Tuning SFA Results for Use in DEA[♣]

Kaoru Tone[a,*], Miki Tsutsui[a,b]

[a] *National Graduate Institute for Policy Studies*

7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan

[b] *Central Research Institute of Electric Power Industry*

2-11-1 Iwato Kita, Komae-shi, Tokyo 201-8511, Japan

[♣] This research is supported by Grant-in-Aid for Scientific Research (c) Japan Society for the Promotion of Science.

[*] Corresponding author.
E-mail addresses: tone@grips.ac.jp (K. Tone), miki@criepi.denken.or.jp (M. Tsutsui).

Tuning SFA Results for Use in DEA

Abstract

After pointing out shortcomings of the traditional adjustment scheme for combining SFA results for use in DEA in the three stage approach, we propose a new scheme. We demonstrate the effect of this adjustment formula using an electric utility data set.

Keywords: DEA, SFA, data adjustment, multi-stage approach

Tuning SFA Results for Use in DEA

1. Introduction

Data envelopment analysis (DEA) has been widely utilized for evaluating relative efficiency of organizations with multiple input resources and output products. DEA employs mathematical programming techniques and mainly deals with data set that are supposed deterministic. Since the objective organizations, called Decision Making Units (DMUs), may belong to several different operational environments and their data may subject to statistical noise, it is strongly demanded that the true managerial efficiency should be identified after accounting (deleting) the operating environment effects and statistical noise on the data. For this purpose, Fried et al. (2002) proposed a three-stage procedure that combines DEA and stochastic frontier analysis (SFA) as follows. At the first stage, they employ DEA for finding slacks of each DMU that constitute the elements of inefficiency. At the second stage, they apply SFA to explain these slacks in terms of the operating environment, statistical noise and managerial efficiency. Then, they adjust the first-stage data set by purging the influence of the operating environment and statistical noise. Lastly, they apply DEA to the adjusted data set at the third stage. Avkiran and Rowlands (2006) further developed Fried et al. (2002) within the non-radial DEA model, i.e., the slacks-based measure (SBM) introduced by Tone (2001).

This paper focuses on their data adjustment schemes. Firstly, we point out irrationality of their adjustment formulae in that their adjustments consist of positive translation of the regressed terms so that the adjusted data should be non-negative, since most DEA models require non-negative data set. However, this operation causes serious bias in the third stage DEA scores. We will demonstrate this fact using examples. Then we propose a new procedure for tuning SFA results for use in the third stage DEA.

This paper unfolds as follows. In Section 2, we briefly survey the multi-stage use of DEA and SFA. Readers are recommended to refer to Fried et al. (2002) and Avkiran and Rowlands (2006) for detailed discussions on the motivation of the multi-stage approach. In Section 3, we will demonstrate the irrationality of their adjustment scheme that combine the SFA results with the original data set. Then, we propose a new tuning scheme for adjusting the SFA results for use in the third stage DEA in Section 4. Comparisons of our proposed scheme with the previous one is presented in Section 5. Some concluding remarks follow in Section 6.

2. Multi-stage Use of DEA and SFA

2.1 Multi-stage approach

We deal with n DMUs with the input matrix $X \in R_+^{m\times n}$ and output matrix $Y \in R_+^{s\times n}$, where m and s are numbers of inputs and output, respectively. For the target DMU (x_o, y_o), where $x_o \in R_+^m$ and $y_o \in R_+^s$ are input and output of the DMU, we express them in terms of X, Y, the intensity vector $\lambda \in R_+^n$, the input slacks $s^- \in R_+^m$ and the output slacks $s^+ \in R_+^s$ as follows:

$$\begin{aligned} x_o &= X\lambda + s^- \\ y_o &= Y\lambda - s^+ \end{aligned} \tag{1}$$

Both Fried et al. (2002) and Avkiran and Rowlands (2006) evaluate the input slacks $s^- \in R_+^m$ and output slacks $s^+ \in R_+^s$, which represent inefficiency of DMU (x_o, y_o), by means of DEA models. Difference exists in the DEA models utilized as follows.

