds on the admissible values of variables. This has been done for both primal and dual variable values. See Arnold et al. (1995) and Cooper et al. (1994). Here, however, we follow the Assurance Region approach first developed in Thompson et al. (1986)[38] and defined more precisely in Thompson et al. (1990). We confine attention to the dual. In this 1986 paper, the assurance region took forms like

$$\begin{aligned} \alpha_r \le \frac{v_r}{v_{r_o}} \le \beta_r, \qquad & r = 1, \ldots, s, \\ \delta_i \le \frac{\mu_i}{\mu_{i_o}} \le \gamma_i, \qquad & i = 1, \ldots, m, \end{aligned} \tag{49}$$

where v_{r_o} and μ_{i_o} represent dual variables which serve as "numeraires" in establishing the upper and lower bounds represented here by α_r, β_r and by δ_i, γ_i for the dual variables associated with each output and input and where $\alpha_{r_o} = \beta_{r_o} = \delta_{i_o} = \gamma_{i_o} = 1$.[39] Uses of such bounds are not restricted to prices and may extend to "utils" or any other evaluations that are regarded as pertinent. For an example of the former, see chapter 14 by Zeng which reports an application that utilizes assurance region approaches to impose bounds on proposed vehicle designs and production schedules for Chinese automobiles. See also chapter 15 by Zhu, which uses this approach to establish bounds on the weights obtained from uses of Analytic Hierarchy Processes in Chinese textile manufacturing.

There is another approach called the "cone-ratio envelopment approach" which can also be used for this purpose. See Cooper et al. (1994). We do not examine this approach in detail, but rather only note that the assurance region approach can also be given an interpretation in terms of cones. To show how this may be done, we use the following matrix representation:

$$\begin{bmatrix} D & O \\ O & C \end{bmatrix} \begin{bmatrix} \mu \\ v \end{bmatrix} \le 0, \tag{50}$$

[37] Evaluations of air force activities have involved consideration of inputs like "weather", where no markets could be referenced.

[38] This application is of interest in its own right since it dealt with evaluations of contributions to "fundamental knowledge" in physics, where neither price nor similar guides could be obtained.

[39] Assurance Regions described by (49) are a special case of "cone ratios" in "intersection form". See Charnes et al. (1990, pp. 77 and 78). Later, Thompson et al. (1995, p. 112) described (49) as "cone ratios in intersection form defined by pairwise comparisons where the first member of each pair is a common numerator".

where μ and ν are non-negative vectors with component values to be determined. The following example taken from chapter 15 by Zhu can serve to illustrate a use of the submatrix D:

$$\begin{bmatrix} -4 & 0 & 1 \\ 2 & 0 & -1 \\ 0 & 8 & -7 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \leq 0; \quad \text{i.e.} \quad \begin{aligned} -4\mu_1 \quad &+ \quad \mu_3 \leq 0, \\ 2\mu_1 \quad &- \quad \mu_3 \leq 0, \\ 8\mu_2 \; &- 7\mu_3 \leq 0, \end{aligned} \tag{51}$$

so that using μ_3 as "numeraire" we have $1/4 \leq \mu_1/\mu_3 \leq 1/2$ and $0 \leq \mu_2/\mu_3 \leq 7/8$, in conformance with (49).

Evidently, $D\mu \leq 0$ with $\mu \geq 0$ defines an "input cone". Similarly $C\nu \leq 0$ with $\nu \geq 0$ in (50) defines an output cone. This is not the end of the line for uses of these assurance region concepts, however, because Thompson et al. (1990) subsequently extended this to include formulations like

$$\begin{bmatrix} D & O \\ O & C \\ F_1 & F_2 \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \tag{52}$$

in which the sub-matrices F_1 and F_2 serve as "linkage constraints" – so-called because they link input conditions to output conditions – but not in cone ratio forms.[40] See Thompson et al. (1995) for a chronology and a diagrammatic portrayal of the various assurance region (and related) approaches that are now available.[41]

The generality of these formulations is evident. They also provide flexibility in use. Prices, utils and other measures may be accommodated and so can mixtures of such concepts. Moreover, one can exploit possible interplays to first examine provisional solutions and then tighten or loosen the bounds until one or more solutions is attained that appears to be reasonably satisfactory. Such bounds might take forms like $\underline{p_i} \leq \mu_i/\mu_r \leq \overline{p_i}$, where $\underline{p_i}$ and $\overline{p_i}$ represent lower and upper bounds on the ith relative price. For the $i = 1, 2$ inputs in figure 3, these bounds might continue to identify a point on the segment A–B as optimal. Alternatively, they might locate a point on the segment A–C associated with a value k_i which is very close to k_1 and, in any case, these bounds might be tightened or loosened in order to explore the properties of other solutions.

