Kittelsen, Sverre A. C.

**Working Paper**

# Monte Carlo simulations of DEA efficiency measures and hypothesis tests

Memorandum, No. 1999,09

**Provided in Cooperation with:**
Department of Economics, University of Oslo

*Suggested Citation:* Kittelsen, Sverre A. C. (1999) : Monte Carlo simulations of DEA efficiency measures and hypothesis tests, Memorandum, No. 1999,09, University of Oslo, Department of Economics, Oslo

This Version is available at:
https://hdl.handle.net/10419/63113

# MEMORANDUM

**No 09/99**

Monte Carlo Simulations of DEA Efficiency Measures and Hypothesis Tests

***By***
***Sverre A.C. Kittelsen***

ISSN: 0801-1117

Department of Economics
University of Oslo

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

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

In co-operation with
**The Frisch Centre for Economic Research**

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

List of the last 10 Memoranda:

| | |
|---|---|
| No 27 | Wage and employment Effects of Payroll Taxes and Investment Subsidies. 27 p. |
| No 28 | By Hilde Bojer: Equivalence Scales and Intra-household Distribution. 12 p. |
| No 01 | By Gabriela Mundaca and Jon Strand: Speculative attacks in the exchange market with a band policy: A sequential game analysis. 45 p. |
| No 02 | By Pedro P.Barros and Tore Nilssen: Industrial Policy and Firm Heterogeneity. 18 p. |
| No 03 | By Steinar Strøm: The Economics of Screening Programs. 8 p. |
| No 04 | By Kai Leitemo and Øistein Røisland: Choosing a Monetary Policy Regime: Effects on the Traded and Non-Traded Sectors. 39 p. |
| No 05 | by Eivind Bjøntegård: The Composite Mean Regression as a Tool in Production Studies. 9 p. |
| No 06 | By Tone Ognedal: Should the Standard of Evidence be reduced for White Collar Crime? 27 p. |
| No 07 | Knut Røed and Tao Zhang: What Hides Behind the Rate of Unemployment? Micro Evidence from Norway. 39 p. |
| No 08 | Geir B. Asheim, Wolfgang Buchholz and Bertil Tungodden: Justifying Sustainability. 24 p. |

# Monte Carlo simulations of DEA efficiency measures and hypothesis tests[†]

**By**

**Sverre A.C. Kittelsen**[††]

**Frisch Centre, Oslo.**

**Abstract**

The statistical properties of the efficiency estimators based on Data Envelopment Analysis (DEA) are largely unknown. Recent work by Simar et al. and Banker has shown the consistency of the DEA estimators under specific assumptions, and Banker proposes asymptotic tests of whether two subsamples have the same efficiency distribution. There are difficulties arising from bias in small samples and lack of independence in nested models. This paper suggest no new tests, but presents results on bias in simulations of nested small sample DEA models, and examines the approximating powers of suggested tests under various specifications of scale and omitted variables.

**JEL Classification: D24, C44, C15**

**Keywords: Data Envelopment Analysis, Monte Carlo simulations, Hypothesis tests, Non-parametric efficiency estimation**

---

[†] I thank Finn R. Førsund, Leopold Simar, Arne Torgersen, Tore Schweder, Rajiv Banker and Shawna Grosskopf, as well as participants at various presentations for valuable comments on previous versions. This paper is part of the project "The saving potential in the public sector" financed by the Research council of Norway (NFR).

[††] Frisch Centre, Gaustadalleen 21, N-0371 Oslo, Norway. Tel:+47-22958815, Fax:+47-22958825, Email:s.a.c.kittelsen@frisch.uio.no

## *1. Introduction*

In the literature on the measurement of technical efficiency of production the non-parametric deterministic frontier method of characterising production technology known as Data Envelopment Analysis (DEA) has gained popularity. Most studies report DEA efficiency without any evaluation of the model specification or of the significance of the estimates, although there are exceptions. Rank order tests have been used to compare efficiency in different groups or subsamples. In contrast, however, to the parametric cost- and production function approaches there have been few attempts at constructing statistical tests of the model specification. Some authors have extended the sensitivity analysis of operations research to DEA, while others have been concerned with the theoretical conditions for one specification to be equivalent to another.

The assumption of no measurement error in the variables implies that the DEA technique is deterministic since each observed point is assumed feasible, but does not imply that the efficiency measures that are calculated are without error. Since these measures are calculated from a finite sample of observations they are liable to sampling error. While it has previously been uncommon to refer to the DEA measures as estimators, it is increasingly recognised that these measures have statistical properties that deserve attention (see e.g. Simar, 1996).

The extent of bias is of interest in order to get better estimates of the level of efficiency and the position of the frontier. Hypothesis tests are necessary to asses alternative model specifications such as variable or constant returns to scale, omitted variables, permissible aggregation and convexity, and also to compare the efficiency of different unit subsets such as privately vs. publicly owned firms. While tests such as the Man-Whitney rank-order tests have been used for subset comparisons[1], the assumptions underlying most tests are not fulfilled when testing model specification since such models generally will be nested.

---

[1] See e.g. Valdmanis (1992) or Magnussen (1996).

In recent developments, Banker (1993) has proven the consistency of the DEA estimators under specific assumptions and suggested statistical tests of model specification, while Korostelev, Simar and Tsybakov (1995a, 1995b) have been concerned with the rate of convergence of non-parametric frontier estimators. Kneip, Park and Simar (1996) extends these results to a more general model. Simar and Wilson (1995) suggests a bootstrap method for estimating the bias and confidence intervals of efficiency estimates and Simar and Wilson (1997) extend this to suggest a test of returns to scale[2]. Even though this approach seems feasible, it would be advantageous if simpler techniques were available.

So far, no tests have been suggested that can be shown analytically to able to discriminate between competing models, especially in small samples. While suggesting some of the tests analysed below, Banker (1993, p.1272) warns that "... the results should be interpreted very cautiously, at least until systematic evidence is obtained from Monte Carlo experimentation with finite samples of varying sizes". Banker (1996) has summarised a series of Monte Carlo runs, some of which are similar to the ones in the present article, concluding that his tests outperform COLS tests and the Welch means tests in many situations. Although results are promising, Simar (1996, p.181) points out that the "...number of replications are definitely too small to draw any conclusions...". In Banker's studies, there are 10-30 samples in each trial, while the simulations reported below are based on 1000 samples in each trial. Furthermore, Banker generally only provides one estimate of power (lack of Type II errors) in each experiment, while this paper plots power curves based on five or ten such estimates.

In addition to the major undertaking of providing enough simulations to draw clear conclusions about the usefulness of the suggested approximate hypothesis tests, this paper aims at providing some simulation evidence on the bias of the DEA efficiency estimators. After a brief review of the efficiency measurement literature, the subsections of section 3 describe the data generating process, the DEA efficiency estimators, and the suggested tests, and these are followed by three result subsections describing the basic

---

[2] See Grosskopf (1996) for a survey of statistical inference in nonparametric models.

results for bias and the returns to scale tests, variations on the basic assumptions, and a section on testing for variable inclusion. The paper does not propose new tests or bias correction methods; one needs first a proper evaluation of those allready sugggested. Some substantive findings do however give grounds for conclusions in empirical work. The simulations show that bias is important, and that the suggested tests are all of incorrect size because of this bias and the lack of independence in nested models. Some of the tests do, nevertheless, pass the size criterium and retain considerable power in most simulations.

## *2. Efficiency measurement*

The idea of measuring technical efficiency by a radial measure representing the proportional input reduction possible for an observed unit while staying in the production possibility set stems from Debreu (1951) and Farrell (1957) and has been extended in a series of papers by Färe, Lovell and others[3] . Farrell's specification of the production possibility set as a piecewise linear frontier has also been followed up using linear programming (LP) methods by Charnes, Cooper et al[4]. The decomposition of Farrell's original measure relative to a constant returns to scale (CRS) technology into separate measures of scale efficiency and technical efficiency relative to a variable returns to scale (VRS) technology is due to Førsund & Hjalmarsson (1974) and has been implemented for a piecewise linear technology by Banker, Charnes and Cooper (1984). Their DEA formulation has served as the main model of most recent efficiency studies and is the basic model in this paper.

In parallel with the non-parametric or mathematical programming approach to efficiency measurement, considerable research has been conducted in a parametric tradition originating with Aigner and Chu (1968). Their deterministic approach was to estimate a smooth production frontier with residuals restricted to be non-negative and interpreting these residuals as a measure of inefficiency. Like in the non-parametric models, this

---

[3] E.g. Färe & Lovell (1978) and Färe, Grosskopf & Lovell (1985).

[4] E.g. Charnes, Cooper & Rhodes (1978) who originated the name DEA. For an overview of the literature on DEA see e.g. Seiford & Thrall (1990).

interpretation is vulnerable to measurement errors and model mis-specification. Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) incorporate a stochastic error term in addition to the inefficiency term in a composed error model of a production or cost frontier. To identify these terms separately in cross-section studies explicit assumptions are needed for the functional form of the distribution of each term[5]. With explicit distributions it is possible to construct statistical tests in such models.

The major drawbacks of the non-parametric methods is precisely that they have not had access to tests that are known to have desirable properties in small samples, and in addition have not been able to account for measurement errors. The drawbacks of the parametric stochastic frontier approach are chiefly the structure imposed on the data by the choice of functional forms, both for the frontier and for the separate error distributions, and in the context of production functions the difficulty of modelling multiple-input multiple-output technology[6].

Recently some work has been done on developing stochastic non-parametric frontier models (e.g. Petersen and Olesen, 1995). The methods suggested so far require, however, either panel data or exogenous estimates of vital parameters such as constraint violation probabilities. The aim of the present paper is less ambitious. Firstly, no account is taken of measurement errors. Secondly, the paper suggests no new statistical tests, but scrutinises tests present in the literature that try to take the analysis of model mis-specification into the realm of the widely used non-parametric models. Hopefully, the experimental evidence presented here point to research directions that may result in better tests in the future.

One attraction of the non-parametric frontier methods is that the functional form is perfectly flexible. On the face of it, the models are solveable when the number of

---

[5] Panel studies often replace this with similarly strong assumptions on the time pattern of inefficiency.

[6] For an overview of the stochastic frontier parametric approach see e.g. Bauer (1990) or Greene (1993b). Both approaches have seen an active and extensive literature in recent years, often written by researchers working in both subfields. It is beyond the scope of this paper to give a full discussion of the relative merits of the competing methods, see e.g. Fried, Lovell and Schmidt (1993) and Diewert and Mendoza (1996).

dimensions becomes large, even when parametric methods would exhaust the degrees of freedom. A full set of disaggregated inputs and outputs and the inclusion of all potentially relevant variables does however create problems even in DEA. Firstly, the DEA method will measure as efficient all units that in some senses have extreme values of the variables; in the variable returns to scale (VRS) specification this includes those that have the lowest value of an input or the highest value of an output. These units are measured as efficient by default. A variable that in fact is irrelevant to the analysis could therefore destroy the efficiency measures for some units, even if the average efficiency is not much affected. Secondly, a related phenomena is that inclusion of an extra variable, as will be shown, increases the mean bias in the efficiency estimators.