Fried et al (2002) employs the input-oriented BCC model (Banker et al. (1984)):

$$
\begin{aligned}
&\min \theta \\
&\text{subject to } \theta x_o = X\lambda + s^- \\
&\qquad y_o = Y\lambda - s^+ \\
&\qquad e\lambda = 1 \\
&\qquad \lambda \geq 0,\ s^- \geq 0,\ s^+ \geq 0,
\end{aligned} \tag{2}
$$

where $e \in R^n$ denoted a row vector in which all elements are equal to 1.

Avkiran and Rowlands (2006) utilizes the non-radial slacks-based model (SBM) introduced by Tone (2001):

$$
\begin{aligned}
&\min \rho = \frac{1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^-}{x_{io}}}{1 + \frac{1}{s}\sum_{r=1}^{s} \frac{s_r^+}{y_{ro}}} \\
&\text{subject to } x_o = X\lambda + s^- \\
&\qquad y_o = Y\lambda - s^+ \\
&\qquad e\lambda = 1 \\
&\qquad \lambda \geq 0,\ s^- \geq 0,\ s^+ \geq 0.
\end{aligned} \tag{3}
$$

Refer to Avkiran and Rowlands (2006) for comparisons of these two approaches. We will not go into the details but just denote the optimal slacks obtained by s^- and s^+.

Both papers regard these slacks as the sources of inefficiencies. However, actual performances are likely to be attributable to some combination of managerial inefficiencies, environmental effects and statistical noise. Thus, they tried to isolate these three effects using stochastic frontier analysis (SFA) in the second stage. The general function of the SFA regressions is represented in Eq. (4) below for the case of input slacks.

$$
s_{ij}^- = f^i(z_j; \beta^i) + v_{ij} + u_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n, \tag{4}
$$

where s_{ij}^- is the stage 1 slack in the *i*th input for the *j*th unit, z_j the environmental variables, β^i the parameter vectors for the feasible slack frontier and $v_{ij} + u_{ij}$ the compounded error structure where $v_{ij} \approx N(0, \sigma_{vi}^2)$ represents statistical noise and $u_{ij} \geq 0$ represents managerial inefficiency.

2.2 Adjustments of Original Data by SFA Results: Previous Studies

Fried et al. (2002) and Avkiran-Rowlands (2006) proposed the following adjustment schemes.

(a) Fried et al. (2002) adjust the input data by deleting significant environmental effects and statistical noises as follows:

Input adjustment

$$x_{ij}^A = x_{ij} + \left[\max_{k\in N}\left\{z_k\hat{\beta}^i\right\} - z_j\hat{\beta}^i\right] + \left[\max_{k\in N}\left\{\hat{v}_{ik}\right\} - \hat{v}_{ij}\right] \quad (5)$$

(b) Avkiran and Rowlands (2006) adjust the output data as follows:

Output adjustment

$$y_{rj}^A = y_{rj} + \left[z_j\hat{\beta}^r - \min_{k\in N}\left\{z_k\hat{\beta}^r\right\}\right] + \left[\hat{v}_{rj} - \min_{k\in N}\left\{\hat{v}_{rk}\right\}\right] \quad (6)$$

The role of max and min in the above formulas is to ensure the adjusted data $\left\{x_{ij}^A\right\}$ and $\left\{y_{rj}^A\right\}$ to be positive, since most DEA models demand the data set to be positive. This operation is a translation of the SFA results. Actually, in the input adjustment case, let us define $\hat{z}_i \equiv \max_{k\in N}\left\{z_k\hat{\beta}^i\right\}$ and $\hat{v}_i \equiv \max_{k\in N}\left\{\hat{v}_{ik}\right\}$. Then $\hat{z}_i$ and $\hat{v}_i$ are fixed (constant) for all DMUs within the input item *i*. Thus, (5) can be written as

$$x_{ij}^A = x_{ij} - z_j\hat{\beta}^i - \hat{v}_{ij} + \hat{z}_i + \hat{v}_i$$

As this formula indicates, the SFA results are translated by $\hat{z}_i + \hat{v}_i$ for each *i*. In the next section, we point out the troubles that this translation induces.

3. Shortcomings of Previous Adjustments

We will demonstrate irrationality of the above adjustment scheme using two examples as follows.

3.1 Two DMUs with single input and single output case

The adjustment formulae (5) and (6) are introduced so that the adjusted values are assured to be non-negative or positive. This means a positive translation of the adjusted data. Now, we investigate how a positive translation effects DEA efficiency scores using a simple example. This example deals only with translation issues but not with environmental and noise issues.