Even ab initio, the Assurance Region approach greatly relaxes the conditions and widens the scope for use of a priori conditions. In some cases, the conditions to be

[40] The expression (52) does not have separate input and output cones and hence is not in the form of a cone ratio.

[41] See also their discussion of modifications of (3) and (5), which can be used to measure profit potential as well as efficiency.

comprehended may be too complex for explicit articulation, in which case additional possibilities are available from other recent advances. For instance, *instead* of imposing bounds on allowable variable values, the cone-ratio envelopment approach *transforms* the data. Brockett et al. provide an example in which a bank regulatory agency wanted to evaluate "risk coverage" as well as the "efficiency" of the banks under its jurisdiction. Bounds could not be provided on possible tradeoffs between risk coverage and efficiency, so this was accomplished by using a set of banks identified as "excellent" (even when they were not members of the original (regulatory) set) and then, employing data from these excellent banks, a cone-ratio envelopment was used to transform the data into improved values that could be used to evaluate each of the regulated banks operating under widely varying conditions. This avoided the need for jointly specifying what was meant by "adequate" risk coverage and efficiency not only in each detail, but also in all of the complex interplays between risk and efficiency that are possible in bank performances. The non-negativity imposed on the slacks in standard DEA models was also relaxed. This then made it possible to identify deficiencies which were to be repaired by increasing expense items such as "bad loan allowances" (as needed for risk coverage) even though this worsened efficiency as evaluated by the transformed data.

Again, this is not the end of the line. In one very recent effort, Thannasoulis and Allen (1994) note that adding *restrictions* like (52) to "multiplier models" implies adding *variables* to the "envelopment models".[42] Proceeding further on this line, they then show how to synthesize new (artificially contrived) DMUs to obtain still more information than is obtained from straightforward uses of assurance region approaches. This opens possibilities for joint uses of this approach to envelopment models in combination with assurance region approaches. Opportunities also exist for joint uses of Assurance Region and Cone Ratio Envelopment approaches which, in turn, can be extended by introducing bounds on allowable values for the variables in the primal (envelopment) model of DEA as well as on the dual (multiplier) model. See Arnold et al. (1996).

10 Conclusion: chapter summaries and suggestions for further research

We have now covered much ground in which past formulations have been extended to provide new uses for DEA. We have also outlined issues for further research to which we now make additions with suggestions pointing toward the other papers in this volume.

One such addition could extend DEA for use in evaluating "returns to scope" as well as "returns to scale". A beginning in this direction may be found in Maindiratta (1990). Zeng's theorem as given in chapter 14 of the present volume of the *Annals of Operations Research* could also be useful because the evidence from the zero

[42] See also Roll and Golany (1991).

outputs in one or more DMUs could play a critical role in evaluating returns-to-scope possibilities that could extend his studies in Chinese vehicle production. See also chapter 15 by Zhu, and see also the still different approach opened in chapter 8 by Banker and Morey.[43)]

It would be even better if this could be accompanied by dynamic extensions which could make it possible to evaluate trends in "mix" as well as "productivity" (technical efficiency) possibilities. See the Moving Frontier Analyses and the related Isocost Curves in chapter 9 by Sinha.[44)] Needless to say, such dynamic extensions could prove useful in many other ways. The Malmquist Index studies reported in chapter 10 by Althin, Färe and Grosskopf add important new dimensions for index number uses by making it possible to distinguish between (a) new frontier possibilities that may be opened over time; (b) efficient and inefficient uses of such new frontiers; (c) identification of when these new possibilities and their uses occur; (d) the identities of the DMUs that used (and failed to use) these possibilities.