Thirdly, in common with the problem of multicollinearity in parametric methods, any two variables that are highly correlated and therefore have much of the same information value will tend to destroy the rates of transformation and substitution on the frontier[7]. Any use of these marginal properties such as returns to scale, relative shadow prices or marginal costs will therefore be affected. Finally, on a more practical level, a model could become unmanageable and not easy to understand if the dimensionality is very high.

Tulkens & Vanden Eeckaut (1991) reject the frontier concept altogether replacing it with a concept of dominance. In the context of the Free Disposal Hull (FDH) specification suggested by Deprins, Simar & Tulkens (1984) they take the view that the non-parametric methods can be interpreted as measuring the relative efficiency of the observed units with no reference to an underlying production possibility set. In such a setting the properties of the frontier are by definition of no interest, nor are estimates of bias in measured efficiency since the *relative* efficiency is observed without bias. This still leaves open the question of model misspecification, and the problem of units being efficient by default.

In contrast this paper proceeds on the assumption that both the extent of bias and the properties of the underlying production or cost possibility set are of interest. The basic

---

[7] See Olesen and Petersen (1991, 1996) for a discussion of multicollinearity in the DEA model, and e.g. Koutsoyiannis (1977) in econometric models.

assumption is that there is a possibility set defined not only by technology in a narrow sense, but also by the common constraints given by nature, custom, work practice, government regulations, knowledge and organisational technology, including the set of incentive mechanisms available to owners and management of the units under observations. Another important assumption is that there are variations between these units in the objectives of the agents involved, and perhaps also in some of the constraints facing them. To the extent that the differences in constraints are in some sense unchangeable, these should be included in the model specification. If differences in objectives, the use of incentive mechanisms or other changeable constraints, leads to differences in actual behaviour between units so that some of them are not on the frontier of the possibility set, these units are deemed inefficient. The distribution of inefficiency between firms is therefore not truly random. If we were able to model these behavioural differences we would also have more information on how to eliminate inefficiency. Since the Industrial Organisation literature has not so far come up with models that can be tested empirically, we must instead model inefficiency *as if* it was generated randomly[8].

## *3. The model*

### 3.1 The Data Generating Process

Given a vector $\mathbf{y}$ of K outputs and a vector $\mathbf{x}$ of L inputs the production possibility or technology set is defined by

$$P = \left\{(\mathbf{y},\mathbf{x}) \in \Re_{+}^{K+L} \middle| \mathbf{y} \text{ can be produced from } \mathbf{x}\right\} \qquad (1)$$

which can be equivalently be described by the Shephard (1970) input requirement set

$$L(\mathbf{y}) = \left\{\mathbf{x} \middle| (\mathbf{y},\mathbf{x}) \in P\right\} \qquad (2)$$

[8] See e.g. Førsund and Hjalmarsson (1987).

The border of the input set for $\mathbf{y} \geq 0, \mathbf{y} \neq 0$ is known as the production isoquant, defined by those points from which a proportional reduction in input usage is not possible for a given output level:

$$\partial L(\mathbf{y}) = \left\{ \mathbf{x} \middle| \mathbf{x} \in L(\mathbf{y}), \theta\, \mathbf{x} \notin L(\mathbf{y}), \theta \forall \quad \in [0,1) \right\} \tag{3}$$

The properties of these sets and their output equivalents are extensively discussed in Shephard (1970).

The data generation process used for the simulations below follow assumptions A1 to A4 of Kneip, Park & Simar (1996). These are briefly

A1) that the *n* observations are independently and identically distributed (i.i.d.) random variables on the set *P*,

A2) that the support of the density of outputs $\mathbf{y}$ is compact. Further, by

A3) the input mix has a density function conditional on output levels, and the input vector length has a density conditional on output levels and input mix. This assumption implies that inefficiencies are radially generated and input-oriented. Finally, by

A4) the density of the modulus must be such that one will observe points arbitrarily near the frontier when the number of observations is sufficiently large.

In the simulations below, power curves are generated for the tests under examination. These power curves consists of 5-10 runs of 1000 samples, generated with different true values of a parameter, mainly the elasticity of scale. The null hypothesis is that one of these values is true, e.g. constant returns. In addition to a basic trial A), some central assumption is varied in subsequent trials, e.g. the sample size in trial B), the efficiency level in trial C), the inefficiency distributional form in trial D), the distribution of output in trial E) and the number of inputs in trial F). Finally the G) trial tests for the inclusion of an extra variable rather than the returns to scale.

The assumptions A1)-A4) above are operationalized by specifying a technology set defined by a production function with one output, a single scale parameter and a Cobb-Douglas core:

$$P = \{(y,\mathbf{x}) | F(y,\mathbf{x}) \le 0\}, \quad F(y,\mathbf{x}) = y - \left[\prod_{l=1}^{L} x_l^{\alpha_l}\right]^{\beta}, \sum_{l=1}^{L} \alpha_l = 1. \tag{4}$$

It follows that the frontier of the set is defined by $F(y,\mathbf{x}) = 0$ and the isoquant by

$$\partial L(y) = \left\{ \mathbf{x} \middle| y = \left[\prod_{l=1}^{L} x_l^{\alpha_l}\right]^{\beta}, \sum_{l=1}^{L} \alpha_l = 1 \right\}. \tag{5}$$

The elasticity of scale is equal to the scale parameter $\beta$, but in all cases the null hypothesis will assume constant returns to scale ($\beta = 1$). In the base trial A and most of the others the frontier is a simple function with one input and one output, $y = x^{\beta}$, while in trial F and G there will be multiple inputs. As is usual in Monte Carlo studies, the base case under the null hypothesis is very simple, but any other base case would be more ad hoc, and the variations below point to the direction of the change in results in more realistic settings.

In each of the trials[9] reported in this paper there is one run with the null hypothesis true and 5-10 runs with the null hypothesis false, each run having *s*=1000 samples, each sample $j = 1 \ldots s$ with a different set of observations $N_j$, but the same sample size *n* of i.i.d. generated observations $(\mathbf{y}^{ij}, \mathbf{x}^{ij}) \in P, i \in N_j$, fulfilling assumption A1) above. The sample size *n* is 100 in most trials, but varies in trials B[10].

By A2), the output quantity *y* is generated randomly from a distribution with a common mean and variance

---

[9] For simplicity I omit subscripting the trials.

[10] The simulations of the 168000 samples were carried out partly in GAUSS on an IBM RS 6000 running UNIX, partly in GAUSS on a Pentium 90 PC and partly in a Borland Delphi 3.0 Pascal program calling the XA solver on a Pentium II 233 PC. The latter ran about 5 times as fast as each of the Pentium 90 PC and the RS 6000, while the Pascal/XA combination performed about 60 times as fast as GAUSS. The largest trials (B6) each took 5 hours on Pascal/Pentium II but algorithms could be further optimised for the purpose. The basic trial A has been run on all platforms to check the consistency of results. Random Uniform numbers are drawn using internal 32-bit generators, while algorithms for Normal, Lognormal, Exponential and Gamma distributions are from Press et al. (1989).

$$y^{ij} \sim f(y), \mu_f = 10, \sigma^2_f = 2 \quad (6)$$

where *f* in all trials except E is the normal distribution[11], $y^{ij} \sim$ N(10,2), truncated at 0 and 20 to comply with compactness and non-negativity[12].

Less general than A3), the input mixes are generated independently of output level *y* as proportional to two numbers drawn from the same distribution as *y*,

$$\left(\frac{x_l^{ij}}{x_m^{ij}}\right) = \frac{\tau_l^{ij}}{\tau_m^{ij}}, \tau_l^{ij}, \tau_m^{ij} \sim f(\tau), l, m \in 1, \ldots, L \quad (7)$$

When there is only one input, (7) is, of course, redundant. Together, (5) - (7) determine a unique frontier point on the isoquant, $\left(y^{ij}, \mathbf{x}^{*ij}\right) \in \partial L(y^{ij})$. Fulfilling the second part of assumption A3), the actual observed values are generated by multiplying the input quantities by a multiplicative inefficiency term for each observation

$$\mathbf{x}^{ij} = \left(1 + \gamma_{ij}\right)\mathbf{x}^{*ij}, \quad \gamma_{ij} \sim g(\gamma) \quad (8)$$

where the inefficiency term $\gamma$ is generated randomly from a one-sided distribution that is usually halfnormal $\gamma_{ij} \sim |N(0,0.25)|$, but where the inefficiency level varies in trial C, and the functional form varies in trial D. The trials are restricted to inefficiency distributions $g(\gamma)$ that fulfil assumption A4) and have a positive density arbitrarily close to the frontier ($\gamma \to 0$).

Figure 1 shows the generated observations in one typical sample with *n*=100. In the output direction the observations are normally distributed, while the inputs are halfnormally distributed away from the frontier which represents efficient input quantities proportional with the output quantities.

---

[11] See Johnson & Kotz (1970a ,1970b) for a full account of the properties of the distributions used.

[12] This had no practical consequence, since several million draws were made before these bounds were breached the first time.

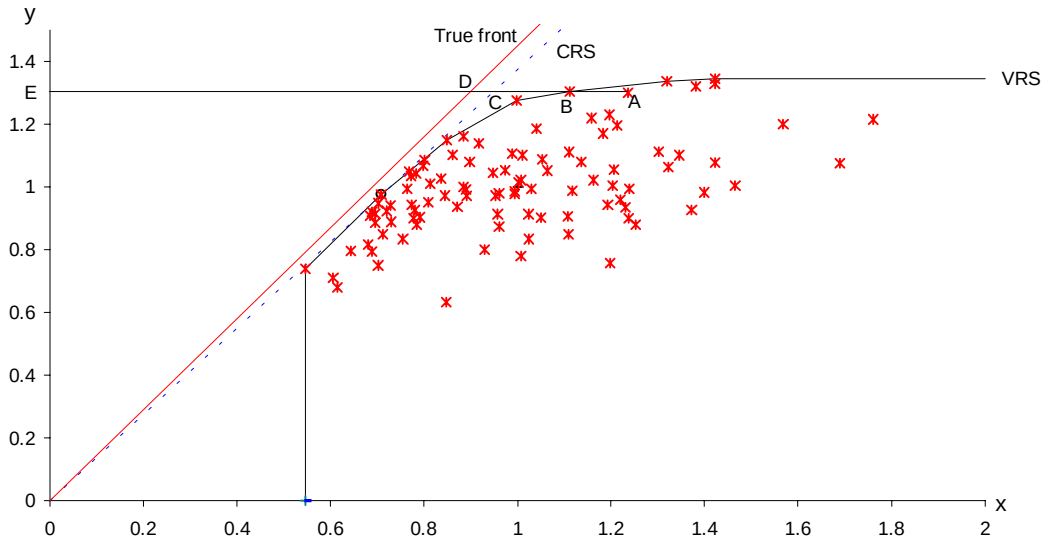



*Figure 1: Simulated data, n=100, g~|N(0,0.5)|, $\beta = 1$, with x and y normalised around their mean. True efficiency for observation A is $E_{Aj} = \mathrm{DE}/\mathrm{AE}$, CRS estimate $\hat{E}_{Aj}^{CRS} = \mathrm{CE}/\mathrm{AE}$, VRS estimate $\hat{E}_{Aj}^{VRS} = \mathrm{BE}/\mathrm{AE}$. The distance between the true frontier and the CRS front is exaggerated.*