Table 1 exhibits two DMUs A and B with a single input x and a single output y. We translate the input x by k. Thus, A's input is $1+k$ while B's is $2+k$. Figure 1 depicts these shifts from A to A' and from B to B'. We translate only input values but keep the output values unchanged.

Table 1. A simple example

	Input	Output	Translated Input	Output
	x	y	$x+k$	Y
A	1	2	$1+k$	2
B	2	1	$2+k$	1

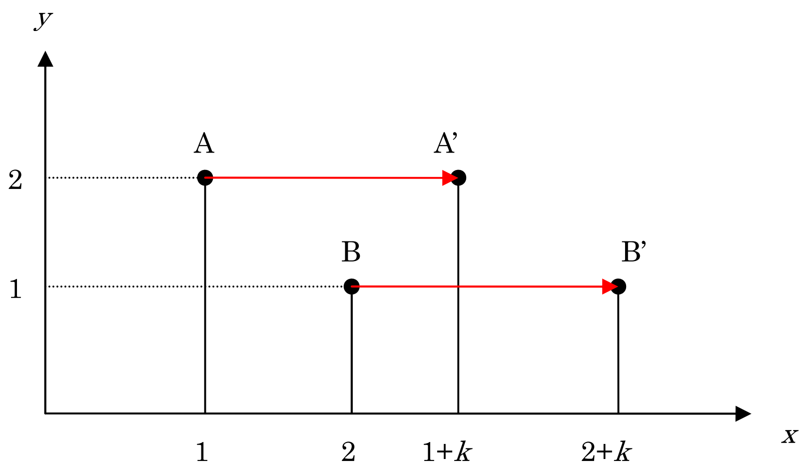

Figure 1. Input Translation

In both cases, i.e., the original and the translated cases, A and A' are efficient and B and B' are inefficient compared with A and A', respectively.

The radial and input-oriented DEA efficiency scores of B' are calculated in terms of k as follows:

Under the constant returns-to-scale assumption (CRS) (CCR-I)

$$\theta_C(k) = \frac{1+k}{2(2+k)}. \tag{7}$$

Under the variable returns-to-scale assumption (VRS) (BCC-I)

$$\theta_V(k) = \frac{1+k}{2+k}. \tag{8}$$

We notice that under this single input and single output case the input-oriented SBM models give the same efficiency value with the radial models. They are monotone increasing in k and hence the difference in efficiency between A' and B' is monotone decreasing in k. Actually, the BCC-I score of B' tends to unity (that of A') as k tends to infinity. This simple example demonstrates that the input translation factor *k* effects the efficiency score significantly and indicates that the adjustment formulae (5) and (6) suffer from the max and min values included that are translation terms in the respective formula. The next example will evidence this fact.

3.2 A multi-stage example

We demonstrate irrationality of the adjustment formula (5) using an actual data set.

3.2.1 Data and statistics

We employed the data from U.S. and Japan electric utilities (48 U.S. and 8 Japan) during the years 1990-2001. We count a utility at a certain year as an independent DMU and, after deleing outliers, we obtained 351 utilities as our DMUs. We employed three inputs and one output as follows:

Input

Input 1: The total nameplate capacity of electric power plants measured in Mega Watts (MW)

Input 2: The consumed fuel converted to British Thermal Units (BTU)

Input 3: The number of employees

Output

Output 1: The generated electric power measured in Mega Watt hours (MWh)

Statistics on the data are displayed in Table 2.

Table 2 Statistics of the Data

	Input 1: Name Plate Capacity (MW)	Input 2: Fuel (BTU)	Input 3: Employee (1/10,000)	Output 1: Generation (MWh)
Average	0.7575	0.7958	1.6174	1.1797
Min	0.1310	0.0994	0.1352	0.1825
Max	2.2682	2.4959	7.3860	3.4374
S. D.	0.5146	0.5477	1.4848	0.7866

3.2.2 DEA model

We employed the input-oriented SBM under the variable returns-to-scale (VRS) assumption.

3.2.3 First stage DEA

The results of the 1st stage input-oriented SBM are summarized in Table 3.

Table 3: 1st Stage SBM Results

	Average	Min	Max	S.D.
SBM score	0.7188	0.4509	1	0.1378

3.2.4 Second stage SFA

We applied SFA for the optimal input slacks obtained in the 1st stage SBM. We employed several environmental factors consisting of non-discretionary, discretionary and dummy variables which are out of control of DMUs. We utilized LIMDEP 8.0 for this purpose.