Another range of topics could involve choices of weights for use in DEA models along lines like those suggested in chapter 5 by Thrall. As noted in our earlier discussion, this, in turn, raises more general questions of desirable properties of measures to be used along lines like those dealt with in chapters 3 and 4 by Pastor and Thrall.

In a somewhat different direction, one might concentrate research on difficult-to-conceptualize methods of measuring some of the more subtle properties of inputs and outputs. "Quality" of inputs and outputs is one example. Another is "complexity", which is dealt with in the DEA formulations used by Sinha in chapter 9. One can also easily add to this list and include topics like "flexibility", etc. *En route*, one might study how DEA responds to misspecifications arising from either (i) an omission of pertinent variables, or (ii) an inclusion of variables which are not pertinent. See chapter 11 by Banker, Chang and Cooper in this volume.

In earlier discussions, we cited chapter 7 by Thompson, Dharmapala, Diaz, González-Lima and Thrall which exploits new methods of sensitivity analyses in which *all* data can be varied simultaneously. A related research direction involves suitable methods for statistical inference and probabilistic characterizations (and interpretations) for use with DEA. One part of the DEA literature has dealt with properties of statistical consistency as discussed in chapter 11 by Banker, Chang and Cooper. Other parts of the literature have been directed to developing nonparametric statistics that can match the non-parametric nature of DEA itself. See Brockett and Golany (1995).

Other problems and possibilities arise very naturally in such extensions. One class of problems concerns distinguishing between managerial error and statistical error. This is dealt with in chapter 12, where Arnold, Bardhan, Cooper and Kumbhakar study ways in which DEA and statistical regressions might be combined.

[43)] See also Ray (1995).

[44)] See also the recently released book by Sengupta (1995).

There is also the problem of how to deal with performance evaluations when risk and uncertainty are present in significant ways. Chapter 13 by Cooper, Huang and Li addresses this topic from the side of both the "evaluator" and the "evaluated" by suggesting "satisficing" rather than "optimizing" models for performance evaluations in a chance-constrained programming framework for DEA.[45)]

More generally, one must be concerned with properties of DEA solutions and the methods (e.g., algorithms) used to obtain them or to study their properties. This is dealt with in detail in the chapters by Thrall and by González-Lima, Tapia and Thrall which appear in Part II of this volume. The interior point algorithm developed in the latter chapter goes beyond achieving solutions in ways that are not restricted to uses of extreme points. It also opens possibilities for other uses. For instance, the "analytic center" solutions achievable by this algorithm are put to use in exhibiting the stability (= robustness) of DEA solutions in the immediately following chapter 7 by Thompson, Dharmapala, Diaz, González-Lima and Thrall. These robustness properties refer to situations in which *all data are varied simultaneously*. These kinds of variations are more general than the ones used in outlier analyses in statistics and the algorithms used for the sensitivity analyses in parametric linear programming where data variations are studied on only one observation at a time. Proceeding on these lines, data from international oil companies are used in chapter 7 to show the robustness of DEA in distinguishing between efficient and inefficient performance and, of course, the methods of analysis are formulated so they can be used in other studies as well.

Evidently, richness and variety are being added to DEA in the chapters we have been describing. There is also a need for unification of the various DEA models, and this is serviced by Yu, Wei and Brockett in chapter 2, who show how families of different DEA models can be obtained from one basic model formulation. Finally, there is the very important topic of new applications and the new problems and possibilities flowing from them. Chapters 14 and 15 by Zhu and Zeng, respectively, in Part V of this volume provide examples from the mixed economy of China which show how bounding techniques (such as "assurance region" approaches) can be used to deal with problems arising because non-market data and subjective evaluations must be dealt with.

Applications of DEA have been a source of new and important insights into policy shortcomings, which in turn have suggested new topics for research. An example is provided in a paper by Arnold et al. (1996), where their use of DEA in a study of Texas public schools found that not a single "excellent" school was also "efficient". Generally located in more favored districts, these "excellent" schools expended excessive resources, at least in part because the excellence evaluations did not take resource consumption into account. This inattention to resource consumption as part of an excellence evaluation is not confined to Texas. It is to be found in the "Clinton

[45)] This has subsequently been extended to include "joint" as well as "marginal" chance constraints. See Cooper et al. (1996).