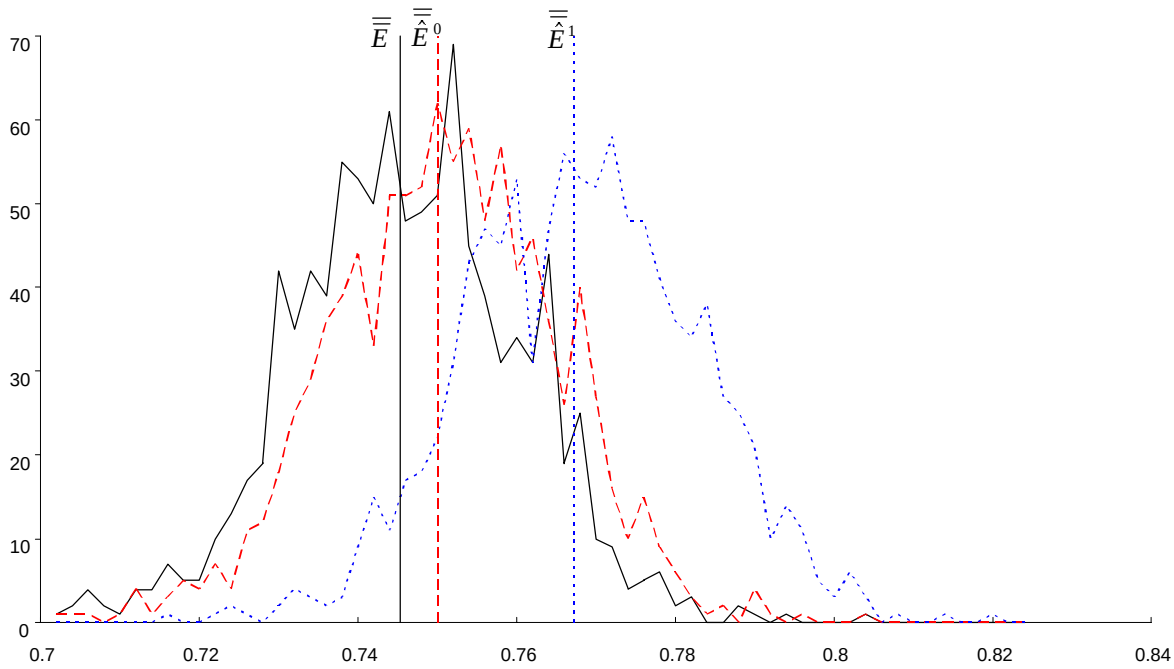



*Figure 2: Frequency distribution in intervals of 0.002 of sample means of true efficiency and estimates in basic trial A, n=100, g~|N(0,0.5)|, with true null hypothesis $\beta = 1$. The solid line represents the mean and sampling distribution of the true mean efficiency $\overline{E}_j^0$, the dashed lines are mean and sampling distribution of the CRS mean estimated efficiency $\overline{\hat{E}}_j^0$, and the dotted lines are mean and sampling distribution of the VRS mean estimated efficiency $\overline{\hat{E}}_j^1$.*

### 3.2 The DEA Efficiency Estimators

Farrell (1957) technical input efficiency can be defined by

$$E(\mathbf{y},\mathbf{x},P) = \text{Min}_{\theta}\left\{\theta|(\mathbf{y},\theta\mathbf{x}) \in P\right\} \tag{9}$$

which, for a feasible point $(\mathbf{y},\mathbf{x}) \in P$ is a number in the interval (0,1] corresponding to the proportional scaling of all inputs necessary to bring the observation to the frontier (isoquant). As noted by e.g. Färe & Lovell (1978) this is the reciprocal of the definition of the Shephard (1970) input distance function, and results could equally well have been represented by this measure. Among the properties of *E(.)* is the homogeneity of degree -1 in inputs, and that it provides an equivalent characterisation of technology since the efficiency measure for a point is 1 if and only if the point is on the isoquant, $\mathbf{x} \in \partial L(\mathbf{y}) \Leftrightarrow E(\mathbf{y},\mathbf{x},P) = 1$ (Shephard, 1970, 67-68).

The efficiency can be calculated relative to the true technology if known, which in our case is

$$E_{ij} \equiv E(\mathbf{y}^{ij},\mathbf{x}^{ij},P) = E(\mathbf{y}^{ij},(1+\gamma_{ij})\mathbf{x}^{*ij},P) = \frac{1}{(1+\gamma_{ij})}E(\mathbf{y}^{ij},\mathbf{x}^{*ij},P) = \frac{1}{(1+\gamma_{ij})} \tag{10}$$

where the last two equalities follows from the homogeneity of the efficiency measure and the fact that the point $\left(y^{ij}, \mathbf{x}^{*ij}\right)$ is on the isoquant. The efficiency measures can also be calculated relative to an estimate of the technology such as the DEA variable returns to scale estimate from sample *j*

$$\hat{P}_j^{\text{VRS}} = \left\{ \mathbf{Y}_j \lambda \geq y, \mathbf{x} \geq \mathbf{X}_j \lambda, \lambda \geq 0, \sum_{i \in N_j} \lambda_i = 1 \right\} \forall j \in 1 \ldots s \tag{11}$$

where $\mathbf{Y}_j, \mathbf{X}_j$ are the vectors or matrices of observed outputs and inputs in sample *j* and $\lambda$ is a vector of reference weights. This corresponds to the formulation in Banker, Charnes & Cooper (1984), and is the minimum extrapolation estimator of the technology satisfying convexity, free disposability of inputs and outputs and feasibility of observed units (Banker, 1993). Adding a homogeneity requirement gives the DEA constant returns to scale estimator of technology

$$\hat{P}_j^{\text{CRS}} = \left\{ (y, \mathbf{x}) \middle| (\Lambda y, \Lambda \mathbf{x}) \in \hat{P}_j^{\text{VRS}}, \Lambda > 0 \right\} = \left\{ \mathbf{Y}_j \lambda \geq y, \mathbf{x} \geq \mathbf{X}_j \lambda, \lambda \geq 0 \right\} \forall j \in 1 \ldots s \tag{12}$$

where the removal of the restriction that referencing weights add to unity corresponds to the formulation in Charnes et al. (1985).

In each trial DEA efficiency estimates $\hat{E}_{ij}^k$ are calculated under a null hypothesis (*k=0*) and under an alternative hypothesis (*k=1*). Except in last trial G, the null hypothesis is that the true technology exhibits constant returns to scale, and the alternative hypothesis is one of variable returns to scale. One can define a shorthand for the estimated efficiencies under the null and alternate hypothesis as

$$\hat{E}_{ij}^0 = \hat{E}_{ij}^{CRS} = E(\mathbf{y}^{ij}, \mathbf{x}^{ij}, \hat{P}_j^{CRS}), \; \hat{E}_{ij}^1 = \hat{E}_{ij}^{VRS} = E(\mathbf{y}^{ij}, \mathbf{x}^{ij}, \hat{P}_j^{VRS}) \tag{13}$$

Only input saving efficiency estimates are calculated, although in CRS the input and output estimates will be the same. Figure 1 shows graphically the CRS and VRS efficiency estimates for a unit A. In the figure, the number of VRS reference points is in fact eight (some very close together), each being a vertex of the VRS frontier. For CRS in the figure there is only one referencing observation, as will generally be the case with one input and one output.

In reporting the results of the simulations, the arithmetic mean of efficiencies for a set of observations are subscripted by their common index, i.e.

$$\bar{\hat{E}}_j^k = \sum_{i \in N_j} \hat{E}_{ij}^k \Big/ n, \quad \bar{\bar{\hat{E}}}^k = \sum_{j=1}^{s} \bar{\hat{E}}_j^k \Big/ s, \quad k \in \{0,1,\cdot\} \tag{14}$$

including the dot $\cdot$ to indicate no $k$ superscript index or hat for the average true generated efficiency, and similarly for other measures such as the estimates of inefficiency terms and its averages

$$\hat{\gamma}_{ij}^k = \frac{1}{\hat{E}_{ij}^k} - 1, \quad \bar{\hat{\gamma}}_j^k = \sum_{i \in N_j} \hat{\gamma}_{ij}^k \Big/ n, \quad \bar{\bar{\hat{\gamma}}}^k = \sum_{j=1}^{s} \bar{\hat{\gamma}}_j^k \Big/ s, \quad k \in \{0,1,\cdot\} \tag{15}$$

### 3.3 The Bias

Korostelev, Simar & Tsybakov (1995a) show that when the true frontier is nonconvex and the inefficiency is uniformly distributed over an interval, although the FDH estimator is a maximum likelihood estimator, the rate of convergence is slow. In Korostelev, Simar & Tsybakov (1995b) they extend the results to a convex technology where the DEA estimator is a maximum likelihood estimator. In this case they find a rate of convergence higher than in the FDH case, but the rate is still decreasing in the number of dimensions (number of inputs plus number of outputs).

Banker (1993) also proves that the DEA output estimator is a maximum likelihood estimator for a convex technology in a model with an additive inefficiency term that is distributed with mode at 0 (i.e. at the frontier). He further proves that the DEA estimator is consistent (i.e. asymptotically unbiased and with a vanishing variance) as long as the cumulative density function is positive for all inefficiencies greater than 0, even without mode at this value. Kneip, Park and Simar (1996) extend these results to a more general multi-input multi-output, proving consistency if the assumptions A1-A4 above are satisfied. They also investigate the rate of convergence, which they find depend on the smoothness of the frontier and deteriorate for higher dimensionality (i.e. number of outputs plus inputs). Gijbels et al. (1996) derive the asymptotic distribution of the DEA frontier estimator, and suggest a bias-corrected estimator, but only for the one-input one-output case. It is shown there that the bias is much more important than the standard deviation for the mean square error of the estimates.

The problem of bias in DEA follows from the fact that the probability of observing a truly efficient unit in a sample is less than one, and for most of the commonly specified distributional forms in fact zero even if one will observe a unit arbitrarily close to the frontier as the sample size increases. It will (almost) always be possible to be more efficient than the most efficient of the observed units. In figure 1 one can see that the CRS frontier lies to the right of the true frontier, even though CRS is the correct model specification in this case. Maintaining the assumption of no measurement error, units will generally be estimated as more efficient than they actually are if the model is correctly specified. By construction, if the model k is correctly specified, the error $B_{ij}^k$ of the estimate will be greater or equal to zero[13]:

$$\hat{E}_{ij}^k \geq E_{ij},\ B_{ij}^k \equiv \hat{E}_{ij}^k - E_{ij} \geq 0 \qquad (16)$$

[13] Allthough intuitive, the proof of this and subsequent statements on ranked and nested models require some tedious definitions and manipulations, and are therefore in an appendix. Diewert and Mendoza (1996) has an informal discussion of some of these results, using the term Le Chatelier Principles.