3.2.5 Adjustments

We adjusted the slacks and hence the inputs using the SFA results by means of the formula (5). In this formula, the terms $\hat{z}_i \equiv \max_{k \in N}\{z_k \hat{\beta}^i\}$ and $\hat{v}_i \equiv \max_{k \in N}\{\hat{v}_{ik}\}$ are fixed (constant) for all DMUs within the input i. Hence, the adjustment formula (5) becomes to a translation as we denoted in the preceding section.

We record these max terms for each input item in Table 4.

Table 4: The Max Values

	Input 1	Input 2	Input 3
$\hat{z}_i \equiv \max_{k \in N}\{z_k \hat{\beta}^i\}$	0.1137	-0.001973	1.3990
$\hat{v}_i \equiv \max_{k \in N}\{\hat{v}_{ik}\}$	0.5241	0.5746	4.8942

Statistics of the adjusted data are summarized in Table 5.

Table 5: Statistics of the Adjusted Data

	Input 1: Name Plate Capacity (MW)	Input 2: Fuel (BTU)	Input 3: Employee (1/10,000)	Output 1: Generation (MWh)
Average	1.2591	1.2651	7.3291	1.1797
Min	0.7794	0.7239	6.4356	0.1825
Max	3.1366	3.0890	14.9754	3.4374
S. D.	0.4642	0.5120	1.0646	0.7866

3.2.6 Third stage DEA

We applied the input-oriented SBM under variable returns-to-scale assumption to the adjusted data set. Statistics of the efficiency score are recorded in Table 6.

Table 6: 3rd Stage SBM Results

	Average	Min	Max	S.D.
SBM score	0.9852	0.9132	1	0.0158

Comparisons of Table 3 and Table 6 demonstrate a big change in the average score: from 0.7188 to 0.9852. Figure 2 compares the distributions of the efficiency scores at the 1st and 3rd stage SBM. This level up might be caused by the adjustment formula (5) using the max values for preventing negative input values. The results of the 3rd stage SBM almost lost the discriminating power in efficiency evaluation and are unacceptable. Although we described our experiences with the VRS model, we have experienced similar odd results under the constant returns-to-scale (CRS) assumption.

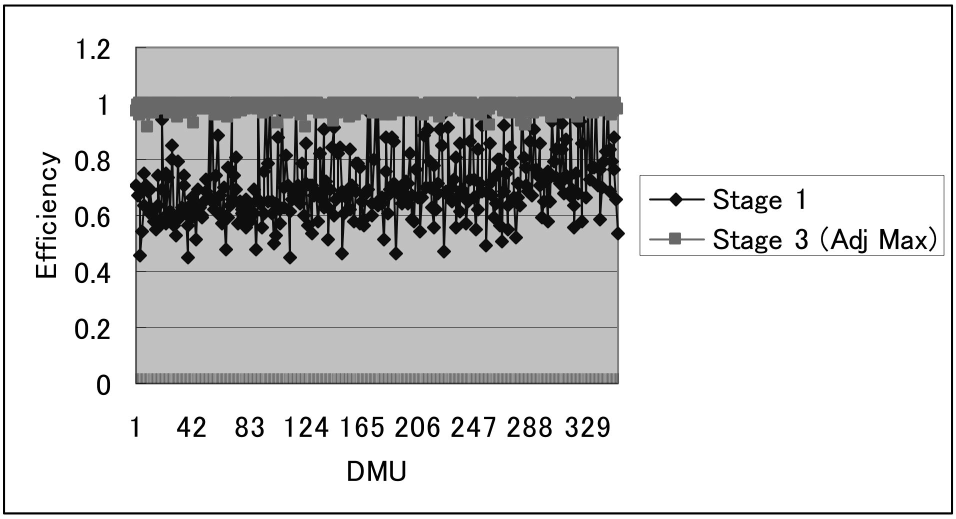

Figure 2: Comparisons of Stage 1 and Stage 3 Efficiency Scores

4. A New Tuning of SFA Results

In this section, we propose a new adjustment scheme.

4.1 Re-adjustments

First, we employ the SFA formula for adjustment with no recourse to max or min as follows .

