# PREFACE TO TOPICS IN DATA ENVELOPMENT ANALYSIS*

A. CHARNES

*University of Texas System and John P. Hardin Centennial Chair, University of Texas at Austin, Austin, Texas 78712-1170, USA*

and

W.W. COOPER

*Graduate School of Business and IC² Institute, University of Texas at Austin, Austin, Texas 78712-1170, USA*

## Abstract

This paper serves as an introduction to a series of three papers which are directed to different aspects of DEA (Data Envelopment Analysis) as follows: (1) uses and extensions of 'window analyses' to study DEA efficiency measures with an illustrative applications to maintenance activities for U.S. Air Force fighter wings, (2) a comparison of DEA and regression approaches to identifying and estimating sources of inefficiency by means of artificially generated data, and (3) an extension of ordinary (linear programming) sensitivity analyses to deal with special features that require attention in DEA. Background is supplied in this introductory paper with accompanying proofs and explanations to facilitate understanding of what DEA provides in the way of underpinning for the papers that follow. An attempt is made to bring readers abreast of recent progress in DEA research and uses. A synoptic history is presented along with brief references to related work, and problems requiring attention are also indicated and possible research approaches also suggested.

## Keywords and phrases

Efficiency, Pareto optimality, production functions, returns to scale, nondiscretionary inputs, linear programming, fractional programming.

*This research was partly supported by the National Science Foundation and USARI Contract MDA 903-83-K0312, with the Center for Cybernetic Studies, the University of Texas at Austin. It was also partly supported by the IC² Institute at the University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

## 1. Introduction

This paper serves as an introduction to the three that follow. All are devoted to 'Data Envelopment Analysis' (DEA), and directed to its uses in evaluating the efficiency of not-for-profit entities. DEA arrives as these efficiency evaluations without requiring either an a priori choice of weights or explicit specification of functional relations between the multiple outputs and inputs.

The next section of this introductory paper will provide the models and methods by which DEA accomplishes this. More than single scalar evaluations of efficiency are possible with DEA, however, and the papers that follow will also examine some of the additional possibilities. Extensions of currently available methods are also required for use in DEA and some of these topics, too, are developed in the papers that follow.

The initial paper contains a study of the use of DEA in measuring the efficiency of aircraft maintenance operations of wings in the Tactical Air Command (TAC) of the U.S. Air Force. In this example, adapted from [24]*, each wing represents an entity to be evaluated, called a DMU (= Decision Making Unit), although smaller entities such as squadrons or larger ones such as Numbered Air Force units might also have been chosen.

An important technical consideration in choosing the units to serve as DMUs relates to the number of degrees of freedom as determined by the number of DMUs relative to the number of outputs and inputs to be included in a study. A variety of possibilities are available for augmenting the number of DMUs and one of these is provided by the 'window analysis' used in the first paper. In this approach, the number of DMUs is increased by treating each DMU as though it is a different (time labelled) DMU for each period in the window where it appears. As will be seen, this kind of window analysis also provides a number of additional insights that can be gained by studying trends of the behavior of each DMU over the time periods considered, along with stability and other properties of the efficiency measures.

Using actual Air Force data (masked in this publication), the study reported in this first paper☆ was submitted to a review by U.S. Air Force personnel. More than overall (scalar) evaluations of efficiency were required to facilitate these reviews and the ability of DEA to identify sources and estimate amounts of inefficiency in each of the multiple outputs and inputs at each DMU proved to be helpful in effecting the wanted reviews of the DEA efficiency ratings.

*The numbers in square brackets are keyed to the references listed at the end of this series of papers.
☆Full details on this study may be found in [47].

The second paper compares DEA with alternative approaches such as regression estimation of sources and amounts of inefficiency. This is done by reference to a *hypothetical* example involving 15 'hospitals', each with 3 inputs and 3 outputs related by 3 independent linear equations which (uniquely) determine the efficient amounts of inputs needed to support whatever outputs might be specified.* Inefficiencies are introduced into the inputs☆ at some of the hospitals (= DMUs) and regression approaches as well as DEA are examined for their ability to identify and estimate these inefficiencies. Commonly used regressions of cost against outputs are used and the results are compared to ratio analyses and other possibilities as well as DEA.

Although these regressions do surprisingly well in their overall evaluations of efficiency for each DMU, they do not do as well as DEA. The performance of these regressions in identifying the underlying sources and amounts of inefficiencies, on the other hand, is poor and unreliable. DEA generally performs very well in identifying these sources and amounts in all except a very few cases where, it so happens, DEA also signals that something may be wrong with the efficiency characterizations that it has provided.

It might be of help to note that these statistical regressions and DEA use different principles of optimization. Thus, in contrast to the one 'overall' optimization used in arriving at the regression estimates, DEA uses $n$ optimizations – one for each DMU. To state the matter differently, DEA optimizes on *each* observation, whereas the usual statistical regression optimizes across *all* observations. Hence, as might be expected, DEA provides a better fit to each observation and a better basis for identifying and estimating the sources of inefficiency associated with the operations of each of the hospitals (= DMUs) included in this study.

The differences in optimizing principles used in DEA and ordinary statistical (least squares) regression estimates suggest that one might be preferred to the other for uses in certain contexts and problems. For instance, the usual least squares regression might be used when general characterizations are of interest for purposes of policy analysis and prediction of future behavior of the entire ensemble of observations. DEA might then be used when interest centers on individual observations and the institutions (= DMUs) to which they relate. It might also be favored when it is reasonable to suppose that identified inefficiencies can be eliminated, while statistical regression might be used when it is assumed that these inefficiencies cannot be removed and will continue into the future.

*The model was used as a further check on results secured in a study of Massachusetts hospitals where, as reported in [76], all results were reviewed by a committee of experienced hospital administrators, physicians, surgeons and state regulators.

☆A statistically designed study based on underlying translog and piecewise Cobb–Douglas technologies with inefficiencies only in the outputs is reported in [10].

Other criteria for choosing between regression and DEA are also possible and the two approaches may also be used together as when DEA is used to adjust or refine the data prior to forming regression estimates. This kind of usage is not confined to least squares regression estimation. Schinnar in [72], for example, used DEA to arrive at estimates of 'efficient' input-output coefficients for use in place of the 'averages' that are usual in Leontieff-type 'inter-industry analyses'. Evidently, additional alternatives are also possible, up to and including the use of DEA to obtain entire efficiency frontiers and to supply 'dual variable' values by means of which tradeoffs and adjustments may be affected along these frontiers*.

As will be seen in the next section, DEA can be given a linear programming formulation with accompanying powers of computation and interpretation. Computer codes for ordinary linear programming (simplex) computations can thus be used, but more efficient codes are available for dealing with the $n$ optimizations involved in DEA. One such code developed by I. Ali, D. Divine and J. Stutz is available from the Center for Cybernetic Studies at the University of Texas at Austin. As seen in table 7 of the first paper, this code provides printouts with supplemental information for interpreting the results for each DMU.

Theoretical and methodological extensions are also required for particular uses of DEA. One such extension revolves around the topic of sensitivity analysis. As noted in the third paper, variations in the data of the DMU being evaluated are associated with variations in one of the basis vectors. Hence it is not possible to use ordinary sensitivity analyses which proceed on the assumption that the data variations being studied do not affect elements of the basis inverse.

This extension of sensitivity analysis can have other uses as well. For instance, as reported in [16], it was desired to use DEA as a guide for effecting changes in the individual program units (= DMUs) in San Antonio's (20 000 student) community college. Such changes could have effects on the efficiency evaluations of *other* DMUs. For purposes like these it is also desirable to have efficient means for sensitivity analyses along with other aspects of the procedures described in the 3rd paper.

All three papers provide numerical examples which are intended to help explain and illustrate what is occurring. They can also be helpful in other ways. For instance, we have used the data in paper number two for checking proposed alternatives and extensions to DEA. Others may find it similarly useful☆ and undoubtedly other uses of the data in this and the other two papers are also possible.

*See [36] for a detailed mathematical development.

☆See the discussion in Bowlin [19].

## 2. Models and notation*

We now formally introduce DEA via the following model and its associated extremal principles which, as noted in [41], extend the usual single output to single input efficiency definitions employed in the natural sciences:

$$\text{Maximize } h_{j_0} = \frac{\sum_{r=1}^{s} u_r y_{rj_0}}{\sum_{i=1}^{m} v_i x_{ij_0}} .$$

Subject to:

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leqslant 1; \quad j = 1, 2, \ldots, j_0, \ldots, n$$

$$u_r > 0; \qquad r = 1, \ldots, s ,$$

$$v_i > 0; \qquad i = 1, \ldots, m , \tag{1}$$

where $x_{ij}$ = the observed amount of input of the $i$th type for the $j$th DMU ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$) and $y_{rj}$ = the observed amount of output of the $r$th type for the $j$th DMU ($r = 1, 2, \ldots, s$; $j = 1, 2, \ldots, n$).