2,000 Plan" for public-school education, and the "Bush–Alexander" plan (which preceded it) also focused on outputs without attention to the resources used to achieve specified standards of excellence. Something might be said for an alternative approach in which efficiency is also rewarded, perhaps, in a conditional fashion, so that schools functioning in difficult circumstances might be helped to achievable levels of excellence in a step-by-step manner. The evolution of practical methods for effecting such evaluations and accompanying rewards could also be a topic for further research along lines like those indicated in this volume.

Appendix

Proof of theorem 10

For this purpose, we replace (35) with the following equivalent:

$$\begin{aligned} &\text{minimize} \quad \alpha/\beta \\ &\text{subject to} \quad \beta Y_o \leq \sum_{j=1}^{n} Y_j \lambda_j, \\ &\qquad \alpha X_o \geq \sum_{j=1}^{n} X_j \lambda_j, \\ &\qquad 1 = \sum_{j=1}^{n} \lambda_j, \\ &\qquad 0 \leq \alpha, \beta, \lambda_j, \ j = 1, \ldots, n. \end{aligned} \tag{A.1}$$

This is a fractional programming problem so, proceeding as in (4), we introduce the new variables

$$\begin{aligned} &\hat{\beta} = t\beta = 1 \quad \text{so} \quad t = \frac{1}{\beta} > 0, \\ &\theta = \alpha/\beta, \\ &\hat{\lambda}_j = t\lambda_j = \lambda_j/\beta. \end{aligned} \tag{A.2}$$

Thus, multiplying all constraints by $t > 0$ in (A.1) gives

$$\begin{aligned} &\text{minimize} \quad \theta \\ &\text{subject to} \quad \hat{\beta} Y_o = Y_o \leq \sum_{j=1}^{n} Y_j \hat{\lambda}_j = \sum_{j=1}^{n} Y_j \lambda_j/\beta, \\ &\qquad \theta X_o = \frac{\alpha}{\beta} X_o \geq \sum_{j=1}^{n} X_j \hat{\lambda}_j = \sum_{j=1}^{n} X_j \lambda_j/\beta, \\ &\qquad \frac{1}{\beta} = \sum_{j=1}^{n} \hat{\lambda}_j = \sum_{j=1}^{n} \lambda_j/\beta, \end{aligned} \tag{A.3}$$

with all variables also constrained to be non-negative. Since $\beta > 0$ is defined by $\sum_{j=1}^{n} \hat{\lambda}_j = \sum_{j=1}^{n} \lambda_j/\beta$, this last constraint is redundant and may be omitted to obtain

$$\begin{aligned} &\text{minimize} \quad \theta \\ &\text{subject to} \quad Y_o \leq \sum_{j=1}^{n} Y_j \hat{\lambda}_j, \\ &\qquad\qquad \theta X_o \geq \sum_{j=1}^{n} Y_j \hat{\lambda}_j, \\ &\qquad\qquad 0 \leq \hat{\lambda}_j, \ j = 1, \ldots, n. \end{aligned} \tag{A.4}$$

This is in the same form as (5). So, via the theory of fractional programming, as given in Charnes and Cooper (1962), we find that a solution to (A.4) is optimal for (A.1) with $\min \alpha/\beta = \min \theta$. We may also move back and forth via the transformations given in (A.2) as claimed in (i) of theorem 9. Proof of part (ii) follows an analogous route and the necessity and sufficiency conditions of theorem 6 are an immediate consequence of the conditions for efficiency of a solution to (5), as specified in (7). □

Finally, we show that the converse is also true. Consider any optimal solultion to (5) with $\sum_{j=1}^{n} y_j \lambda_j^* = 1/\beta^*$. Setting $\theta^* = \alpha^*/\beta^*$ and multiplying all constraints by $\beta^* > 0$, we have a solution which is optimal for (A.1). This follows because assuming a solution to (A.1) with $\alpha/\beta < \alpha^*/\beta^*$ contradicts the assumption that θ^* is optimal for (5). □

Acknowledgements

We are grateful to J.T. Pastor and R. Färe for comments and suggestions. Support from the IC^2 Institute of The University of Texas at Austin is also gratefully acknowledged.