Input adjustment

$$x_{ij}^{A} = x_{ij} - z_j \hat{\beta}^{i} - \hat{v}_{ij} \tag{9}$$

Output adjustment

$$y_{rj}^{A} = y_{rj} + z_j \hat{\beta}^{r} + \hat{v}_{rj} \tag{10}$$

Then we re-adjust them into x_{ij}^{AA} or y_{rj}^{AA} using the following formulas.

Re-adjustment

Input

$$x_{ij}^{AA} = \frac{x_{i\max} - x_{i\min}}{x_{i\max}^{A} - x_{i\min}^{A}} (x_{ij}^{A} - x_{i\min}^{A}) + x_{i\min} \qquad (i = 1, \ldots, m : j = 1, \ldots n) \tag{11}$$

where

$$x_{i\max} = \max_{k \in N} \{x_{ik}\}, x_{i\min} = \min_{k \in N} \{x_{ik}\}, x_{i\max}^{A} = \max_{k \in N} \{x_{ik}^{A}\}, x_{i\min}^{A} = \min_{k \in N} \{x_{ik}^{A}\}.$$

Output

$$y_{rj}^{AA} = \frac{y_{r\max} - y_{r\min}}{y_{r\max}^{A} - y_{r\min}^{A}} (y_{rj}^{A} - y_{r\min}^{A}) + y_{r\min} \qquad (r = 1, \ldots, s : j = 1, \ldots n) \tag{12}$$

where

$$y_{r\max} = \max_{k \in N} \{y_{rk}\}, y_{r\min} = \min_{k \in N} \{y_{rk}\}, y_{r\max}^{A} = \max_{k \in N} \{y_{rk}^{A}\}, y_{r\min}^{A} = \min_{k \in N} \{y_{rk}^{A}\}.$$

4.2 Rationale

The proposed re-adjustment scheme has the following properties:

(1) x_{ij}^{AA} increases in x_{ij}^{A}. Thus, the re-adjusted data have the same ranking with the adjusted data. Actually x_{ij}^{AA} is a linear transformation of x_{ij}^{A} with a positive coefficient. The coefficient and the constant term of this linear transformation are constant within the respective input item *i*.

(2) At $x_{i\max}^{A}$, $x_{i\max}^{AA}$ attains the maximum value $x_{i\max}^{AA} = x_{i\max}$.

(3) At $x_{i\min}^{A}$, $x_{i\min}^{AA}$ attains the minimum value $x_{i\min}^{AA} = x_{i\min}$.

Hence, the re-adjusted data set $\{x_{ij}^{AA}\}$ remains in the range $[x_{i\min}, x_{i\max}] (\forall i)$, and the maximum and minimum values are the same between $\{x_{ij}^{AA}\}$ and $\{x_{ij}\}$.

For the output side, we have the same property: the re-adjusted data set $\{y_{rj}^{AA}\}$ remains in the range $[y_{r\min}, y_{r\max}] (\forall r)$, and the maximum and minimum values are the same between $\{y_{rj}^{AA}\}$ and $\{y_{rj}\}$.

These properties are appealing in that they eliminate ambiguity regarding the range of adjusted input and output values that effect the DEA scores significantly as we have shown in the previous examples. Furthermore, when we start the first stage DEA, we usually confirm that the ranges of input and output values are appropriate for the chosen DEA model. (We delete outliers before going into the first stage.) Therefore, it is not odd to keep the ranges status quo and re-evaluate the DEA efficiency score at the third stage using the re-adjusted data set.

5. Numerical Comparisons

We re-adjust the US electric utility data set and compare the results.

Using the formula (9) (but not using the max in (5)), we adjusted the input data, and then re-adjusted the data by the formula (11). Table 7 displays the statistics of the re-adjusted data. As expected, the min and max values are the same with the original data in Table 2.

Table 7 Statistics of the Re-adjusted Data

	Input 1: Name Plate Capacity (MW)	Input 2: Fuel (BTU)	Input 3: Employee (1/10,000)	Output 1: Generation (MWh)
Average	0.5659	0.6478	0.8938	1.1797
Min	0.1310	0.0994	0.1352	0.1825
Max	2.2682	2.4959	7.3860	3.4374
S. D.	0.4209	0.5188	0.9039	0.7866

The 3rd stage SBM was applied to this data set and the results are summarized in Table 8.