The $j_0$th unit being evaluated in the objective is also part of the constraint set so that $0 \leqslant h_{j_0} \leqslant 1$ with existence of solutions assured for any (or all) $j_0$ for which an evaluation is sought. The $u_r$ and $v_i$ values are determined directly from the data via the above model with its associated extremal principle as each $j = j_0$ designates a $\text{DMU}_{j_0}$ to be inserted in the functional. Evidently, $\max h_{j_0} = h^*_{j_0} = 1$ is required for

*The developments in this section are drawn from the original version of [38], which was edited to form the third paper in the sequence that follows.

efficiency so $h_{j_0}^{\star} < 1$ means efficiency has not been achieved. The value of $h_{j_0}^{\star}$ is independent of the units in which the observed inputs and outputs are stated. This is formalized and proved as follows:*

THEOREM 1 (units invariance):

The optimal $h_{j_0} = h_{j_0}^{\star}$ is independent of the units in which the observed inputs and outputs are measured so long as the units are the same for every DMU.

*Proof:*☆

Let an optimal solution to (1) be given by $h_{j_0}^{\star}$, $u_r^{\star}$, $v_i^{\star}$. Then replace the original $y_{rj}$, $x_{ij}$ by $\rho_r y_{rj}$, $\delta_i x_{ij} > 0$ with $\rho_r y_{rj} = y_{rj}$ and $\delta_i x_{ij} = x_{ij}$ for those $r$ and $i$ where no change is made. Choosing $u_r' = u_r^{\star}/\rho_r$ and $v_i' = v_i^{\star}/\delta_i$, we have $h_{j_0}'^{\star} \geqslant h_{j_0}^{\star}$. Suppose we could have $h_{j_0}'^{\star} > h_{j_0}^{\star}$. However, $u_r = u_r'^{\star}\rho_r$ and $v_i = v_i'^{\star}\delta_i$ satisfy the original constraints and this contradicts the assumed optimality of $h_{j_0}^{\star}$, $u_r^{\star}$, $v_i^{\star}$. Thus the assumption $h_{j_0}'^{\star} > h_{j_0}^{\star}$ leads to a contradiction. The only remaining possibility is $h_{j_0}'^{\star} = h_{j_0}^{\star}$, as asserted in the theorem. Q.E.D.

The formulation in (1) is a nonlinear-nonconvex problem involving a linear fractional functional with linear fractional constraints. As shown in [41], it can be replaced by what looks like an ordinary linear programming problem (ignoring the *required* positivity of the variables) and associated with common methods of solution (e.g. the simplex method) which provide optimal bases of 'efficient DMUs' for effecting the evaluation of the $j_0$th DMU. The number of times the model must be solved can be reduced, since each such solution locates a subset of efficient DMUs as a by-product of the computations. It also provides a basis for assessing the sources of inefficiency (if any) in the DMU being evaluated along with optimal values of the associated dual variables that can be used to determine tradeoff possibilities along the efficiency frontier.

We have indicated how (1) generalizes the concept of efficiency (single output to single input ratio form) that is used in the natural sciences. By moving to the corresponding linear programming form, we can relate it to the concept of 'Pareto efficiency' or 'Pareto-Koopmans efficiency'† by taking account of any non-zero slack that may be present in the optimal solutions.

Using vector notation, we therefore now replace (1) with a reciprocal 'inefficiency' ratio form, namely:

*Other invariance properties are given in [69].
☆An alternative proof is given in [30].
†See chapter IX of [28].

$$\text{Minimize } h_0 = \frac{v^T X_0}{u^T Y_0}$$

subject to

$$\frac{v^T X_j}{u^T Y_j} \geqslant 1, \quad j = 1, 2, \ldots, n$$

$$(u^T Y_0)^{-1} v^T \geqslant \epsilon \cdot e^T > 0,$$

$$(u^T Y_0)^{-1} u^T \geqslant \epsilon \cdot e^T > 0. \tag{2}$$

Since we shall have

$$u^T Y_0 > 0,$$

we can achieve the new linear programming form via (2) ff. in what follows. Thus we can proceed to relate this to (1) by observing that $Y_0$ and $X_0$ contain as components the $y_{rj_0}$ and $x_{rj_0}$ that appear in the functional of (1). Similarily, the $X_j$ and $Y_j$ vectors contain the same $x_{ij}$ and $y_{rj}$ data as in (1) for each $j = 1, \ldots, n$. The components of the row vectors $u^T$ and $v^T$ are to be determined via the indicated optimization to obtain the $u^\star$ and $v^\star$ values which are wanted. Finally, $e^T$ is the transpose of the column vector $e$ which has all of its elements equal to unity*.

The symbol $\epsilon$ represents the infinitesimal we use to generate the non-Archimedean ordered extension field we shall use. In this extension field, $\epsilon$ is less than every positive number in our base field but greater than zero. Its usage guarantees that optimal solutions to the new (extended field) linear programming problems are at finite non-zero extremal points.

To transform the ratio problem of (2) to a linear programming form, we make the change of variables:

$$t = \frac{1}{u^T Y_0}$$

$$\omega^T = t v^T, \quad \mu^T = t u^T. \tag{3}$$

*To avoid still further notation, we shall use the same symbol $e$ for vectors of different length.

Then multiplying numerators and denominators in (2) by $t$ and adding the consistency condition, $tu^{\mathrm{T}}y_0 = 1$, problem (2) becomes

$$\min \omega^{\mathrm{T}} X_0$$

subject to

$$\begin{aligned} \omega^{\mathrm{T}} X_j - \mu^{\mathrm{T}} Y_j &\geqslant 0, \qquad j = 1, 2, \ldots, n \\ \mu^{\mathrm{T}} Y_0 &= 1, \\ \omega^{\mathrm{T}} &\geqslant \epsilon e^{\mathrm{T}}, \\ \mu^{\mathrm{T}} &\geqslant \epsilon e^{\mathrm{T}}. \end{aligned} \tag{4}$$

The linear programming dual to (4) is:

$$\max (z_0 + \epsilon e^{\mathrm{T}} s^- + \epsilon e^{\mathrm{T}} s^+)$$

subject to

$$\begin{aligned} \sum_j X_j \lambda_j + s^- &= X_0 \\ Y_0 z_0 - \sum_j Y_j \lambda_j + s^+ &= 0 \\ \lambda_j, s^+, s^- &\geqslant 0, \end{aligned} \tag{5}$$

and $z_0$ unconstrained.

We may note that (5) can be interpreted as a problem in which one maximizes the 'intensity' $z_0$ of the output vector $Y_0$ subject to envelopement from above and below. The envelopment from above is by reference to the outputs and the envelopment from below is by reference to the inputs of $\mathrm{DMU}_0$, as can be seen by rewriting the expressions for (5) in equivalent inequality form. The envelopment is tightened to the maximal extent possible via $\max z_0 = z_0^\star$, with, as we shall see, $z_0^\star \geqslant 1$ and $z_0^\star = 1$ occurring only when $\mathrm{DMU}_0$ is efficient.

Before proceeding to a proof of this last statement, we might observe that the name 'Data Envelopment Analysis' derives from what has just been described, but is intended to cover this kind of 'envelopment process' in other situations where such inequality representations and extremal principles are employed for analogous purposes.

We now prove what we shall refer to as the Non-Archimedean Efficiency Theorem:*

THEOREM 2:

$DMU_0$ is efficient in (4) if and only if

$$z_0^\star = 1 \quad \text{and} \quad s^{+\star} = s^{-\star} = 0,$$

i.e. the intensity is unity and all slacks equal to zero, where an optimal solution to (5) is denoted by the vector $(\lambda^\star, z_0^\star, s^{-\star}, s^{+\star})$.

*Proof:*

$DMU_0$ is efficient if and only if $\omega^{\star T} X_0 = 1$ in (4). By equality of the dual functionals at an optimum, $1 = \omega^{\star T} X_0 = z_0^\star + \epsilon e^T s^{-\star} + \epsilon e^T s^{+\star}$.

Since the constraint system of (5) contains no non-Archimedean quantities, it follows that no basic solution can contain non-Archimedean quantities. But

$$1 = z_0^\star + \epsilon e^T (s^{+\star} + s^{-\star})$$

if and only if the coefficients of $\epsilon$ are all zero, since $z_0^\star$, $s^{+\star}$, and $s^{-\star}$ do not contain $\epsilon$. Q.E.D.

The introduction of a non-Archimedean infinitesimal has thus provided access to a simple but rigorous linear programming model having the required positivity of the $u^T$ and $v^T$. The theorem also shows why, in addition to $z_0^\star = 1$, the slacks $s^{+\star}$ and $s^{-\star}$ must equal zero. Notice further that it is precisely the tagging of the slack in (5) with this infinitesimal which allows us to recover the effects of any positivity requirements on the $u^T$ and $v^T$ in (2). Concomitantly, any $u^T$ or $v^T$ component in (2) which has this non-Archimedean infinitesimal involved in its value will be uniquely identified with a slack which is positive in (5). Thus each slack value in (5) is unambiguously

*See [38] for further discussions and interpretations.

related to the appearance of a non-Archimedean infinitesimal in (2). The slack which is apparent in (2), but not in (5), requires nothing further in the way of treatment or interpretation since it very naturally represents the complement of the corresponding efficiency or inefficiency value, i.e. the deviation from unity for this constraint, in each particular evaluation.