References

Ahn, T., A. Charnes and W.W. Cooper, A note on the efficiency characterizations obtained in different DEA models, Socio-Economic Planning Sciences 23, 1989, 253–257.

Aida, K., W.W. Cooper and J.T. Pastor, Evaluating water supply services in Japan with RAM – A range-adjusted measure of efficiency, Omega, 1997, to appear.

Ali, I. and L.M. Seiford, Translation invariance in Data Envelopment Analysis, Operations Research Letters 9, 1990, 403–405.

Arnold, V., I. Bardhan and W.W. Cooper, A two-stage approach for identifying and rewarding efficiency in Texas secondary schools, in *IMPACT Essays in Honor of George Kozmetsky*, W.W. Cooper, D. Gibson, F.Y. Phillips and S. Thore, eds., Greenwood Press, Boston, 1996.

Arnold, V., I. Bardhan, W.W. Cooper and A. Gallegos, Primal and dual optimality in computer codes using two-stage solution procedures in DEA, in *Operations Research: Methods, Models and Applications*, Jay Aranson and S. Zionts, eds., Kluwer Academic, Boston, 1996.

Banker, R.D., Estimating most productive scale size using Data Envelopment Analysis, European Journal of Operational Research 17, 1984, 35–44.

Banker, R.D., I. Bardhan and W.W. Cooper, A note on returns to scale in DEA, European Journal of Operational Research 88, 1996, 583–585.

Banker, R.D., H. Chang and W.W. Cooper, Equivalence and implementation of alternative methods for determining returns to scale in DEA, European Journal of Operational Research, 1995.

Banker, R.D., A. Charnes and W.W. Cooper, Models for estimating technical and scale efficiencies in DEA, Management Science 30, 1984, 1078–1092.

Banker, R.D. and W.W. Cooper, Validation and generalizations of DEA and its uses, TOPS, Sociedad Española de Estadistica e Investigación Operativa, Madrid, 1994.

Banker, R.D. and R.M. Thrall, Estimating most productive scale size using Data Envelopment Analysis, European Journal of Operational Research 62, 1992, 74–84.

Bardhan, I., W.F. Bowlin, W.W. Cooper and T. Sueyoshi, Models and measures for efficiency dominance in DEA, Part I: Additive models and MED measures, Journal of the Operations Research Society of Japan, 1995.

Bardhan, I., W.F. Bowlin, W.W. Cooper and T. Sueyoshi, Models and measures for efficiency dominance in DEA, Part II: FDH and Russell measures, Journal of the Operations Research Society of Japan, 1996.

Brockett, P.L., A. Charnes, W.W. Cooper, Z. Huang and D.B. Sun, Data transformations in DEA cone-ratio approaches for monitoring bank performance, European Journal of Operational Research, 1997, to appear.

Brockett, P.L., W.W. Cooper, H-C. Shin and Y. Wang, Inefficiency and congestion in Chinese production before and after the 1978 economic reforms, Socio-Economic Planning Sciences, 1997, to appear.

Brockett, P.L. and B. Golany, On some applications of rank statistics in DEA, Management Science, 1995.

Byrnes, P., R. Färe and S. Grosskopf, Measuring productive efficiency: An application to Illinois strip mines, Management Science 30, 1984, 671–681.

Charnes, A. and W.W. Cooper, *Management Models and Industrial Applications of Linear Programming*, Wiley, New York, 1961.

Charnes, A. and W.W. Cooper, Preface to topics in Data Envelopment Analysis, Annals of Operations Research 2, 1985, 59–94.

Charnes, A. and W.W. Cooper, Programming with linear fractional functionals, Naval Research Logistics Quarterly 9, 1962, 181–186.

Charnes, A. and W.W. Cooper, The strong Minkowski–Farkas–Weyl theorem for vector spaces over ordered fields, Proceedings of the National Academy of Sciences 44, 1958, 1–3.

Charnes, A., W.W. Cooper, B. Golany, L. Seiford and J. Stutz, Foundations of Data Envelopment Analysis for Pareto–Koopmans efficient empirical production functions, Journal of Econometrics 30, 1985, 91–107.