Furthermore, as is commonly observed in the literature (e.g. Diewert and Mendoza, 1996), the VRS estimates will show greater or equal efficiency than the CRS estimates. In figure 1 the VRS frontier is on or to the right of the CRS frontier. Färe & Primont (1987) show that if an aggregated model is nested within a disaggregated model, the measured $\hat{E}^{VRS}$ in the disaggregated model will be greater or equal than the measured $\hat{E}^{VRS}$ in the aggregated model. These relationships follow from the principle that the optimal value of a minimised variable can never become lower if a restriction is added to the optimisation problem. This principle also implies that a model in which a variable is included will give an efficiency estimate that is as least as high as in the same model with the variable omitted, since including a variable is the same as adding an extra restriction to the optimisation problem in (9) when *P* is replaced with (11) or (12). In general therefore if model 0 is nested within model 1, in the sense that model 0 can be obtained from model 1 as a special case:

$$\hat{E}_{ij}^{1} \geq \hat{E}_{ij}^{0}, \quad B_{ij}^{1} > B_{ij}^{0} \tag{17}$$

if model 1 assumes the feasibility of all observations (see proposition 1 in the appendix).

Since the true model is equal to or nested within the null hypothesis model in the true simulations in this paper, and since (16) and (17) holds for each individual observation, a complete ranking exists also for the average efficiencies in each sample generated under the true null hypothesis:

$$\overline{\hat{E}}_{j}^{1} \geq \overline{\hat{E}}_{j}^{0} \geq \overline{E}_{j}, \quad \overline{B}_{j}^{1} \geq \overline{B}_{j}^{0} \tag{18}$$

The average efficiency estimates $\overline{\hat{E}}_{j}^{k}$ in each sample are also the sample estimate of true mean efficiency. In each trial there are 1000 samples, and averaging over these gives the Monte Carlo estimate of the expected value of these mean efficiency estimates $\overline{\overline{\hat{E}}}^{k}$ and their bias $\overline{\overline{B}}^{k}$. These obey the same ranking as in (18).

Thus there is not only bias in estimators of each unit's efficiency , but also in the estimators of average efficiency, and furthermore the bias is at least as great in more restricted models, i.e. with increasing dimensionally. If the null hypothesis is false, the estimates are no longer nescessarily larger than the true efficiencies, but the ranking of the two estimates remain valid[14].

### 3.4 The Tests

On the basis of his consistency results basis, Banker (1993) suggests asymptotic tests for comparing the efficiencies of to subsets of observations, $N^a, N^b$, and a null hypothesis that the single parameter of the inefficiency distributions are equal. If the inefficiency estimates $\hat{\gamma}_{ij}$ are independently and identically distributed (i.i.d.). and the underlying true inefficiency distribution is halfnormal, the statistic

$$F_j^{\,H} = \frac{\sum_{i\in N^a}\left(\hat{\gamma}_{ij}^{\,a}\right)^2 \Big/ n^a}{\sum_{i\in N^b}\left(\hat{\gamma}_{ij}^{\,b}\right)^2 \Big/ n^b} \tag{19}$$

is asymptotically F-distributed with ( $n^a, n^b$ ) degrees of freedom.

If the inefficiency estimates are i.i.d. and the underlying distribution is exponential, the statistic

$$F_j^{\,E} = \frac{\sum_{i\in N^a}\hat{\gamma}_{ij}^{\,a} \Big/ n^a}{\sum_{i\in N^b}\hat{\gamma}_{ij}^{\,b} \Big/ n^b} \tag{20}$$

---

[14] In fact, when $\beta > 1$ the true technology defined by (4) is not convex, and does not therefore fulfil the assumptions underlying the alternate hypothesis estimate $\hat{P}_j^{VRS}$. A parallel run of trial A below with a local linearisation reveals that this has only negligible effect on the bias and power results for this range of $\beta$. Even in the extreme case of a scale elasticity of 1.5 there are only 4% of the observations with a negative bias $B_{ij}^1$. Since the interest lies around the true null, the much simpler formulation in (4) is chosen.

is asymptotically F-distributed with $(2n^a, 2n^b)$ degrees of freedom.

If no parametric assumptions are maintained about the inefficiency distributions, Banker further suggests using a Kolmogorov-Smirnov type of nonparametric test of the equality of two distributions. Applied to the distributions of i.i.d. efficiency estimates $\hat{E}_{ij}^a, \hat{E}_{ij}^b$, and denoting the estimated cumulative distribution function of these as $S_j^a(E), S_j^b(E)$, the statistic

$$D_j^+ = \text{Max}_E \left\{ S_j^a(E) - S_j^b(E) \right\} \tag{21}$$

is asymptotically distributed with a rejection probability of

$$\Pr\left( D_j^+ > \left( \frac{n^a n^b}{n^a + n^b} \right)^{-\frac{1}{2}} z \right) = e^{-2z^2}, \quad z > 0 \tag{22}$$

which makes it applicable for testing one-sided hypotheses (Johnson & Kotz, 1970b).

For comparison, the simple T-statistic[15] for the equality of group means is reported:

$$T_j = \frac{\text{Mean}_{i \in N^b}\left(\hat{E}_{ij}^b\right) - \text{Mean}_{i \in N^a}\left(\hat{E}_{ij}^a\right)}{\sqrt{\frac{n^b \text{ Var}_{i \in N^b}(\hat{E}_{ij}^b) + n^a \text{ Var}_{i \in N^a}(\hat{E}_{ij}^a)}{n^b + n^a - 2}\left[\frac{1}{n^b} + \frac{1}{n^a}\right]}} \tag{23}$$

which, if sample means are i.i.d. normal, is T-distributed with $n^a + n^b - 2$ degrees of freedom. By the central limit theorem the sample means will be approximately normal unless sample size is very small. The expression greatly simplifies when $n^a = n^b$ as is the case in the reported simulations. Finally the T-test for paired observations is also reported:

---

[15] See e.g. Bhattacharyya & Johnson (1977, p.295-296).

$$T_j^P = \frac{\text{Mean}_{i\in N}(\hat{E}_{ij}^b - \hat{E}_{ij}^a)}{\sqrt{\frac{\text{Var}_{i\in N}(\hat{E}_{ij}^b - \hat{E}_{ij}^a)}{n-1}}} \tag{24}$$

which, if mean difference in efficiency is normal with zero expected value, is T-distributed with *n*-1 degrees of freedom.

Banker (1993) investigates the asymptotic properties of the tests (19)-(20) for disjoint groups, but does not consider the usefulness of the tests for nested models, and leaves open their approximating powers in small samples. In the Monte Carlo studies summarised in Banker (1996), however, these tests are explicitly applied to nested models, where they are formulated with full sample estimates both for *a* and *b* above. The *full sample tests* appear by substituting into (19)-(24) for each sample *j*

$$\begin{aligned} &\hat{E}_{ji}^a = \hat{E}_{ij}^0, \hat{\gamma}_{ij}^a = \hat{\gamma}_{ij}^0, \\ &\hat{E}_{ij}^b = \hat{E}_{ij}^1, \hat{\gamma}_{ij}^b = \hat{\gamma}_{ij}^1, \\ &N^a = N^b = N_j, n^a = n^b = n \end{aligned} \tag{25}$$

Since DEA estimators are generally not independently and identically distributed (i.i.d.), there are theoretical problems with all five tests. The first four assume independence between all observations of $E$ (or $\gamma$), which is obviously not fulfilled for nested models if all observations are included in the calculations under both the null and the alternative hypothesis. There will then be a strong dependence resulting from measuring efficiency for the same observations under both models. This strong dependence will not be present however, if the sample is split into two equal size sets $N_{j0} \cup N_{j1} = N_j, N_{j0} \cap N_{j1} = \varnothing$ so that e.g. half the observations are used when calculating the null hypothesis variables, and the other half are used when calculating the alternative hypothesis variables. The full sample is still used as reference sets in the calculation of the technology estimates in (11) and (12). In the simulations, *split sample tests* are calculated by substituting in (19)-(23)

$$\begin{aligned}
&\hat{E}_{ij}^{a} = \hat{E}_{uj}^{0},\ \hat{\gamma}_{ij}^{a} = \hat{\gamma}_{uj}^{0},\ u \in N_{j0} \\
&\hat{E}_{ij}^{b} = \hat{E}_{vj}^{1},\ \hat{\gamma}_{ij}^{b} = \hat{\gamma}_{vj}^{1},\ v \in N_{j1} \\
&N^{a} = N_{j0}, N^{b} = N_{j1}, n^{a} = n^{b} = n/2
\end{aligned} \tag{26}$$

and superscripting the estimate means and their biases with *S* e.g. $\overline{\hat{E}}_{j}^{Sk} = \text{Mean}_{i \in N_{jk}}\left(\hat{E}_{ij}^{k}\right)$. Splitting the sample can not be done for the paired T-test in (24), because there would be no pairing of estimates.

Even if one has removed the strong dependence by splitting the sample, there is a weak dependence between estimated efficiencies, since they can be calculated relative to the same referencing observations. This weak dependence will diminish as the sample size increases, but to avoid it one can partition the observations also in the technology estimates. Let the technology estimate for hypothesis *k* be $\hat{P}_{j}^{Rk}$ calculated from (11) or (12), but so that only the observations in each subset enter the matrices $\mathbf{Y}_{j}^{k}, \mathbf{X}_{j}^{k}$. Then in the simulations, *separate reference set tests* are calculated by substituting in (19)-(23)

$$\begin{aligned}
&\hat{E}_{ij}^{a} = \hat{E}_{uj}^{R0} = E_{uj}(y^{uj}, x^{uj}, \hat{P}_{j}^{R0}),\ \hat{\gamma}_{ij}^{a} = 1/\hat{E}_{uj}^{R0} - 1,\ u \in N_{j0} \\
&\hat{E}_{ij}^{b} = \hat{E}_{vj}^{R1} = E_{vj}(y^{vj}, x^{vj}, \hat{P}_{j}^{R1}),\ \hat{\gamma}_{ij}^{b} = 1/\hat{E}_{vj}^{R1} - 1,\ v \in N_{j0} \\
&N^{a} = N_{j0}, N^{b} = N_{j1}, n^{a} = n^{b} = n/2
\end{aligned} \tag{27}$$

again superscripting averages with *R*.

In addition to their dependence, the estimators of nested models are not identically distributed since, as is shown in (17) above, by adding an extra restriction the model specification itself makes the bias of a model greater than the bias of a model that is nested within it. If the samples are split, this difference holds only in expected values, and not necessarily for all samples. As both estimators are consistent, this effect should diminish as sample size increases. The simulations in the next section aim to shed light on how serious the bias and dependence affect the applicability of the various tests in finite samples.

Finally, all but the paired T-test in (24) are based on comparisons of the magnitude of the sample average of the estimates or their squares, rather than on some measure of the differences between the individual unit estimates. They do not use the information contained in the paired nature of the estimates, and will therefore generally not have the full potential power.