As already noted, the value of $h_0^*$ and hence the values of $z_0^*$ and $\omega^{*\mathrm{T}} X_0$ do not depend on the units in which the inputs and outputs are measured. The slacks do not have this same property and it is important to arrange computations so that their values will not affect the $z_0^*$. This can be done by inserting an extra row in the simplex tableaus to accommodate the non-Archimedean infinitesimals, as described on page 176 ff. in [28]. Alternatively, one may proceed in a two-phase manner by optimizing on $z_0$, in (5) say, and then maximizing the slack with $z_0$ fixed at the $z_0^*$ value achieved in the preceding phase. Other possibilities also exist, of course, but in any case efficiency requires that the slacks $s^{+*}$ and $s^{-*}$ must all be zero and this condition does not depend on the units of measure used.

Drawing this all together, we have

$$h_0^* = \omega^{*\mathrm{T}} X_0 = z_0^* + \epsilon e^{\mathrm{T}} s^{-*} + \epsilon e^{\mathrm{T}} s^{+*}, \tag{6}$$

with Theorems 1 and 2 holding in all of these cases so that, in particular, if $h_0^* = 1$, then non-Archimedean elements do not appear in the optimum value of (2). Any or all of these formulations may be used for computation or interpretation. Additional formulations are also possible.

In (5), normalizing is applied to the outputs. We can supply normalizing to the inputs instead, as is done in the problem on the left in the following dual pair:

$$\begin{array}{l|l}
\max \mu^{\mathrm{T}} Y_0 & \min \theta - \epsilon e^{\mathrm{T}} s^{+} - \epsilon e^{\mathrm{T}} s^{-} \\
\text{subject to} & \text{subject to} \\
\nu^{\mathrm{T}} X_0 = 1 & Y\lambda - s^{+} = Y_0 \\
\mu^{\mathrm{T}} Y - \nu^{\mathrm{T}} X \leqslant 0 & \theta X_0 - X\lambda - s^{-} = 0 \\
-\mu^{\mathrm{T}} \leqslant -\epsilon e^{\mathrm{T}} & \lambda, s^{+}, s^{-} \geqslant 0 \\
-\nu^{\mathrm{T}} \leqslant -\epsilon e^{\mathrm{T}} &
\end{array} \tag{7}$$

Formally, the model on the right is obtained from (5) by dividing all constraints by $z_0$ and reorienting the objective as indicated for the intensity (or scale) variable $\theta = 1/z_0$.* The symbols $X$ and $Y$ are the matrices $X = [X_1, \ldots, X_n]$, $Y = [Y_1, \ldots, Y_n]$. The problem on the right is in 'envelopment form'. The problem on the left is in (normalized) 'production' or 'efficiency technology' form where the objective is to maximize $DMU_0$'s rate of output,☆ given the unit rate of input normalization in the first constraint, subject to non-Archimedean positivity and the 'efficiency technology' condition that outputs cannot exceed inputs in any of the other constraints.

We can do more than simply test for efficiency. We can also adjust the inefficient operations by means of formulae that we can develop as follows.

Let $\theta^*$, $\lambda_B^*$, $s_B^{+*}$, $s_B^{-*}$ designate an optimal *basic* solution with an associated collection of coefficient vectors and matrices from the problem on the right of (7):

$$\begin{bmatrix} 0 \\ X_0 \end{bmatrix}, \quad \begin{bmatrix} Y_B \\ -X_B \end{bmatrix}, \quad \begin{bmatrix} -I_B^+ \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ -I_B^- \end{bmatrix}. \tag{8}$$

From this, as the basis of this optimal solution, we obtain the reduced system of equations:

$$\begin{aligned} Y_B\lambda_B^* - s_B^{+*} &= Y_0 \\ \theta^* X_0 - X_B\lambda_B^* - s_B^{-*} &= 0. \end{aligned} \tag{9}$$

If one replaces $(X_0, Y_0)$ by $X_0' = \theta^* X_0 - s_B^{-*}$ and $Y_0' = Y_0 + s_B^{+*}$, then $\theta = 1$, $\lambda = \lambda_B^*$, $s_B^{-*} = 0$, $s_B^{+*} = 0$ is a feasible basic solution for $(X_0', Y_0')$ with this same DMU column basis.

Now $\mu_B^{*T}$, $\nu_B^{*T}$ are the dual evaluators for the old basis. Applying these to (9) we obtain

$$\begin{aligned} \mu_B^{*T} Y_B \lambda_B^* &= \mu_B^{*T}(Y_0 + s_B^{+*}) = \mu_B^{*T} Y_0', \\ \nu_B^{*T} X_B \lambda_B^* &= \nu_B^{*T}(\theta^* X_B - s_B^{-*}) = \nu_B^{*T} X_0', \end{aligned} \tag{10}$$

*The variables in (5) can be related to those in (6) by writing the latter as $\hat{\lambda} = \lambda/z_0$ and $\hat{s}^+ = s^+/z_0$, $\hat{s}^- = s^-/z_0$, but we do not do this in order to avoid additional notation.

☆The reference is to 'virtual' outputs and inputs as defined in the next section.

But $\mu_B^{\star T} Y_B - \nu_B^{\star T} X_B = 0$ for the dual inequalities designated by $\lambda_B^\star$. Hence

$$\mu_B^{\star T} Y_0' = \mu_B^{\star T} Y_B \lambda_B^\star = \nu_B^{\star T} X_B \lambda_B^\star = \nu^{\star T} X_0', \tag{11}$$

i.e. the new inequality replacing $\mu^T Y_0 - \nu^T X_0 \leqslant 0$ is also satisfied by $\mu_B^{\star T}$, $\nu_B^{\star T}$. Further

$$(\mu_B^{\star T} Y_0')/(\nu^{\star T} X_0') = 1 . \tag{12}$$

Thus $\tilde{\mu}_B^T = \mu_B^{\star T}/(\nu^{\star T} X_0'), \tilde{\nu}_B^T = \nu_B^{\star T}/(\nu^{\star T} X_0')$ is a basic feasible solution to the $(X_0', Y_0')$ problem with functional value $\tilde{\mu}_B^T Y_0' = 1$ equal to the dual problem functional value $\theta = 1$. Thus the 'projection', which we refer to as the CCR projection,*

$$\begin{aligned} X_0 &\to X_0' = \theta^\star X_0 - s^{-\star} , \\ Y_0 &\to Y_0' = Y_0 + s^{+\star} \end{aligned} \tag{13}$$

is efficient. That is, the $X_0'$, $Y_0'$ obtained in this manner from the original $X_0$, $Y_0$ is efficient and the differences

$$\begin{aligned} \Delta X_0 &= X_0 - X_0' , \\ \Delta Y_0 &= Y_0' - Y_0 , \end{aligned} \tag{14}$$

represent the estimated *amounts* of input and output inefficiencies, respectively, in the $X_0$, $Y_0$ observed for $DMU_0$.

*As published in [41] which was also the first work to associate an efficient ( 'Pareto efficient') input-output vector with *each* given $(X_j, Y_j)$ vector. See also the alternate development provided there.

## 3. Application and interpretations

We now draw from the discussion of audit and evaluations of managerial performance on pp. 45–46 in [50] in order to distinguish activities in each of the following categories:

| | |
|---|---|
| (1) Propriety of | (a) Objectives pursued<br>(b) Methods used |
| (2) Effectiveness in | (a) Stating objectives<br>(b) Attaining objectives |
| (3) Efficiency of | (a) Benefits achieved<br>(b) Resources utilized |

Some managerial performance measures may comprehend more than one category. Profit, for example, may reflect the improved effectiveness achieved by a business firm in changing from producing steel to producing oil. It may also reflect the efficiency with which oil is produced. Hence, the total profit may include both efficiency and effectiveness quite apart from whether the change from steel to oil was a proper undertaking.

Distinctions between effectiveness and efficiency need not be emphasized in evaluating private enterprise activities. They are of importance in gauging the activities of public enterprises where a change from one type of activity to another (i.e. a change in the direction of its activities) often requires specific legislation or voter approval.

Our concern here, and in the papers that follow, is with efficiency. We lay aside the more difficult problem of effectiveness and assume that this has been decided in the choice of inputs (resources) to be used and outputs (benefits) to be achieved, as well as the ways in which the inputs and outputs are to be measured. Theorem 1 on 'units invariance' provides a certain amount of latitude and other devices may also be used. Generally speaking, however, 'augmentation' is the desired direction for outputs and 'diminution' is the desired direction for inputs.

Flexible uses of these definitions are possible. For instance, the reciprocal of an input amount may be formalized as an output and placed in the numerator rather than the denominator (along with other inputs) in (1). Flexibility is also allowed in the choice of DMUs. These choices are of basic importance, however, and so a certain amount of checking is always advisable.

We may now formalize our definition of efficiency as follows:

100% efficiency is attained for any DMU only when

(a) None of its outputs can be increased without either

  (i) increasing one or more of its inputs or
  (ii) decreasing some of its other outputs.

(b) None of its inputs can be decreased without either

  (i) decreasing some of its outputs or
  (ii) increasing some of its other inputs.