Table 8: Results of 3rd Stage SBM using the Re-adjusted Data

	Average	Min	Max	S.D.
SBM score	0.9232	0.6812	1	0.0750

Figure 3 compares the efficiency scores of the 1st and the new 3rd stage SBM. The upgrade of the average score from 0.7188 (1st stage) to 0.9232 (New 3rd stage) reflects the effects of environmental factors and statistical noises identified in the 2nd stage SFA. Compared with the Figure 2 which resulted from the adjustments using max, the new 3rd stage results are more acceptable for efficiency evaluations.

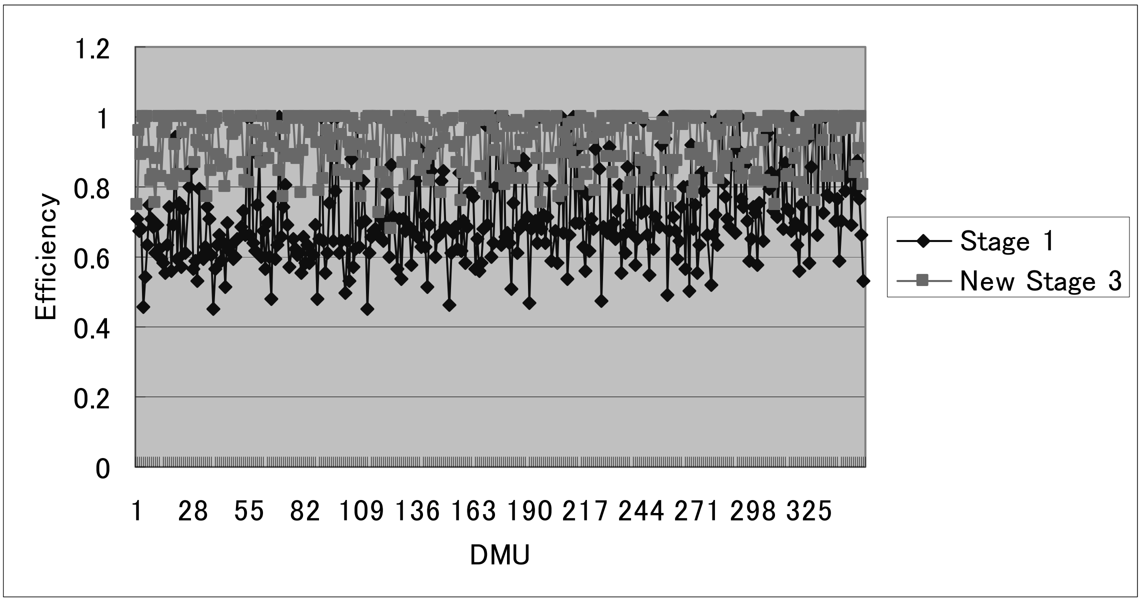

Figure 3: Comparisons of Stage 1 and New Stage 3 Scores

6. Concluding Remarks

In the DEA studies, many authors have tried to identify the true managerial efficiency after accounting for the operational environment effects and statistical noises on the data. The three stage approach proposed by Fried et al. (2002) is a remarkable advance on this line. They combined DEA with SFA in the manner that the slacks obtained in the 1st stage DEA was regressed by means of the environmental effects, statistical noises and managerial efficiency in the data. Then they adjust the original input data using the regression results.

In this paper, we have pointed out shortcomings in their data adjustment and proposed a new adjustment scheme of SFA results for use in DEA. This scheme was applied to U.S. and Japan electric utilities and proved its superiority over the traditional one. Combining non-parametric DEA with parametric SFA may arouse several fundamental problems. The data adjustment problem is an important issue among them. We hope our method serves as a stepping stone to the final resolution.

References

Fried HO, Lovell CAK, Schmidt SS, Yaisawarng S. Accounting for environmental effects and statistical noise in data envelopment analysis. Journal of Productivity Analysis 2002; 17: 157-74.

Avkiran NK, Rowlands T. How to better identify the true managerial performance: State of the art using DEA, OMEGA 2006 (forthcoming).

Tone K. A slacks-based measure of efficiency in data envelopment analysis. European Journal of Operational Research 2001; 130: 498-509.

Banker RD, Charnes A, Cooper WW, Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 1984, 30: 1078-1092.