The null hypothesis in each of the simulations below is that the assumptions underlying $P^0$ are true. Since model 0 is nested in model 1, this latter could equally well describe technology when the null is true, implying $P = P^1 = P^0$. Even if the null is not true, it is always assumed that $P = P^1$. Since we know from the nested character of the models and proposition 2 of the appendix that $P^1 \subseteq P^0$, the null and alternate hypothesis can be formalised as:

$$H_0: P = P^1 = P^0, \quad H_1: P = P^1 \subset P^0 \tag{28}$$

This is equivalent to equal true efficiencies $H_0: E_{ij}^1 = E_{ij}^0, \quad H_1: E_{ij}^1 > E_{ij}^0$ for all observations i in the sample j, and the tests are based on comparisons of the estimates of these efficiencies. This implies one-sided tests where the null/hypothesis is rejected if the test statistic exceeds the critical value of the theoretical distribution. The rejection rate $r_t(\cdot) = \Pr\left(t > t^*\right)$ for some statistic *t* and critical value $t^*$ is generally a function of the true characteristics of the technology, which in the simulations is manipulated by a parameter.

Two types of error can occur by this procedure, rejection of a true null hypothesis (Type I error), or the non-rejection of a false null hypothesis (Type II error). The *size* (significance level) of a test is defined as the rejection probability if the null hypothesis is true, $r_t$(true null)=Pr(Type I error), and the *power* of the test defined as rejection probability when it is false, $r_t$(false null)=1-Pr(Type II error). While much of the

**Table 1. Results from trial A. Mean efficiencies, estimates and bias with different scale parameters.**

Common conditions: Number of samples s=1000, Sample size n=100, One input/one output, Normal distribution of output $y \sim N(10,2)$. Halfnormal distributions of inefficiencies $\gamma \sim |N(0,0.25)|$ implying E(γ)=0.399 and E(*E*)=0.745.

| Scale parameter $\beta$ | 0.6 | 0.8 | 1 | 1.2 | 1.4 |
|---|---|---|---|---|---|
| True generated variables | | | | | |
| $\overline{\overline{E}} = \text{Mean}_j(\overline{E}_j)$ | 0.7450 | 0.7448 | 0.7452 | 0.7443 | 0.7451 |
| $\text{SD}_j(\overline{E}_j)$ | (0.0141) | (0.0149) | (0.0152) | (0.0145) | (0.0145) |
| $\text{Mean}_j(\text{SD}_i(E_{ij}))$ | 0.1441 | 0.1436 | 0.1437 | 0.1439 | 0.1439 |
| $\overline{\overline{\gamma}} = \text{Mean}_j(\overline{\gamma}_j)$ | 0.3989 | 0.3992 | 0.3985 | 0.4004 | 0.3988 |
| CRS Estimates | | | | | |
| $\overline{\overline{\hat{E}}}^0 = \text{Mean}_j(\overline{\hat{E}}_j^0)$ | 0.6430 | 0.7197 | 0.7500 | 0.7354 | 0.7208 |
| $\text{SD}_j(\overline{\hat{E}}_j^0)$ | (0.0442) | (0.0227) | (0.0151) | (0.0166) | (0.0198) |
| $\text{Mean}_j(\text{SD}_i(\hat{E}_{ij}^0))$ | 0.1396 | 0.1413 | 0.1446 | 0.1433 | 0.1424 |
| $\overline{\overline{B}}^0 = \text{Mean}_j(\overline{B}_j^0)$ | -0.1021 | -0.0251 | 0.0048 | -0.0089 | -0.0243 |
| $MSE(\overline{\hat{E}}_j^0)$ | 0.0124 | 0.0011 | 0.0003 | 0.0004 | 0.0010 |
| $\text{MSE}(\hat{E}_{ij}^0)$ | 0.0150 | 0.0014 | 0.0000 | 0.0004 | 0.0015 |
| VRS Estimates | | | | | |
| $\hat{E}^1 = \text{Mean}_j(\hat{E}_j^1)$ | 0.7706 | 0.7681 | 0.7668 | 0.7651 | 0.7651 |
| $\text{SD}_j(\hat{E}_j^1)$ | (0.0144) | (0.0145) | (0.0150) | (0.0146) | (0.0148) |
| $\text{Mean}_j(\text{SD}_i(\hat{E}_{ij}^1))$ | 0.1496 | 0.1490 | 0.1488 | 0.1488 | 0.1487 |
| $B^1 = \text{Mean}_j(B_j^1)$ | 0.0256 | 0.0233 | 0.0216 | 0.0208 | 0.0200 |
| $\text{MSE}(\hat{E}_j^1)$ | 0.0009 | 0.0008 | 0.0007 | 0.0006 | 0.0006 |
| $\text{MSE}(\hat{E}_{ij}^1)$ | 0.0021 | 0.0020 | 0.0018 | 0.0017 | 0.0017 |

In the row headers, $\text{Mean}_j(z_j) = \sum_{j=1}^{s} z_j / s$ for a variable *z*, $\text{SD}_j(z_j) = \sqrt{\sum_{j=1}^{s}(z_j - \text{Mean}_j(z_j))^2 / s}$. If the index is *i* it runs over the list of units *1..n* instead. The mean square errors are

$$\text{MSE}(\overline{\hat{E}}_j^k) = \text{Mean}_j\left(\left[\overline{\hat{E}}_j^k - \text{E}(E)\right]^2\right) \text{ and } \text{MSE}(\hat{E}_{ij}^k) = \text{Mean}_j\left(\text{Mean}_i\left(\left[\hat{E}_{ij}^k - E_{ij}\right]^2\right)\right).$$

literature assumes that the size is under the control of the researcher (e.g. Greene, 1993a, p.126), Engle (1984) characterises a test as *best* if it has the maximum power among all tests with size less than or equal to some particular level. In the simulations, only 5% tests will be reported, so the best test will be the among those with $r_t(\text{true null}) \leq 5\%$.

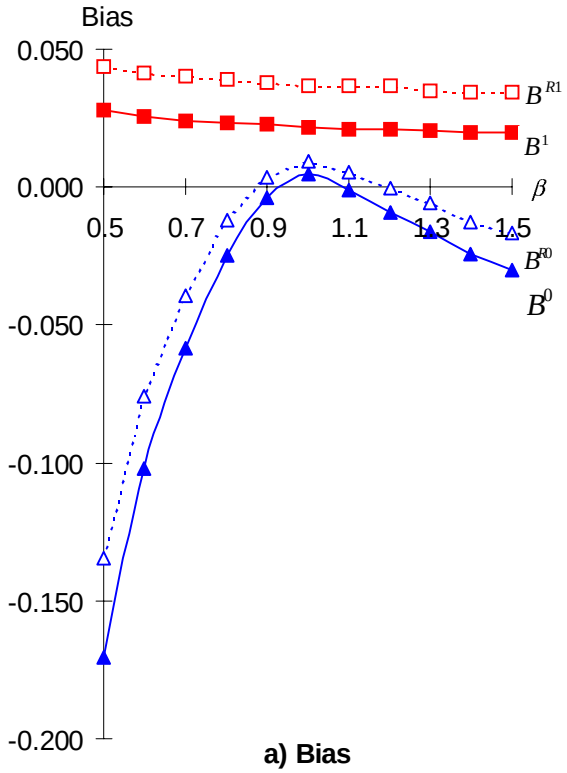
This criteria presuposes that the null hypothesis is the conservative choice since it is the hypothesis which is not rejected if the tests are inconclusive. There are both methodological and economic reasons for choosing the model that is nested within as the null. Firstly this model is simpler and therefore avoids problems of extreme observations being efficient by default and of multicolinearity discussed above. Secondly, it is likely to be statistically more efficient, in that estimators will converge faster. Thirdly, it is in most cases the supposition which is least likely to have undesirable social consequences. In the case of testing for returns to scale, rejecting CRS could mean that there are market imperfections that require costly government intervention. Similarly, when testing for inclusion of irrelevant variable, producers will have a selfinterest in results that show them to be more efficient. Since including more variables is shown above to give efficiency estimates (and bias) that are at least as high, and since producers are normally better organised than consumers, there could well be reasons to counterweigh an inclination to blow up the number of variables.

## *4. The Results*

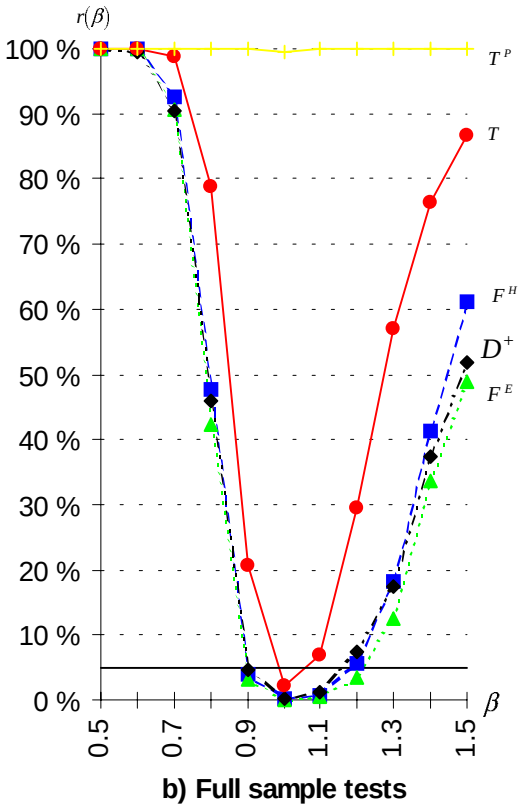
### 4.1 Bias and Testing for Returns to Scale: The Basic Results

*Bias*

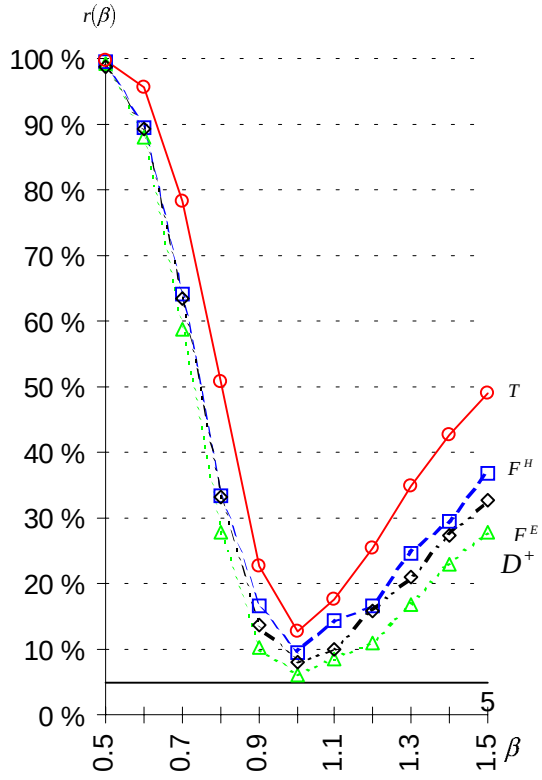
The results of trial A is reported in detail in table 1. This simulation has a sample size of 100 and a halfnormal distribution of inefficiencies with a theoretical mean $\gamma$ of 0.3989, and also serves as a basis of comparison for subsequent trials. The central column represents results when the null hypothesis of constant returns to scale is true ($\beta = 1$). The mean of the 100,000 true efficiencies $\overline{\overline{E}}$ is 0.7452, and this is approximately constant across different values of the scale parameter. Similarly the mean true



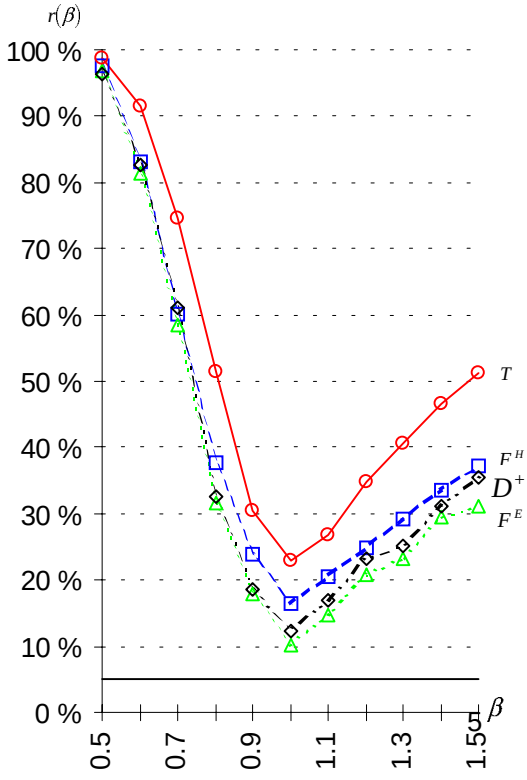

a) Bias




b) Full sample tests




c) Split sample tests




d) Separate reference sets

*Figure 3: Bias and power curves for tests in trial A.*

inefficiency is fairly constant. In the CRS model, mean estimated $\overline{\overline{\hat{E}}}^0$ when the null is true is as expected slightly higher (0.7500) while the VRS estimate $\overline{\overline{\hat{E}}}^1$ is 0.7668.