Charnes, A., W.W. Cooper, Z.M. Huang and D.B. Sun, Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks, Journal of Econometrics 46, 1990, 73–91.

Charnes, A., W.W. Cooper, A. Lewin and L. Seiford (eds.), *Data Envelopment Analysis: Theory, Methods and Applications*, Kluwer Academic, Boston, 1994.

Charnes, A., W.W. Cooper and B. Mellon , Blending aviation gasolines – a study in programming interdependent activities, Econometrica 23, 1954, 307–323.

Charnes, A., W.W. Cooper and E. Rhodes, Measuring efficiency of decision making units, European Journal of Operational Research 1, 1978, 429–449.

Charnes, A., W.W. Cooper, L. Seiford and J. Stutz, A multiplicative model for efficiency analysis, Socio-Economics Planning Sciences 16, 1982, 223–224.

Charnes, A., W.W. Cooper, L. Seiford and J. Stutz, Invariant multiplicative efficiency and piecewise Cobb–Douglas envelopments, Operations Research Letters 2, 1983, 101–103.

Charnes, A., W.W. Cooper and R.M. Thrall, A structure for classifying and characterizing efficiency and inefficiency in Data Envelopment Analysis, Journal of Productivity Analysis 2, 1991, 197–237.

Cooper, W.W., Z. Huang, V. Lelas and S. Li, Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in DEA, Journal of Productivity Analysis, 1996 (submitted).

Cooper, W.W., S. Kumbhakar, R.M. Thrall and X. Yu, DEA and stochastic frontier analyses of the effects of the 1978 Chinese economic reforms, Socio-Economic Planning Sciences 29, 1995, 85–112.

Cooper, W.W. and J.T. Pastor, Global efficiency measures in DEA, Working Paper, Universidad de Alicante, Alicante, Spain, 1995.

Cooper, W.W. and K. Tone, A survey of some recent developments in Data Envelopment Analysis, in *OR: Towards Intelligent Decision Support*, R. Slowinski, ed., Semi-Plenary Papers of the 14th European Conference on Operational Research, Institute of Computing Science, Poznan University of Tech-nology, Poznan, Poland, 1995.

Cooper, W.W., K. Tone, H. Takamori and T. Sueyoshi, DEA: Survey and interpretations, Communications of the Operations Research Society of Japan (August, September, and October issues, 1994). In Japanese; English translations can be secured from the authors.

Debreu, G., The coefficient of resource utilization, Econometrica 19, 1951, 273–292.

Deprins, D., L. Simar and H. Tulkens, Measuring labor efficiency in post offices, in *The Performance of Public Enterprises: Concepts and Measurement*, M. Marchand, P. Pestieau and H. Tulkens, eds., North Holland, Amsterdam, 1984, pp. 243–267.

Dyson, R.G. and E. Thanassoulis, Reducing weight flexibility in Data Envelopment Analysis, Journal of the Operational Research Society 39, 1988, 563–576.

Färe, R. and S. Grosskopf, Measuring congestion in production, Zeitschrift für Nationalökonomie 43, 1983, 257–271.

Färe, R. and S. Grosskopf, Measuring productivity: A comment, International Journal of Operations and Production Management 14, 1994, 83–88.

Färe, R., S. Grosskopf and C.A.K. Lovell, *The Measurement of Productive Efficiency*, Kluwer-Nijhoff, Boston, 1985.

Färe, R. and C.A.K. Lovell, Measuring the technical efficiency of production, Journal of Economic Theory 25, 1978, 150–162.

Farrell, M.J., The measurement of productive efficiency, Journal of the Royal Statistical Society, Series A 120, 1957, 253–290.

Golany, B. and G. Yu, Estimating returns to scale in DEA, Working Paper, Department of Management Science and Information Systems, Graduate School of Business, University of Texas at Austin, Austin, TX, 1994.

Goldfarb, D. and M.J. Todd, Linear programming, in *Handbooks in Operations Research and Management Science*, Vol. 1: *Optimization*, North Holland, Amsterdam, 1989.

Koopmans, T.C., in *Analysis of Production as an Efficient Combination of Activities*, T.C. Koopmans, ed., Wiley, New York, 1951, chapter III.