Thus efficiency is represented by the attainment of Pareto optimality* and conversely. Output or input inefficiency corrections are allowed under this definition without worsening any other input or output and the need for assigning measures of relative importance to the different inputs and outputs is thereby avoided.

The above definition is formulated so that efficiency may be determined relative to prior theoretical knowledge. That is, such knowledge may be available by reference to available theory as in parts of the natural sciences. It can also be arranged artificially, by design, for testing DEA and other approaches to efficiency measurement, as is done by reference to the underlying models in the second of the three papers that follow. Such knowledge of true or theoretical efficiency is not available for other situations, however, as in the Air Force applications reported in the first paper. For such uses, we need to extend the above definition to one which involves only *relative* efficiency as determined from the kind of data that are likely to be available.

> 100% *relative* efficiency is attained by any DMU only when comparisons with other relevant DMUs do not provide evidence of inefficiency in the use of any input or output.

Via this characterization, the preceding definition is adjusted for immediate application to data of the kind we shall be considering. We should also note, however, that other combinations of the above definitions are also possible so that, in addition, pertinent aspects of any theoretically grounded norms or other types of available knowledge may also be used in common with other data when required.

*Also called Pareto–Koopmans optimality in chapter IX of [28].

All of the observed outputs $y_{rj}$, and all of the observed inputs $x_{ij}$ are assumed to be available as known positive constants.* That is, each of the $j = 1, 2, \ldots, n$ DMUs are assumed to have used positive amounts of each pertinent input and to have produced positive amounts of each pertinent output.

Some of the inputs may be varied at the discretion of a manager and some may not. An example of a 'non-discretionary' input is provided by the weather, which can affect the performance of Air Force wings. We do not deal with this topic in the papers that follow, but it is evident that 'better' and 'worse' weather at different bases should be taken into account along with other inputs and outputs that are pertinent to performance efficiency.☆ Thus, both discretionary and non-discretionary inputs were used in the DEA study of Air Force wings, along with a mix of various types of wings (training and operational) which also used different types of aircraft with relatively satisfactory results in all cases.

The optimal $u_r^*$ and $v_i^*$ as determined from (1) and (2) ff. have a variety of uses in their own right as when, for instance, they are employed to determine further tradeoff possibilities after efficiency has been attained. They are also called 'virtual rates of transformation'. More generally, the $u_r$, $v_i$ define a 'virtual output'

$$y_0 = \sum_{r=1}^{s} u_r y_{r0} , \tag{15}$$

and a 'virtual input'

$$x_0 = \sum_{i=1}^{m} v_i x_{i0} , \tag{16}$$

so that also

$$h_0 = y_0/x_0 , \tag{17}$$

*Methods for relaxing this positivity requirement for the observed inputs and outputs in every DMU are described in [46].

☆Banker and Morey [14] provide detailed models and methods for treating non-discretionary inputs and report a test in which the economics of scale effects were reversed by their treatment. See also Bowlin [19] for an application of DEA to Air Force base maintainance activities in which the use of the Banker–Morey models had no substantial effects.

with

$$h_0^\star = y_0^\star / x_0^\star \tag{18}$$

when an optimum is achieved. In other words, our definitions as given above were motivated by the classical ratio definitions of efficiency in engineering, physics (and other fields), while accommodating multiple output and multiple input situations. The definitions are arranged so that they also make contact with the definitions of Pareto efficiency in economics and, as illustrated in the papers that follow, we can move from an overall (scalar) evaluation of efficiency and track the sources and amounts of inefficiency into the underlying components. Finally, if wanted, we can also construct the entire efficiency surface and arrange to determine the tradeoff possibilities that are associated with movement along these efficiency surfaces. See [36] for a rigorous development by means of which the production function (efficiency) surfaces may be constructed (and validated) empirically.

## 4. Refinements and extensions

To relate the foregoing definitions more specifically to DEA we refer to the problem on the right in (7) and say that a DMU = $DMU_0$ is 'DEA efficient' if and only if both of the following are satisfied:

$$\text{(i)} \quad \min \theta = \theta^\star = 1, \quad \text{and}$$

$$\text{(ii)} \quad s^{+\star} = s^{-\star} = 0 \text{ in all alternative optima.} \tag{19}$$

In addition to the positivity assured for the solution to the problem on the left in (7) by $\epsilon > 0$, we can now bring into view the role these constants play for the problem on the right. For any choice of $\theta$, the problem on the right maximizes the sum of the slacks. Because of the non-Archimedean character of $\epsilon > 0$, achievement of $\min \theta = \theta^\star$ with all slacks equal to zero ensures that the slacks must also be zero in all alternative optima.

Figure 1 provides an illustration in which 4 DMUs are represented by points $P_1, P_2, P_3, P_4$ with inputs given by their coordinates $(x_1, x_2)^T$. All have produced one unit of output so inefficiencies, if any, are in the input amounts utilized.

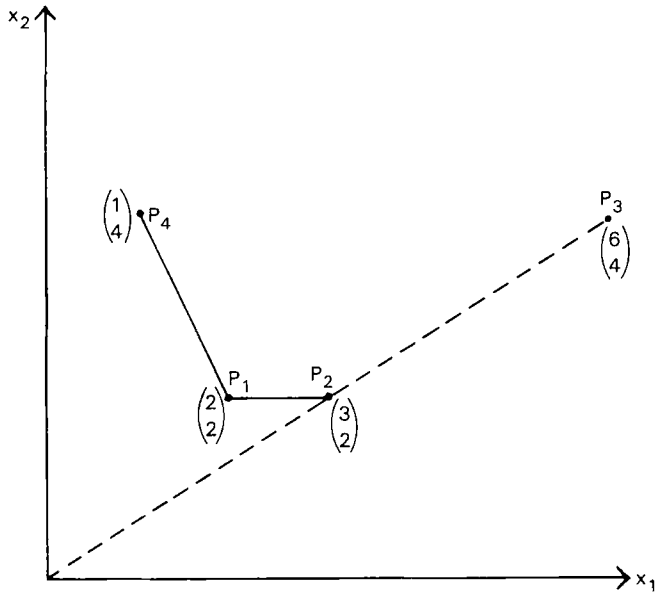


Fig. 1. DEA efficiencies.

To evaluate $P_3$, for example, we insert its coordinates in (7) to obtain

$$\min \theta - \epsilon s_1^- - \epsilon s_2^-$$

subject to

$$6\theta = 2\lambda_1 + 3\lambda_2 + 1\lambda_3 + 6\lambda_4 + s_1^-$$

$$4\theta = 2\lambda_1 + 2\lambda_2 + 4\lambda_3 + 4\lambda_4 + s_2^-$$

$$1 = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$$

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4, s_1^-, s_2^- \geqslant 0 .$$

The coefficients in the first 2 constraints are obtained from the coordinates shown for the similarly indexed point in fig. 1. The third constraint represents the unit output that resulted from the inputs utilized by each DMU. For the single output case, it is easy to show that $s^{+*} = 0$, which means that no output inefficiency is present in the

single output case.* Hence, we have simply omitted the slack variable from the output constraint.

The value min $\theta = \theta^* = 1/2$ in the above problem is compatible with $\lambda_1 = 1$ and it is also compatible with $\lambda_2 = 1$ and all other $\lambda_j = 0$. The choice $\lambda_1 = 1$ is associated with $s_1^- = 1$, while $\lambda_2 = 1$ has $s_1^- = s_2^- = 0$. Hence, $\lambda_1^* = 1$, $s^{-*} = 1$ and all other variables at zero value are optimal and $P_3$ fails to satisfy either of the conditions for DEA efficiency in (19).

To evaluate $P_2$ we simply replace the $\theta$ coefficients in the above problem with the $P_2$ coordinates from fig. 1, i.e. we replace $6\theta$ with $3\theta$ and $4\theta$ with $2\theta$. The other parts of the problem are not altered.

Carrying out the minimization for the thus altered problem, we obtain $\theta^* = 1$ and this, too, is compatible with $\lambda_1 = 1$ and $\lambda_2 = 1$. The second of these two choices has both slacks at zero, while the first has $s_1^- = 1$. Hence the optimal choice is $\theta^* = 1$, $\lambda_1^* = 1$, $s_1^{-*} = 1$, and all other variables equal to zero. Thus (i) is satisfied for (18) but not (ii), and DEA efficiency is not attained by $P_2$.

Continuing with similar $\theta$-coefficient replacements for $P_1$ and $P_4$, it is found that both are DEA efficient since the twin conditions in (19) are both satisfied for the DMUs associated with these points. To interpret these results, we may return to fig. 1 and imagine that the solid line connecting $P_4$ and $P_1$ is part of a level line obtained by passing a plane through the three-dimensional production surface at unit output level and projecting the results down into the two dimensions portrayed in fig. 1. The line connecting $P_1$ and $P_4$ is then distinguished as the 'isoquant', i.e. the portion of the level line containing the combinations of $(x_1, x_2)$ that will produce one unit of output with DEA efficiency.

The point $P_3$ as shown in fig. 1 is not efficient because it is not on the level line.☆ That is, $P_3$ is not a 'frontier point', which means that this point does not lie on the 3-dimensional production surface.