The frequency distributions of the sample means $\overline{E}_j, \overline{\hat{E}}_j^0, \overline{\hat{E}}_j^1$ are shown in figure 2. The figure shows the magnitude of the bias in each model, which is clearly greater for the VRS estimators than the CRS estimators, in accordance with (18). The figure also shows that the distributions are nearly normal in shape and have approximately the same spread. The Kolmogorov-Smirnov test is not able to reject the normality of any of these distributions[16]. Even though the underlying distribution of $E_{ij}$ is decidedly non-normal, with a sample size of 100 the central limit theorem seems to have some strength.

The second row of table 1 is the standard deviation of the sample mean true efficiencies, which is the Monte Carlo estimate of the standard error of the sample mean. As expected, the mean of the standard deviations in each sample divided by the $\sqrt{n} = 10$ provides a reasonable estimator of this standard error (See e.g. Greene, 1993a, p.91). For the T-test at least, the problem lies in the bias and dependence and not in non-normality of the mean estimates.

The estimator bias as a function of the scale parameter is listed in table 1, and is also shown in panel a) of figure 3. The CRS estimator full sample bias $\overline{\overline{B}}^0$ has a slightly positive maximum when CRS is true at $\beta = 1$, but drops to negative values away from the null both with true increasing and decreasing returns to scale. Ideally, the bias should be

---

[16] The adjusted Kolmogorov-Smirnov statistic $D\sqrt{n}$ for the two-sided test has a value of 0.579, 0.568 and 0.651 for the distributions of $E_j, \hat{E}_j^0, \hat{E}_j^1$ respectively, which compares with a critical value of 0.819 at the 10% significance level.

zero at $\beta = 1$, but otherwise this is satisfactory. The problem lies more in the bias of the VRS estimators, which although stable across scale parameters, are consistently high at 2-2.5%. A common criteria for evaluating estimators is the mean square error (MSE), which is also reported in table 1. The CRS efficiency measures are

| **Table 2. Results from trial A. Correlations and power curves for tests.** | | | | | |
|---|---|---|---|---|---|
| Scale parameter, $\beta$ | 0.6 | 0.8 | 1 | 1.2 | 1.4 |
| Estimate correlation | | | | | |
| $\text{Mean}_j\left(\hat{\rho}_i\left(\hat{E}_{ij}^0, \hat{E}_{ij}^1\right)\right)$ | 0.8586 | 0.9439 | 0.9642 | 0.9604 | 0.9493 |
| $\hat{\rho}_j\left(\overline{\hat{E}}_j^0, \overline{\hat{E}}_j^1\right)$ | 0.3503 | 0.6214 | 0.8776 | 0.8387 | 0.7505 |
| $\hat{\rho}_j\left(\overline{\hat{E}}_j^{S0}, \overline{\hat{E}}_j^{S1}\right)$ | 0.0182 | 0.0652 | -0.0044 | -0.0142 | 0.0296 |
| Split sample bias | | | | | |
| $\overline{\overline{B}}^{S0} = \text{Mean}_j\left(\overline{B}_j^{S0}\right)$ | -0.1024 | -0.0254 | 0.0051 | -0.0088 | -0.0243 |
| $\overline{\overline{B}}^{S1} = \text{Mean}_j\left(\overline{B}_j^{S1}\right)$ | 0.0263 | 0.0236 | 0.0216 | 0.0206 | 0.0197 |
| Separate reference set bias | | | | | |
| $\overline{\overline{B}}^{R0} = \text{Mean}_j\left(\overline{B}_j^{R0}\right)$ | -0.0762 | -0.0120 | 0.0097 | -0.0008 | -0.0126 |
| $\overline{\overline{B}}^{R1} = \text{Mean}_j\left(\overline{B}_j^{R1}\right)$ | 0.0414 | 0.0390 | 0.0365 | 0.0365 | 0.0343 |
| Full sample rejection rates in percent (5% tests) | | | | | |
| $F^H$ | 99.9 | 47.6 | 0.2 | 5.6 | 41.3 |
| $F^E$ | 99.9 | 42.2 | 0.0 | 3.4 | 33.6 |
| $D^+$ | 99.6 | 46.0 | 0.2 | 7.3 | 37.4 |
| $T$ | 100.0 | 78.8 | 2.3 | 29.5 | 76.4 |
| $T^P$ | 100.0 | 100.0 | 99.4 | 100.0 | 100.0 |
| Split sample rejection rates in percent (5% tests) | | | | | |
| $F^H$ | 89.4 | 33.3 | 9.6 | 16.6 | 29.5 |
| $F^E$ | 88.0 | 27.7 | 6.2 | 11.0 | 23.0 |
| $D^+$ | 89.3 | 33.1 | 8.0 | 15.8 | 27.2 |
| $T$ | 95.7 | 50.7 | 12.8 | 25.3 | 42.8 |
| Separate reference set rejection rates in percent (5% tests) | | | | | |
| $F^H$ | 83.0 | 37.6 | 16.5 | 25.0 | 33.6 |
| $F^E$ | 81.4 | 31.6 | 10.1 | 20.7 | 29.4 |
| $D^+$ | 82.7 | 32.7 | 12.4 | 23.1 | 31.2 |
| $T$ | 91.6 | 51.5 | 22.9 | 34.9 | 46.5 |

In the row headers, $\hat{\rho}_j\left(z_j, w_j\right) = \sum_{j=1}^{s}(z_j - \overline{z})(w_j - \overline{w}) \Big/ \sqrt{\left[\sum_{j=1}^{s}(z_j - \overline{z})^2\right]\left[\sum_{j=1}^{s}(w_j - \overline{w})^2\right]}$ for two variable *z, w.* If the index is *i* it runs over the list of units *1..n* instead. The definition of mean is given in table 1. The grey marks the test that is best by the criteria of size less than 5% and maximum power.

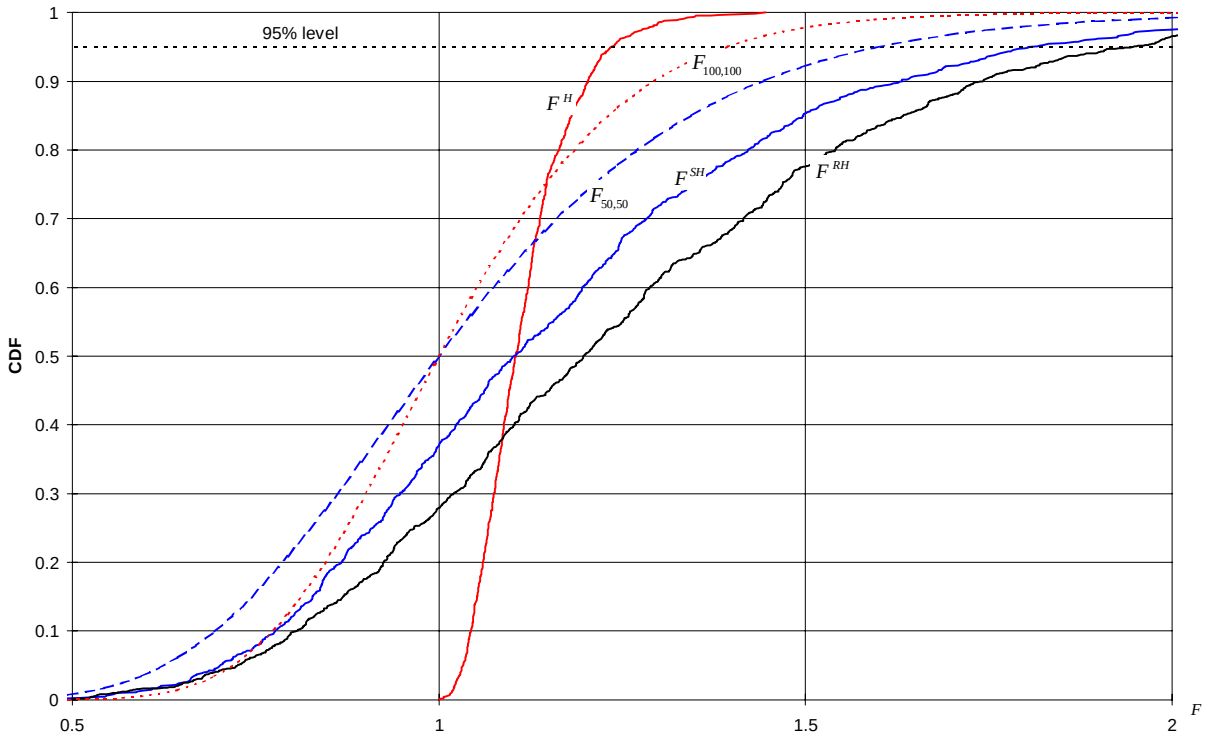



*Figure 4: Cumulative density functions for observed and theoretical halfnormal F-statistic in base trial A n=100, g~|N(0,0.25)|, with true null hypothesis $\beta = 1$. $F^{H}$ is the observed CDF for the full sample statistic which should compare with the theoretical $F_{100,100}$, $F^{SH}$ is for the split sample statistic, and $F^{RH}$ is for the separate reference set statistic, the last two should compare with the theoretical $F_{50,50}$.*