Koopmans, T.C., *Three Essays on the State of Economic Science*, McGraw–Hill, New York, 1957.

Lovell, C.A.K. and J. Pastor, Units invariant and translation invariant DEA models, Operations Research Letters, 1995.

Lovell, C.A.K., J.T. Pastor and J.A. Turner, Measuring macroeconomic performance in the OECD: A comparison of european and non-european countries, European Journal of Operational Research, 1995 (to appear).

Maindiratta, A.J., Largest size-efficient scale and size efficiencies of decision making units in DEA, Journal of Econometrics 46, 1990, 57–72.

Ray, S., Quantity, quality and efficiency for a partially super-additive cost function: Connecticut schools revisited, Working Paper, University of Connecticut, Storrs, CT, 1995.

Rhodes, E., *Data Envelopment Analysis and Related Approaches for Evaluating the Efficiency of Decision Making Units with an Application to Program Follow Through in U.S. Public School Education*, Ph.D. Thesis, School of Urban & Public Affairs, Carnegie-Mellon University, Pittsburgh. Also available from University Microfilms, Inc., Ann Arbor, MI, 1978.

Roll, Y. and B. Golany, Controlling factor weights in DEA, IIE Transactions 23, 1991, 2–9.

Russell, R., On the axiomatic approach to the measurement of technical efficiency, in *Measurement in Economics*, W. Eichhorn, ed., Physica-Verlag, Heidelberg, 1988.

Seiford, L.M., A DEA bibliography (1978–1992), in *Data Envelopment Analysis: Theory, Methodology and Application*, A. Charnes, W.W. Cooper, A.Y. Lewin and L.M. Seiford, eds., Kluwer Academic, Boston, 1994.

Sengupta, J.T., *Dynamics of Data Envelopment Analysis: Theory of Systems Efficiency*, Kluwer Academic, Boston, 1995.

Sueyoshi, T.. Measuring the industrial performance of Chinese cities by Data Envelopment Analysis, Socio-Economic Planning Sciences 26, 1992, 75–88.

Thanassoulis, E. and R. Allen, Simulating output weight restrictions in Data Envelopment Analysis by means of artifical observations, Management Science, 1994.

Thompson, R.G., P.S. Dharmapala and R.M. Thrall, DEA sensitivity analysis of efficiency measures with spplications to Kansas farming and Illinois coal mining, in *Data Envelopment Analysis: Theory, Methods and Applications*, A. Charnes, W.W. Cooper, A. Lewin and L. Seiford, eds., Kluwer Academic, Boston, 1994.

Thompson, R.G., P.S. Dharmapala and R.M. Thrall, Linked-cone DEA profit ratios and technical efficiency with spplication to Illinois coal mines, International Journal of Production Economics 39, 1995, 99–115.

Thompson, R.G., P.S. Dharmapala and R.M. Thrall, Importance for DEA of zeros in data and multipliers, Journal of Productivity Analysis 4, 1993, 337–348.

Thompson, R.G., L. Langemeier, E.T. Lee and R.M. Thrall, The role of multiplier bounds in efficiency analysis with an application to Kansas farming, Journal of Econometrics 46, 1990, 93–108.

Thompson, R.G., R.D. Singleton, R.M. Thrall and B.A. Smith, Comparative site evaluations for locating a high-energy physics laboratory in Texas, Interfaces 16, 1986, 16–26.

Thompson, R.G. and R.M. Thrall, Polyhedral assurance regions with linked constraints, in *New Directions in Computational Economics*, W.W. Cooper and A. Whinston, eds., Kluwer Academic, Dordrecht, The Netherlands, 1994, pp. 121–133.

Tulkens, H. and P. Vanden Eeckaut, Non parametric efficiency, progress and regress methods for panel data: Methodological aspects, CORE Discussion Paper 9316, Center for Operations Research and Econometrics, Université Catholique de Louvain, Voie du Roman Pay 34, B-1348 Louvain-la-Neuve, Belgium, 1993.

Zhu, J. and Z.H. Shen, A discussion of testing returns to scale in DEA, European Journal of Operational Research 81, 1995, 590–596.