The property of being a frontier point is necessary but not sufficient for DEA efficiency, as witness the situation for $P_2$. Even though $P_2$ is a frontier point, it is possible to go from $P_2$ to $P_1$ in a way that reduces one input (in the amount of $s_1^{-*} = 1$) without increasing the other input and without decreasing the output. Hence, $P_2$ is not DEA efficient.

*This is not true for extensions to the case of multiple outputs. See the discussion of K. Laitinen in the next section.

☆An alternative interpretation would make the broken line connecting $P_2$ and $P_3$ part of the level line which is on the 'wrong side' of the production surface and thus exhibit what Byrnes, Färe and Grosskopf [22] refer to as 'congestion'. As noted in [8], we prefer to describe these as 'mix inefficiencies' and reserve the term 'congestion' for use in situations where input reductions are associated with output *increases*. In any case, the segment from $P_1$ to $P_2$ is not part of an isoquant if that term connotes 'efficient production', as in the usual usages in economics.

The situation for the isoquant connecting $P_1$ and $P_4$ represents Pareto (or DEA) efficiency. It is not possible to go from one point to another along this line segment in order to decrement one input without also decrementing the output or incrementing the other input. Holding output constant at one unit, the rate at which $x_1$ and $x_2$ must be (optimally) exchanged for each other is indicated by the isoquant connecting $P_1$ and $P_4$.* No such *tradeoff* is required in going from $P_2$ to $P_1$.

We now write

$$\theta^* x_{i0} = \bar{x}^*_{i0} + s_i^{-*}; \qquad i = 1, \ldots, m, \tag{20}$$

where $\bar{x}^*_{i0}$ is an optimal convex combination of the $x_{ij}$, i.e.

$$\bar{x}^*_{i0} = \sum_{j=1}^{n} x_{ij} \lambda_j^*, \tag{21}$$

with

$$\sum_{j=1}^{n} \lambda_j^* = 1, \lambda_j^* \geqslant 0 \text{ all } j.$$

A rearrangement of (20) then gives

$$\theta^* = \frac{\bar{x}^*_{i0} + s_i^{-*}}{x_{i0}}, \tag{22}$$

and $X_0$, the vector with components $x_{i0}$, will be a frontier point if and only if $\theta^* = 1$.

We illustrate with our previous solution for $P_3$ in fig. 1 where we have $\theta^* = 1/2$, $\lambda_1^* = 1, s_1^{-*} = 1, s_2^{-*} = 0$. Thus

$$1/2 = \frac{x_{11} + s_1^{-*}}{x_{13}} = \frac{2 + 1}{6},$$

where $x_{13}$ and $x_{23}$ are the first and second components in $X_0$ for $P_3$. Evidently, $P_3$ is not on the frontier. The value $\theta^* = 1/2$, as is readily verified, represents the ratio of

*These tradeoffs may be developed from the ratios of the dual variables as in [36].

the Euclidean distances from the origin to $P_2$ and $P_3$, respectively. It brings $P_3$ into $P_2$ with the latter expressed as the sum of $P_1$ plus the slack vectors.

Proceeding next to $P_2$ we again utilize our previous solution $\theta^* = 1$, $\lambda_1^* = 1$, $s_1^* = 1, s_2^* = 0$ and obtain

$$1 = \frac{x_{11} + s_1^{-*}}{x_{12}} = \frac{2 + 1}{3}.$$

Although $P_2$ is on the frontier, it is not efficient because of the presence of $s_1^{-*} = 1$. Stated differently, the condition $\theta^* = 1$ is necessary but not sufficient for DEA efficiency. Output and input slacks must also all be zero, as noted in (19). Finally, we also note that $P_2$ cannot be part of an optimal basis because it is not efficient. Hence, it is $P_1$ (which is efficient) rather than $P_2$ which is used in the optimal basis from which $P_3$ is evaluated. Stated more generally, the fact that a structural vector,* like $P_1$ or $P_4$, enters into an optimal basis suffices to identify it as efficient although, as we shall see in the first of the following three papers, further analyses may be required when it achieves this status by performing *only* as a 'self-evaluator' – i.e. appearing only in its own optimal basis, and no other – which can occur because of its being located in a part of the space where it cannot be represented as a non-negative combination of *other* $P_j$*s* (cf. the situation for $P_4$ in fig. 1).

To refine these developments in a manner that extends the theory and interpretive power of DEA, we draw on Banker, Charnes and Cooper [9] and replace (7) by

$$\min \theta - \epsilon \left[ \sum_{r=1}^{s} \hat{s}_r^+ + \sum_{i=1}^{m} \hat{s}_i^- \right]$$

subject to

$$\sum_{j=1}^{n} y_{rj} \hat{\lambda}_j - \hat{s}_r^+ = y_{r0}$$

$$\theta x_{i0} - \sum_{j=1}^{n} x_{ij} \hat{\lambda}_j - \hat{s}_i^- = 0$$

*We are using the terminology of [28].

$$\sum_{j=1}^{n} \hat{\lambda}_j = 1$$

$$\hat{\lambda}_j, \hat{s}_i^-, \hat{s}_r^+ \geqslant 0\,, \tag{23}$$

where $j = 1, \ldots, n$; $i = 1, \ldots, m$; $r = 1, \ldots, s$.

The adjunction of the last constraint for the $\hat{\lambda}_j$ introduces a new variable in the dual to (23) which we represent by $\mu_0$ in:

$$\max \sum_{r=1}^{s} \hat{\mu}_r y_{r0} - \mu_0$$

subject to

$$\sum_{i=1}^{m} \hat{\nu}_i x_{i0} = 1$$

$$-\sum_{i=1}^{m} \hat{\nu}_i x_{ij} + \sum_{r=1}^{s} \hat{\mu}_r y_{rj} - \mu_0 \leqslant 0$$

$$\hat{\mu}_r, \hat{\nu}_i \geqslant \epsilon > 0 \;\forall\; r, i\,. \tag{24}$$

The new variable $\mu_0$ is not constrained in sign. Hence it may assume optimal values

$$\mu_0^\star \lesseqgtr 0\,. \tag{25}$$

As shown in [9], the value $\mu_0^\star = 0$ may be indentified with the property that returns to scale are locally constant for $DMU_0$, while $\mu_0^\star < 0$ and $\mu_0^\star > 0$ if returns to scale are locally increasing or decreasing, respectively.

To relate these properties to the DEA efficiency conditions of (19) we utilize (13) and bring the observed inputs and outputs for any $DMU_0$ onto the efficiency frontier. We will then have

$$\frac{\sum_{r=1}^{s} \mu_r^* \hat{y}_{r0} - \mu_0^*}{\sum_{i=1}^{m} \mu_i^* \hat{x}_{i0}} = 1,$$

where $\hat{y}_{r0}$, $\hat{x}_{i0}$ represent the thus adjusted values of the observed outputs and inputs. We can also write this as

$$\sum_{r=1}^{s} \mu_r^* \hat{y}_{r0} - \sum_{i=1}^{m} v_i^* x_{i0} - \mu_0^* = 0,$$

which is the equation of a hyperplane with 'intercept' $\mu_0^*$. It is in fact a supporting hyperplane at the point $(\hat{y}_{10}, \ldots, \hat{y}_{s0}, \hat{x}_{10}, \ldots, \hat{x}_{m0})$.

All points in the same facet as the one used to evaluate (and adjust) $DMU_0$ will have the same or alternate optimum bases. They will therefore have the same optimal dual variable values. Hence they will also satisfy (24), which means that the similarly adjusted output and input values for these DMUs will also lie in this hyperplane.*

The portrayal in fig. 2(a) for the one output/one input case makes it possible to visualize what is occurring. Point $A$ is inside the production possibility set with the frontier defined by the efficiency surface represented by the solid line. Obtaining a solution to (23) for the $DMU_0$ associated with $A$ and applying (13) projects $A$ into $B$ on the efficiency surface.

Any point on this efficiency surface with input value $x$ and output value $y$ will satisfy

$$b = \frac{y - a}{x},$$

where $a$ is the intercept and $b$ the slope. In our case, this gives

$$\frac{dy}{dx} = \frac{y}{x} - \frac{\mu_0^*}{x}.$$

*See Charnes, Cooper, Golany, Seiford and Stutz [36] for details.

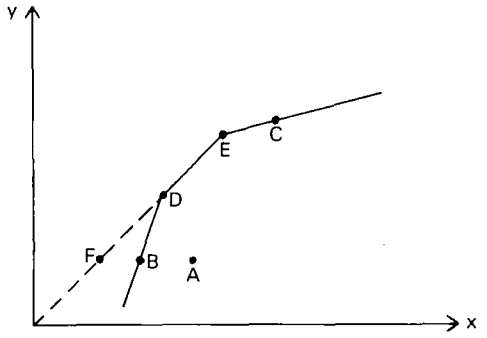


Fig. 2(a)

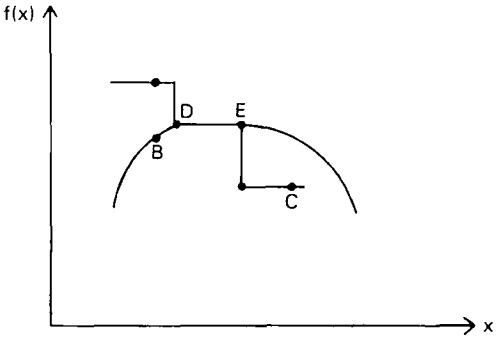


Fig. 2(b). Returns to scale and most productive scale size.