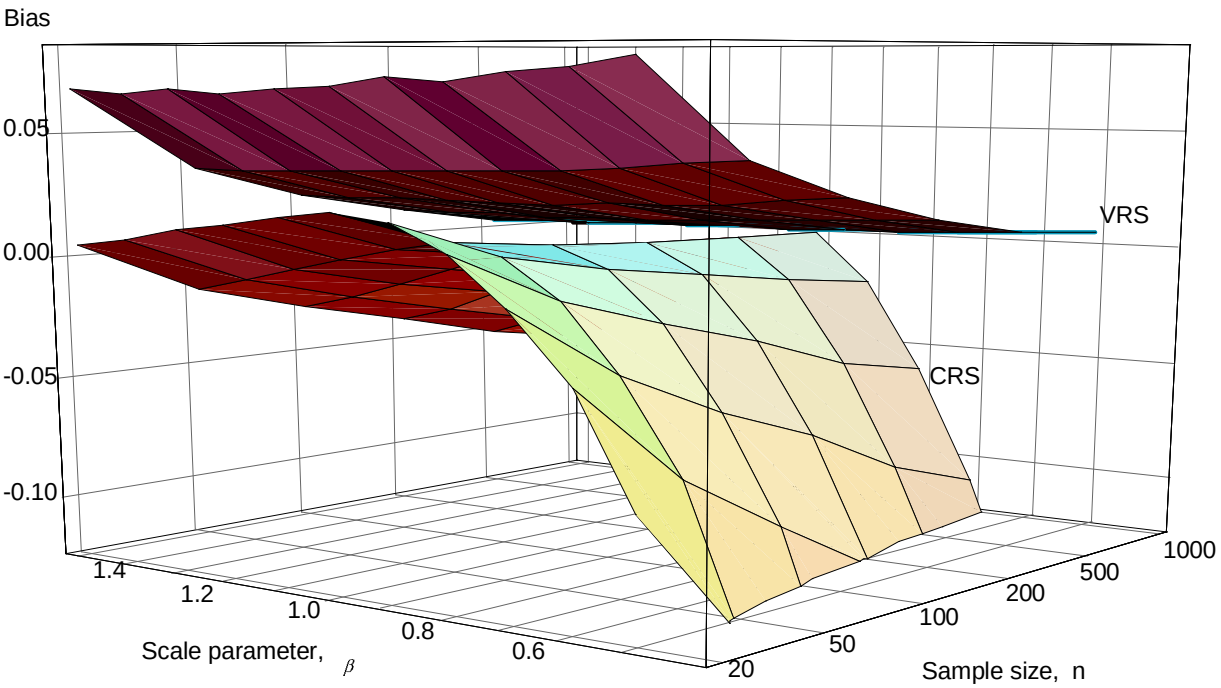



*Figure 5: Bias as function of scale parameter and sample size in trial B.*

good estimators with very low MSE both as estimators $\hat{E}_{ij}^{0}$ of the efficiency of individual units, and as estimators $\overline{\hat{E}}_{j}^{0}$ of the sample mean efficiencies. In fact the CRS estimators are better than the VRS estimators by the MSE criteria even for quite a wide range of non-CRS scale parameter values[17]. This statistical efficiency supports the choice of CRS as the conservative null hypothesis.

*Tests*

All the tests suggested above rely on an expectation that the efficiency distributions will be approximately equal when the null hypothesis is true. The fact that there is a difference in mean bias between the CRS and VRS estimators at $\beta = 1$, will tend to increase the values of the test statistics, and therefore make rejection more likely.

The dependence between the two estimators will work in the opposite direction, since the efficiency estimates will tend to be more equal. The first line of table 2 show the linear dependence of the estimates within the sample, while the second shows the correlation between the sample mean CRS and VRS estimates. It is this latter strong dependence that motivates the split sample tests. The third line shows any remaining dependence between split sample means, which could justify the separate reference set tests. This dependence seems to be negligible or non-existent.

While the bias in the split sample estimators is approximately the same as in the full sample, the halving of the size of the separate reference sets increases noticeably the bias for both the CRS and VRS estimators. This can also be seen in panel a) of figure 3, where both the level and the difference between the biases has increased.

The consequences of these biases and dependencies for the different tests are tabulated in the lower half of table 2 and shown in panels b)-d) of figure 3. The paired T-tests is seriously affected by the bias, and reject far too many under the null hypothesis. All the

[17] The MSE of the individual efficiency estimates becomes less for VRS than CRS again at the scale elasticity $\beta = 1.5$, outside the right of table 1.

**Table 3. Results from trial B. Differing sample sizes.**

| Common conditions: | Number of samples s=1000, One input/one output, Normal distribution of output $y \sim N(10,2)$. Halfnormal distributions of inefficiencies $\gamma \sim \lvert N(0,0.25) \rvert$ implying E(γ)=0.399 and E($E$)=0.745. | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trial | B1 | | | B2 | | | (B3=A) | | | B4 | | | B5 | | | B6 | | |
| Sample size, $n$ | 20 | | | 50 | | | 100 | | | 200 | | | 500 | | | 1000 | | |
| Scale parameter, $\beta$ | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 |
| CRS bias, $B^0$ | 0.010 | 0.023 | 0.016 | -0.011 | 0.009 | -0.001 | -0.025 | 0.005 | -0.009 | -0.037 | 0.002 | -0.016 | -0.052 | 0.001 | -0.022 | -0.060 | 0.000 | -0.027 |
| VRS bias, $B^1$ | 0.072 | 0.070 | 0.067 | 0.038 | 0.037 | 0.035 | 0.023 | 0.022 | 0.021 | 0.014 | 0.012 | 0.011 | 0.007 | 0.006 | 0.005 | 0.004 | 0.003 | 0.002 |
| Full sample rejection rates in percent (5% tests) | | | | | | | | | | | | | | | | | | |
| $F^H$ | 12.2 | 4.7 | 7.2 | 18.4 | 1.3 | 3.5 | 47.6 | 0.2 | 5.6 | 85.1 | 0.0 | 19.3 | 100.0 | 0.0 | 83.3 | 100.0 | 0.0 | 99.8 |
| $F^E$ | 11.5 | 3.6 | 4.2 | 15.1 | 1.1 | 2.6 | 42.2 | 0.0 | 3.4 | 81.8 | 0.0 | 13.4 | 99.8 | 0.0 | 74.9 | 100.0 | 0.0 | 99.7 |
| $D^+$ | 15.5 | 5.5 | 8.0 | 20.2 | 2.5 | 5.1 | 46.0 | 0.2 | 7.3 | 81.2 | 0.0 | 18.8 | 99.6 | 0.0 | 69.9 | 100.0 | 0.0 | 98.7 |
| $T$ | 22.2 | 10.2 | 13.1 | 44.2 | 7.6 | 16.7 | 78.8 | 2.3 | 29.5 | 98.0 | 0.6 | 63.8 | 100.0 | 0.0 | 98.5 | 100.0 | 0.0 | 100.0 |
| $T^P$ | 99.2 | 96.5 | 98.2 | 99.9 | 99.2 | 99.9 | 100.0 | 99.4 | 100.0 | 100.0 | 99.8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Split sample rejection rates in percent (5% tests) | | | | | | | | | | | | | | | | | | |
| $F^H$ | 23.5 | 15.4 | 18.1 | 24.4 | 12.1 | 18.0 | 33.3 | 9.6 | 16.6 | 54.3 | 10.7 | 22.1 | 88.8 | 8.0 | 44.3 | 99.2 | 8.3 | 70.4 |
| $F^E$ | 11.3 | 12.6 | 13.3 | 19.2 | 7.4 | 12.4 | 27.7 | 6.2 | 11.0 | 49.9 | 5.3 | 15.1 | 88.7 | 3.7 | 37.1 | 99.3 | 2.5 | 69.6 |
| $D^+$ | 3.9 | 8.1 | 9.1 | 22.4 | 11.6 | 14.6 | 32.7 | 12.4 | 23.1 | 58.0 | 7.1 | 23.8 | 92.3 | 7.7 | 44.5 | 99.7 | 7.6 | 80.9 |
| $T$ | 21.1 | 15.6 | 16.2 | 31.7 | 16.5 | 23.3 | 50.7 | 12.8 | 25.3 | 73.2 | 12.2 | 37.8 | 97.1 | 10.6 | 63.5 | 99.9 | 9.0 | 89.0 |

other full sample tests shown in panel a) have sizes less than the theoretical 5% significance level. Among these the T-test stands out as clearly more powerful and is therefore best by our criteria. The two F-tests and the Kolmogorov-Smirnov test have essentially the same power curve, but since the true inefficiency distribution is halfnormal, it is not surprising that the halfnormal F-test does slightly better than the other two. All these tests are unbiased in the sense that the rejection rate is lowest at the true hypothesis scale parameter value (Greene, 1993a, p.127) .

The position of the power curves is the result of the interaction of the relative bias in an upward direction and the dependence in a downward direction. It is not therefore surprising that removing dependence by splitting the sample in panel b) of figure 3 shifts the power curves up. Not only are the sizes of tests too high, but the reduction in degrees of freedom flattens the power curves so that at both extremes of the size parameter these tests have less power than the full sample tests.

The separate reference set test are even worse. The reduced estimated technology increases the relative bias without eliminating any real dependency. The result is power curves that reject true null hypothesis too often and more often than the split sample tests with common reference set. The results for all other trial are similar and the separate reference set tests will henceforth not be reported.

The problem with strong dependence between efficiency estimators when calculating the statistics using the full sample for both null and alternative models is illustrated in figure 4. The $F^H$ statistic calculated under the true assumption of halfnormal inefficiency terms has a distribution clearly different from the theoretical *F* value, starting at 1 and increasing much more rapidly than the theoretical F-distribution with 100,100 degrees of freedom[18]. The critical value from the theoretical distribution rejects far too few cases at the 95% level of significance.

[18] Since this statistic by (15) and (17) is bound to be greater or equal to 1, an alternative test would be to truncate the F-distribution. In fact, this truncated distribution fares even worse in the region where rejection takes place.

Figure 4 also shows the halfnormal $F^H$ is shown when calculated with randomly split samples. Here the two distributions are very similar, but with a clear shift outwards for the observed $F^H$. Use of the theoretical critical value would reject the null hypothesis in 10.8% of the iterations, about twice as many as should have been rejected (5%). It is obviously the impact of the different bias in the two models that is creating this shift, and it is clear that if one had a correct estimator of the difference in bias, the tests could also be corrected. As it is, however, the split sample tests fail the size criteria[19].

While such figures suggest that one could tabulate an alternative critical value from the Monte Carlo simulated distributions of the test statistics, the steepness of the of the full sample distribution conveys that the critical values could be quite sensitive to the model assumptions.

### 4.2 Testing for Returns to Scale: Varying the Assumptions

*Sample size*

Table 3 reports the summary results of the B set of trials, where the sample size is varied from 20 to 1000. The bias reduces clearly with sample size in all cases, for the efficiency estimate in the VRS case the reduction is from around seven per cent to less than half of one per cent. In fact, when $\beta = 1$, the logarithm of the bias is almost linear in the logarithm of sample size with a regression equation of

$$\ln(\overline{B}_j^0) = -0.949 - 0.959 \ln(n), \quad R^2 = 0.998 \tag{29}$$

in the CRS case, and

$$\ln(\overline{B}_j^1) = -0.259 - 0.782 \ln(n), \quad R^2 = 0.999 \tag{30}$$

[19] Analogous graphs for the other tests show a very similar picture.

in the VRS case. This supports the finding by Korostelev, Simar & Tsybakov (1995a) that the rate of convergence is a power function of the sample size.