Note that we will then have

$$\frac{dy}{dx} \gtrless \frac{y}{x}$$

according to whether $\mu_0^* \lesseqgtr 0$ as in (25).

Via these simplified relations between output and input we can use these expressions to relate marginal product to average product in a diagrammatic variant of the usual returns-to-scale analysis in economics. In contrast to the usual 'smooth behavior' of the economics textbooks, 'marginal product' here assumes the form of a step function and the curve for 'average product' exhibits 'kinks' at points where changes in the marginal product occur.

Figure 2(b) shows the relations between marginal and average product, with the latter achieving its maximum value for all points in $D-E$, the interval where con-

stant returns to scale prevail. Returns to scale will occur in the order portrayed in fig. 2(a) for any DEA analysis, with changes from one locale to another being signalled by changes in the optimum basis as successive DMUs are brought into the functional of (24), where the sign of $\mu_0^\star$ shows which returns-to-scale situation applies locally. The extension to sensitivity analysis supplied in the third paper of the following three papers may be used to delineate the permissible range of variations before a change in the local returns-to-scale properties will occur.

Via the concept of Most Productive Scale Size (MPSS), Banker in [6] provides a development in which he shows that MPSS (= constant returns to scale) is achieved by any $DMU_0$ if and only if $\theta^\star = 1$. To see what this means, we may consider a $DMU_0$ with input-output vector $(X_0, Y_0)$ which is not at MPSS. In the sense defined by Banker, this means that it is possible to move from $(X_0, Y_0)$ to a new $(\alpha X_0, \beta Y_0) = (\hat{X}_0, \hat{Y}_0)$ with $\alpha, \beta > 0$ and $\beta > \alpha$, where $\hat{X}_0 \geqslant \Sigma_{j=1}^n X_j \hat{\lambda}_j$ and $\hat{Y}_0 \leqslant \Sigma_{j=1}^n Y_j \hat{\lambda}_j$ for some $\hat{\lambda}_j \geqslant 0$, $1 = \Sigma_{j=1}^n \hat{\lambda}_j$. In other words, $(X_0, Y_0)$ is not MPSS if it is possible to (i) achieve a more than proportionate increment in all components of $Y_0$ when $X_0$ is augmented (increasing returns to scale), or (ii) achieve a less than proportionate decrement in all components of $Y_0$ when $X_0$ is decremented (decreasing returns to scale). MPSS is achieved only when neither (i) nor (ii) is possible (constant returns to scale).

We proceed via a different route than the one utilized by Banker in [6] in order to tie these conditions to the ones specified in (25). We therefore first prove:

*Lemma 1:*

If $\min \theta = \theta^\star = 1$ in (7) then, Part One, $\min \hat{\theta} = \hat{\theta}^\star = 1$ in (23) and, Part Two, $\mu_0^\star = 0$ in (24).

*Proof:*

Assume $\theta^\star = 1$.

*Part One:* Since $\hat{\lambda}_{j_0} = 1$ and $\hat{\theta} = 1$ and all other $\hat{\lambda}_j = 0$ satisfies the constraints in (23), we must always have $\hat{\theta}^\star \leqslant 1$. Suppose we could have an optimal solution $\hat{\lambda}_j^\star$, $j = 1, \ldots, n$, for (23) with $\hat{\theta}^\star < 1$. This would also satisfy (7) which differs from (23) only because the latter contains the added constraint $\Sigma_{j=1}^n \hat{\lambda}_j = 1$. Hence the supposition that we could have $\hat{\theta}^\star < 1$ in (23) would contradict the optimality assumed for $\theta^\star = 1$ in (7). This proves Part One.

*Part Two:* Because (24), the dual to (23), contains one more variable than the dual to (7), we must have

$$\sum_{r=1}^{s} \hat{\mu}_r^{\star} y_{r0} - \mu_0^{\star} \geqslant \sum_{r=1}^{s} \mu_r^{\star} y_{r0} . \tag{26}$$

By virtue of the duality theory of linear programming and the preceding result, we have equality holding in the above expression. The expression on the right shows that an optimum is available for (24) with $\mu_0^{\star} = 0$. Furthermore, no other optimum can have $\mu_0^{\star} \neq 0$ when $\hat{\theta}^{\star} = 1$. To see that this is so, observe that $\mu_0^{\star}$ serves as a dual evaluator for $\Sigma_{j=1}^{n} \hat{\lambda}_j^{\star} = 1$ in (23). Thus, $\mu_0^{\star} \neq 0$ means that we could obtain an improvement in the optimal value of the functional by replacing this constraint with $\Sigma_{j=1}^{n} \hat{\lambda}_j = 1 + \delta$ for some $\delta \neq 0$. In particular, if $\mu_0^{\star} < 0$ then choosing $\delta < 0$ would decrease the optimal functional value. Similarly, the choice $\delta > 0$ would decrease the optimal functional value if $\mu_0^{\star} > 0$. Using duality theory again, the indicated reductions would yield a new $\hat{\theta}^{\star} < 1$, and because these solutions to the primal in (23) were already available for (7), we would again have a contradiction with $\theta^{\star} = 1$, the initially assumed optimum for (7). Q.E.D.

Next we start with (24) and show that $\mu_0^{\star} = 0$ implies $\theta^{\star} = \theta^{\star} = 1$:

*Lemma 2:*

If $\mu_0^{\star} = 0$ is part of an optimum for (24) then $\hat{\theta}^{\star} = \theta^{\star} = 1$, where $\hat{\theta}^{\star}$ is optimal for (23) and $\theta^{\star}$ is optimal for (7).

*Proof:*

An optimum with $\mu_0^{\star} = 0$ will also be optimal for the dual to (7) so that, via the dual theorem of linear programming, $\hat{\theta}^{\star} = \theta^{\star}$.

Now consider the following new problem:

$$\min \tilde{\theta} - \epsilon \left[ \sum_{i=1}^{m} \tilde{s}_i^{-} + \sum_{r=1}^{s} \tilde{s}_r^{+} \right]$$

subject to

$$\tilde{\theta} x_{i0} = \sum_{j=1}^{n} x_{ij} \tilde{\lambda}_j + \tilde{s}_i^{-}$$

$$y_{r0} = \sum_{j=i}^{n} y_{rj}\tilde{\lambda}_j - \tilde{s}_i^+$$

$$\tilde{\theta} = \sum_{j=1}^{n} \tilde{\lambda}_j$$

$$\tilde{\lambda},\ \tilde{s}_i^-,\ \tilde{s}_r^+ \geqslant 0, \tag{27}$$

and observe that this admits, but does not require $\tilde{\theta} = 1$. Now suppose we could have $\tilde{\theta}^\star < \hat{\theta}^\star$ where the latter is optimal for (23). This would mean that we could replace the last constraint in (23) with $\tilde{\theta}^\star = \Sigma_{j=1}^n \hat{\lambda}_j$ and achieve a reduction in $\hat{\theta}^\star$ which implies that $\mu_0^\star > 0$ for the corresponding dual evaluator in (24). Similarly, if $\tilde{\theta}^\star > \hat{\theta}^\star$ we would have $\mu_0^\star < 0$. The only other possibility is $\tilde{\theta}^\star = \hat{\theta}^\star$, in which case we would have $\tilde{\theta}^\star$ is optimal for (23) and hence is part of an optimal solution satisfying $\Sigma_{j=1}^n \hat{\lambda}_j^\star = 1$. This can only be true if $\tilde{\theta}^\star = 1$ with $\mu_0^\star = 0$ in (23). Combining this with our first paragraph in this proof we then have $\tilde{\theta}^\star = \hat{\theta}^\star = \theta^\star = 1$ as the only result which is consistent with $\mu_0^\star = 0$ in (24). Q.E.D.

As in other parts of DEA,* we have a variety of alternative characterizations which we list in the following:

*Theorem:* MPSS has been achieved when any of the following are optimal:

| in (7) | in (23) or (24) |
|---|---|
| $\theta^\star = 1$ or $\sum_{j=1}^{n} \lambda_j^\star = 1$ | $\theta^\star = 1$ or $\mu_0^\star = 0$ |

The projection of $A$ onto $B$ in fig. 2(a) assumed that formulae (13) and (14) were applied to optimal solutions to (23). However, there is no trouble in accommodating this to the preceding analysis, since this can be accomplished in the manner suggested by Banker, viz.,

*See the discussion on p. 432 in [41].

☆This broken line is sometimes referred to as the 'Reference Set'. Byrnes, Färe and Grosskopf [22] refer to it as a 'Reference Technology'.

$$\frac{\theta^* X_0 - s^{-*}}{e^{\mathrm{T}}\lambda^*}, \qquad \frac{Y_0 + s^{+*}}{e^{\mathrm{T}}\lambda^*}, \tag{28}$$

where $\lambda^*$ is the optimal solution vector to (7) and $e^{\mathrm{T}}\lambda^* = \Sigma_{j=1}^{n} \lambda_j^*$ with, as Banker [6] notes,

$$\sum_{j=1}^{n} \lambda_j^* \lesseqgtr 1, \tag{29}$$

according to whether returns to scale are increasing, constant, or decreasing, respectively. As the above formulae indicate, however, achievement of MPSS with constant returns to scale does not guarantee achievement of 100% efficiency. The latter also requires $s^{+*} = s^{-*} = 0$, which is automatically achieved in the CCR formulation.

These economies of scale are local, with changes occurring as movement is effected from one facet to another in the course of DEA analysis (see the third paper in the following series). In the kinds of DEA analyses heretofore available, one can only obtain 'concave' production functions (see the basic paper [36] for proofs). A production function possessing convex portions with *large* local economies of scale would thus be totally missed.* Needed methods for uncovering the presence of such possibilities are suggested in [36], along with ways of showing some DMUs to be efficient that would otherwise be rated as inefficient.

Of fundamental importance in applications is the choice of DMUs along with the inputs and outputs to be used in evaluating their activities. Most uses of mathematical programming are for 'planning' managerial activity where the inputs and outputs to be considered are confined, by and large, to those which can be varied at the discretion of management. DEA introduces 'control' features revolving around actual accomplishments in which some of the inputs (and perhaps some of the outputs) are at least partly non-discretionary.

Inputs or outputs are said to be non-discretionary when their values are not subject to management control.☆ Even when inputs are wholly non-discretionary, they may nevertheless need to be considered in arriving at relative efficiency evaluations. In the first paper that follows, for example, the effects of possibly differing weather

*Even an isotone function can have such portions. See [36].

☆Banker and Morey [14] refer to the values of these variables as being determined exogenously.

conditions on the sortie rates of fighter aircraft needs to be taken into account in evaluating the maintenance efficiencies at different bases. This may be done in various ways in a DEA analysis, of course, but the important point is that such non-discretionary inputs should be included whenever they are believed to have significant effects.

One way to approach this problem is provided by the following model due to Banker and Morey [15]. Using $D$ and $N$ to represent the index sets for Discretionary and Non-Discretionary inputs, respectively, this model replaces (22) with

$$\min \theta - \epsilon \left[ \sum_{i \in D} s_i^- + \sum_{r=1}^{s} s_r^+ \right]$$

subject to

$$\text{For } i \in D: \quad \theta x_{i0} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-$$

$$\text{For } i \in N: \quad x_{i0} = \sum_{j=1}^{n} x_{ij}\lambda_j + s_i^-$$

$$\text{For } r = 1, \ldots, s: \quad y_{r0} = \sum_{j=1}^{n} y_{rj}\lambda_j - s^+$$

$$1 = \sum_{j=1}^{n} \lambda_j$$

$$\lambda_j,\ s_i^-,\ s_r^+ \geqslant 0 . \tag{30}$$

Particular attention might be called to the fact that the slacks for $i \in N$ are not represented in the functional. That is, they are treated in the same manner as the slacks in ordinary linear programming.

To bring all this more clearly into view, we formulate the dual to (30) for direct comparison with (24) as follows:

$$\max \sum_{r=1}^{s} \mu_r y_{r0} - \sum_{i \in N} \nu_i x_{i0} - \mu_0$$

subject to

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i \in N} \nu_i x_{ij} - \mu_0 - \sum_{i \in D} \nu_i x_{ij} \leqslant 0$$

$$\sum_{i \in D} \nu_i x_{i0} = 1$$

$$\mu_r \geqslant \epsilon > 0 \quad \text{for} \quad r = 1, \ldots, s$$

$$\nu_i \geqslant \epsilon > 0 \quad \text{for} \quad i \in D$$

$$\nu_i \geqslant 0 \quad \text{for} \quad i \in N . \tag{31}$$

Note that $\epsilon$, the non-Archimedean constant, continues to be used for all outputs, as before. It also continues to apply for the inputs that are discretionary but not for the inputs that are non-discretionary. The latter are constrained only to non-negative ranges so that values of $\nu_i^* = 0$ are possible for $i \in N$ – a condition which is sometimes referred to as corresponding to the assumption of 'free disposability' in the literature of economics.* This assumption does *not* apply to the non-Archimedean terms, however, which are assumed to have 'some (positive) value' that needs to be considered in their use or non-use.

For purposes of further interpretation, we now reverse the transformations in (3) and apply them to (23). This gives

*This term, which seems to have originated with T. Koopmans [61], is employed extensively by R. Färe and his associates who also refer to it as 'strong disposability'. See [22] and the references cited therein.

$$\max \frac{\sum_{r=1}^{s} u_r y_{r0} - \sum_{i \in N} v_i x_{i0} - u_0}{\sum_{i \in D} v_i x_{i0}}$$

subject to

$$\frac{\sum_{r=1}^{s} u_r y_{rj} - \sum_{i \in N} v_i x_{ij} - u_0}{\sum_{i \in D} v_i x_{ij}} \leqslant 1$$

$$v_i \geqslant 0, \quad i \in N;$$

$$\frac{v_i}{\sum_{i \in D} v_i x_{i0}} \geqslant \epsilon > 0, \quad i \in D; \quad \frac{u_r}{\sum_{i \in D} v_i x_{i0}} \geqslant \epsilon > 0, \quad r = 1, \ldots, s. \tag{32}$$

Using the terminology of [26], Banker and Morey [14] interpret the objective in (32) as being directed to maximizing the 'net virtual surplus' – – i.e., the surplus of outputs minus fixed inputs [see (16)–(18)]. In a more complete characterization, it maximizes the *rate* of net virtual surplus per unit discretionary input utilized.

Banker and Morey [14] show how to extend this to the case of non-discretionary outputs and how to decompose the resulting efficiencies into separately identifiable scale and technical efficiencies.* We shall not pursue these topics further except to note that extensions to *partially* controllable inputs and outputs☆ require the introduction of additional constraints while the question of simultaneous treatment of non-discretionary inputs and outputs remains open. Even when some inputs are completely non-discretionary, as in the treatments from Banker and Morey de-

*Banker and Maindiratta [13] show how to extend this analysis to the case of 'price' or 'allocative efficiencies' which take account of prices charged for different inputs and received for different outputs.

☆I.e., outputs or inputs which are discretionary only within certain ranges.

scribed above, one might argue that relative efficiency ratings should be secured only from DMUs with non-discretionary inputs of comparable magnitudes to those for $DMU_0$. This, too, might be handled by adding constraints while retaining the usual objective function. See [36] for suggested treatments of these topics.

## 5. History and related work

The first publication of Data Envelopment Analysis appeared in 1978 in the article [41] which the authors published jointly with E. Rhodes* (see also [42], [69] and [21]). This work had been initiated three years earlier in collaborative research with E. Rhodes to develop more adequate methods of evaluating some of the government supported programs for disadvantaged children that were then being undertaken in public schools. The applications to Program Follow Through, a large-scale experiment in U.S. public school education, which had been used to guide the development of DEA, were published jointly with E. Rhodes in [40]. They also appear elsewhere in fuller detail (see the doctoral thesis of E. Rhodes [69]).

As might be expected, it is possible to relate DEA to on-going developments by others. An indication of DEA's relation to some of this other work may form a fitting close to this preface. Only a brief synopsis will be given, however, and interested persons should consult some of the references we cite for further bibliographical detail.

Leaving aside topics like fractional and linear programming and similar precursors published in the literature of operations research and management science, we can locate two strands of research in the literature of economics that are related to DEA. One strand originates in the publications of R.W. Shepard – [74, 75] – dealing with relations of 'duality'☆ between cost and production functions. The other strand originated in the publication [55] by M.J. Farrell,† which was concerned with developing an improved alternative to the customary measures of 'productivity'.††

Although Farrell's work was heavily empirical from the start, the same was not true for Shephard whose work was almost exclusively concerned with formalizing and rigorously establishing the relations he was studying. This initial emphasis has been continued in subsequent work. Afriat in [1], for instance, has continued to

* A still earlier, synoptic treatment of DEA was published in [31].

☆ We have elsewhere suggested that these might better be called 'transform relations'. See [43].

† Farrell cites Debreu [51] as a source of some of his ideas.

†† Farrell subsequently expanded this work to considerations of 'returns to scale' efficiency in his work with Fieldhouse [56].

emphasize formal analyses in the tradition of Shephard,* although Hanoch and Rothschild [60] and Varian [77], Diewert and Parkan [53] and others have begun to turn this work in other directions, as the latter authors have begun to provide formulations which can be applied to empirical data to see whether cost, profit and production function structures are satisfied as postulated in economics.

Although it can also be related to these kinds of undertakings,☆ Farrell's work was directed mainly toward developing a summary measure of the efficiency of the behavior of different economic entities. This tradition has continued into the present, as represented in the work of R. Färe and his associates (see [23] and [54]), along with Lovell and Schmidt [67] and others. Färe, in particular, has sought to build on Farrell's classifications into technical, scale and allocative (or price) efficiencies in order to devise a scalar measure that can be decomposed for identifying each of these efficiencies separately – as well as jointly – in one measure of 'overall' efficiency. Continuing in the tradition initiated by Farrell, a great deal of this work has been rich in empirical use as well as theoretical development and interpretation.

Almost all of the work in 'Farrell Efficiency' has been restricted to single output situations. Farrell, as well as others working in the tradition he initiated, describe what is to be done in extending their methods for use in the case of multiple outputs (see Farrell [55], p. 257). They do not, however, supply what is needed in the way of precise mathematical details with accompanying definitions and interpretations.

Some of the many new elements (and pitfalls) that enter into extensions to the multi-product case may be indicated by turning to yet another tradition in the economics literature. Almost all of classical 'theory of the firm' literature is heavily oriented to the single output case. In a recently published book [64], K. Laitinen seeks to extend that theory to the multi-product case in a relatively straightforward manner (i.e. with traditional calculus formulations and analyses). This extension is not so easy to come by, however, as can be seen from Laitinen's attempted extensions. Witness, for instance, his definition of the isoquant for any specified output vector as representing all combinations of inputs which are just sufficient to produce this output vector.†

Although satisfactory for the single output case,†† this definition fails to allow for the possible presence of output inefficiencies (and related difficulties) which are encountered when moving from the single to the multiple output case (see table 8

* See McFadden [68] and Diewert [52].

☆ See the discussion on p. 255 ff. in [43].

† See p. 16 ff. in [64].

††At least in situations where the reasoning is from actually observed outputs (i.e. technically achievable outputs), as in DEA, since otherwise one must assume that the chosen (or assumed) output levels can actually be produced.

in the immediately following paper for an example involving output (as well as input) inefficiencies). The relation between outputs and between inputs, as well as between outputs and inputs, all need to be simultaneously considered. Especially when actual applications are to be undertaken, this can be a complicated and cumbersome task which DEA avoids by proceeding non-parametrically.

The use of Pareto optimality to define and characterize DEA efficiency was also a novel element in this literature, as was the introduction of non-Archimedean terms and the slack variables with which they are associated. Turning again to Laitinen [64] we can observe some of the problems that occur when proceeding without attention to these matters. Laitinen (along with others) uses Shephard's 'distance function'* to define the production function in terms of 'unit input distance'. That is, the level line intersecting the production surface is defined as the input combinations at unit distance from the origin which will yield the specified output vector (Laitinen [64, pp. 16–17]). This is not satisfactory for defining the corresponding efficiency surface or production function, however, especially when applications to actual data are involved, since the resulting characterization can confuse frontier properties with efficiency, as was noted in our discussion of fig. 1.☆

The problem of distinguishing between frontier and efficiency properties has been present from the start in the literature we have been citing. There is reason to suppose that Shephard was aware of the problem although he never dealt with it in a specific manner. Farrell dealt with it by introducing 'points at infinity' as part of the production possibility set, but he was unable to give this concept operational form in the models he was using. Färe and Lovell in [54] discuss this in detail, where they attempt to provide a single real number efficiency measure which they refer to as 'Russell Measure'. Färe with others has continued along this route in a manner which generally requires multiple linear programs to obtain such a measure (see [22]).

As already noted, DEA uses a different approach by introducing non-Archimedean elements that maximize the slack without disturbing the value of $\theta$. As a result, an inefficient vector cannot be a member of an optimum basis and the solution of only one linear programming problem provides all that is required for technical and returns-to-scale efficiency characterizations. Working with R. Thrall, the authors of this preface have also shown how to obtain a transformed problem that does not require any non-Archimedean elements although, in general, it is computationally more efficient to work with the direct DEA models and algorithms that we have been describing (see [46]).

Another and probably more striking difference is that DEA is 'empirically based'. The way in which this differs from other approaches is best illustrated by contrast with Laitinen's 'theory based' approach to testing and estimation. Using

*Really a 'gauge function' in the sense of Brunn, Minkowski and Fenchel [58].
☆See also the discussion in Banker, Charnes and Cooper [9].

data generated from artificially selected (known) functions, Laitinen was unable to do anything very much with the troubles he encountered. In particular, his theory based approach left him unable to identify the sources of his troubles or how, and in what manner, they might be repaired in any specific manner. He was able only to conjecture possible problems with modeling or statistical specification which would require another fresh approach without any guarantee of repair. DEA, on the other hand, turns attention to each observation and, as illustrated in the second of the following papers, it provides operationally implementable formulae directed to correcting the troubles identified for each such point relative to the ensemble of observations and the model used to identify these problems.

In a sense, DEA explicitly joins together the two strands of work initiated by Farrell and Shephard. It also goes further by providing machinery not only for testing and estimation, but also for modifying observational data to produce what is required. Thus, DEA provides an 'empirically based' theory of production and related behaviors that can be used for control and correction as well as for prediction and understanding.* As noted in the second of the following three papers, DEA therefore provides a new method of estimation which can be used as an alternative to customary statistical regressions and, of course, combinations of the two can yield additional new alternatives.

Still further possibilities have been brought into view via the concurrent research reported in [36]. One aspect of this research provided the following as a new DEA form with ties to 'goal programming' and the operationally implementable form to test for Pareto optimality that is given in chapter IX of [28]:☆

$$\min - \left[ \sum_{r=1}^{s} s_r^+ + \sum_{i=1}^{m} s_i^- \right]$$

subject to

$$\sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ = y_{r0}$$

*See Charnes, Cooper, Learner and Phillips [37] for further discussion.

☆The material in this chapter was originally presented at the 1958 Econometric Society meetings in Chicago. See [36] for further details.

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{i0}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \ , \tag{33}$$

with all variables constrained to be non-negative just as in (22). As in chapter IX of [28], we have Pareto efficiency for $(Y_0, X_0)$ – with components $y_{r0}, x_{i0}$, $r = 1, \ldots, s$; $i = 1, \ldots, m$ – if and only if the optimal value of the functional is zero in the above model. Furthermore, if the test for Pareto efficiency is not passed then the non-zero slacks in the optimal solution show where (and in what amounts) the needed adjustments may be made.

This identification of efficiency estimation with goal programming allows us to close this preface by referring to yet another strain of work initiated by Aigner and Chu [3], which is concerned with 'frontier estimation' in a still different manner than the other work we have been citing. Using an explicitly assumed Cobb–Douglas form* for the production function, Aigner and Chu in [3] use a goal programming approach with only one-sided deviations permitted (as described in [35] and [28]) to obtain the relevant parameter values from observational data.

The original work by Aigner and Chu assumed a determinsitic frontier. Aigner, Amemiya and Poirier [2], Aigner, Lovell and Schmidt [4], and others, have extended this to the case of 'stochastic frontiers' – i.e. frontiers which reflect the behavior of stochastic elements that can affect the input and output behaviors.

Førsund, Lovell and Schmidt in [59] describe this kind of work in a manner that is still relatively up to date. Incomplete and confined to the case of single outputs, this work is not usable for the multiple output situations that are characteristic of public enterprises. Moreover, DEA has again introduced new directions of work in the form of statistical distributions of DEA efficiency measures which are not covered in this other work.☆ In particular, research conducted by the authors in collaboration with E. Rhodes has resulted in a canonical form for analyses of such statistical distri-

*This was the form also assumed by Farrell in [55] to check and interpret the results he had already secured in the illustrative empirical application he provides on p. 275 ff. See also Farrell and Fieldhouse [56] p. 258.

☆Farrell and Fieldhouse in [56] p. 263 report their use of plots of efficiency measures to study the effects of data groupings, but this work is differently motivated and directed than the treatment of statistical distributions of the efficiency measures per se.

butions of the efficiencies obtained from a collection of DMUs. As reported in [30], this type of analysis forms a part of what is required to obtain comparative evaluations of the efficiency of different 'programs' *after* the effects of 'managerial inefficiencies' have been identified and corrected. The ability to distinguish between 'managerial efficiency' and 'program efficiency', we might also add, depends on the ability to make adjustments such as DEA makes possible by means of formulae such as (13) and (14) and, as indicated in [40] and [41], still other refinements and extensions are also possible.

This distinction between 'managerial' and 'program' efficiency also provides a new type of efficiency (or inefficiency) which should be added to the tri-partite collection introduced by Farrell in [55]: (i) technical efficiency, (ii) scale efficiency and (iii) price or allocative efficiency. Only the last named member of this set has not been dealt with in this preface.* Banker and Maindiratta [13] do deal with this via a DEA approach which extends the work of Varian [77] on this topic. Our own belief is that goal programming offers a better approach in many cases, not only because of the multiple objective character of most not-for-profit entities, but also because of the impossibility of pricing many of their outputs.☆ As noted by Bowlin [20], the joining of DEA and goal programming effected in [36] makes it possible to utilize the coefficients obtained from DEA as an element of the goal programming models used. See also Schinnar [72] for a use of DEA to obtain 'efficient coefficients' for use in input-output analyses of Leontieff type.

A great variety of additional possibilities for research as well as applications are open for attention. The papers that follow should help to push matters along and also indicate further opportunities for both research and use of DEA and like approaches.

*Farrell expressed grave reservations about any results that might be secured from studying price efficiency, even in private sector applications.

☆Cf., e.g. the aircraft sorties that constitute one of the outputs in the paper that follows.