Figure 5 shows the interaction of sample size and scale parameter in determining the bias. The $\beta = 1$ line from lower left to upper right represents the biases under the null hypothesis but varying scale. The n=100 line from lower right to upper left represent the bias across different scale parameters and corresponds to panel a) of figure 3. The VRS bias surface is almost invariant with respect to the scale parameter, but converges clearly towards zero as the sample size increases away and to the right in the figure. The CRS bias surface responds much sharper to changes in the scale parameter, as it should, and particularly for low scale elasticities and large sample sizes to the right in the figure. For the largest sample size of 1000, the two surfaces almost touch at zero bias when the null hypothesis is true, giving support to the asymptotic properties of the estimators and the tests.

The lower parts of table 3 shows how the full sample tests reject the null hypothesis in far too few cases, except when the sample is very small. The strong dependence between estimators destroys the sizes. The paired T-test however rejects in far too many cases for all sample sizes, due to the one-sided nature of the bias. In a sense, this test is too powerful when bias is not corrected.

The split sample F-test with a (true) halfnormal assumption tends to get asymptotically closer to the correct rejection level of 5% as the sample size increases, but always rejects in too many cases. The split sample F-test with an exponential assumption overshoots and rejects too few in the largest samples. The split sample T-test performs similarly to the $F^H$ -test in the sense that it consistently over-rejects, but it has a higher rejection rate than the halfnormal F-test. Even when the split sample tests have sizes below the specified 5%, they are not as powerful as the full sample tests which are still best.

Among the full sample tests the T-test is best for sample sizes above 50, but has too many Type I errors to meet the size criteria in the smallest samples. The halfnormal F-test does best when the sample size is 20, and the $D^+$ test does best for n=50, but as figure 6a) shows, in both cases the power functions are so flat that it is difficult to speak

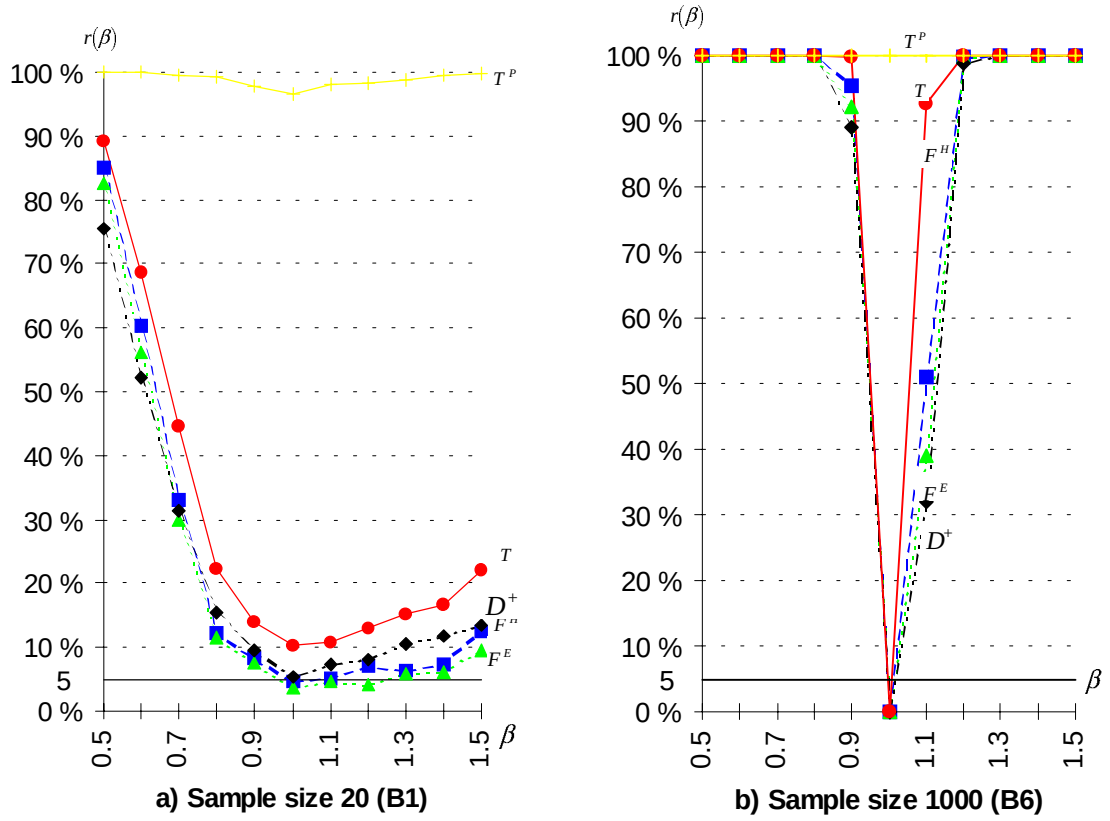



*Figure 6: Power curves for full sample tests in trials B.*

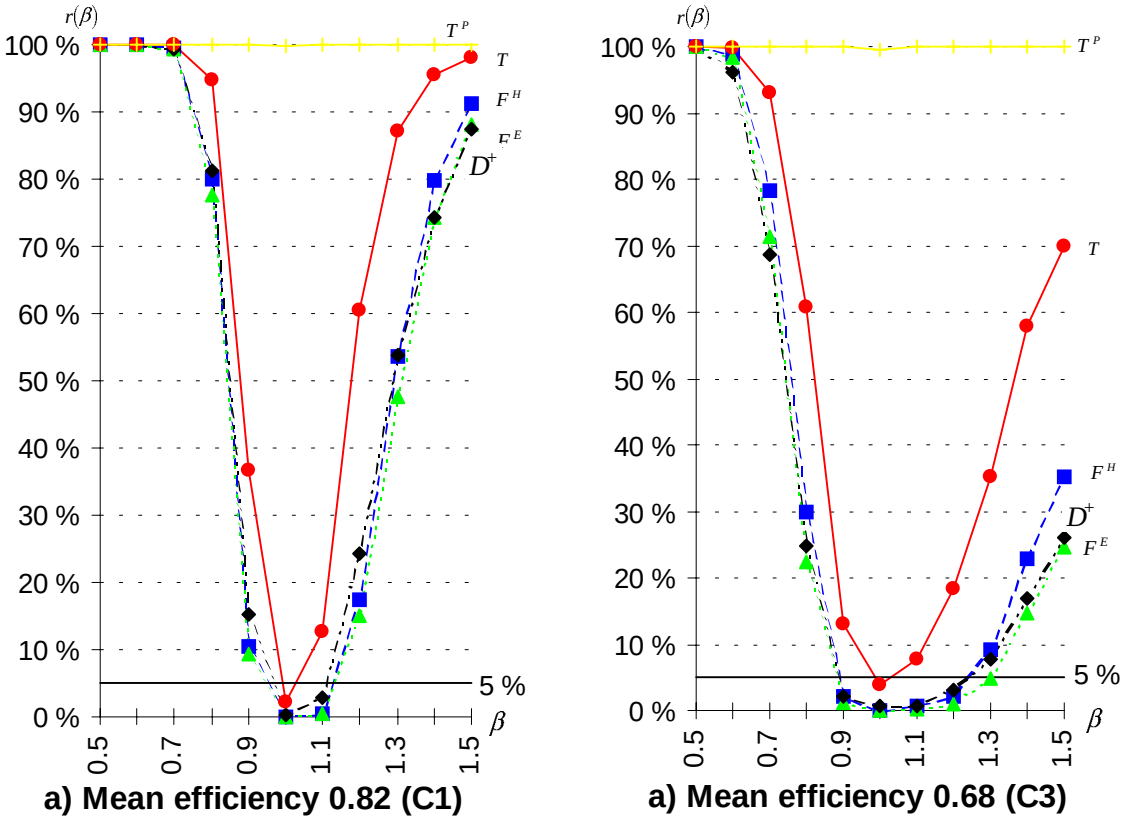



*Figure 7: Power curves for full sample tests in trials C.*

**Table 4. Results from trial C. Different efficiency levels.**

| Common conditions: | Number of samples s=1000, Sample size n=100, One input/one output, Normal distribution of output with $y \sim N(10,2)$ , Halfnormal distributions of inefficiencies. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Trial | C1 | | | C2 (=A) | | | C3 | | |
| Distribution of inefficiencies $\gamma$ | $\lvert N(0,0.1)\rvert$ | | | $\lvert N(0,0.25)\rvert$ | | | $\lvert N(0,0.5)\rvert$ | | |
| Expected efficiency E($E$) | 0.816 | | | 0.745 | | | 0.683 | | |
| Scale parameter, $\beta$ | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 |
| CRS bias, $B^0$ | -0.037 | 0.003 | -0.015 | -0.025 | 0.005 | -0.009 | -0.018 | 0.006 | -0.004 |
| VRS bias, $B^1$ | 0.017 | 0.016 | 0.014 | 0.023 | 0.022 | 0.021 | 0.028 | 0.027 | 0.026 |
| Full sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 79.9 | 0.1 | 17.3 | 47.6 | 0.2 | 5.6 | 29.8 | 0.1 | 2.2 |
| $F^E$ | 77.7 | 0.1 | 15.0 | 42.2 | 0.0 | 3.4 | 22.5 | 0.1 | 1.0 |
| $D^+$ | 81.2 | 0.3 | 24.2 | 46.0 | 0.2 | 7.3 | 24.7 | 0.7 | 3.2 |
| $T$ | 94.7 | 2.1 | 60.5 | 78.8 | 2.3 | 29.5 | 60.7 | 3.8 | 18.2 |
| $T^P$ | 100.0 | 99.8 | 100.0 | 100.0 | 99.4 | 100.0 | 100.0 | 99.6 | 100.0 |
| Split sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 50.3 | 11.0 | 21.9 | 33.3 | 9.6 | 16.6 | 27.5 | 11.6 | 17.8 |
| $F^E$ | 48.0 | 6.1 | 16.7 | 27.7 | 6.2 | 11.0 | 19.5 | 5.9 | 11.2 |
| $D^+$ | 56.0 | 8.4 | 22.0 | 33.1 | 8.0 | 15.8 | 22.6 | 8.7 | 13.3 |
| $T$ | 71.2 | 14.3 | 35.4 | 50.7 | 12.8 | 25.3 | 36.8 | 15.4 | 24.2 |

of any power at all. Returns to scale are not really testable for sample sizes less than 100. For the largest samples, all full sample tests are good.

*Efficiency level*

Table 4 reports the results from the C trials where the level of inefficiency varies. Bias is clearly increasing with inefficiency. However, bias is a fairly constant proportion of the estimated inefficiency term, varying from 2% to 2.5% for the CRS estimator when CRS is true and from 10% to 12% for the VRS estimator. This would indicate that a bias correction term should be multiplicative.

The suggested hypothesis tests in table 4 and figure 7 show a picture very similar to that in the previous trials. Full sample tests, except the paired T-test, pass the size criteria, while the split sample test do not. Even though the exponential F-test almost has a small enough size, the full sample tests are all more powerful with decreasing returns to scale, and the full sample T-test is also more powerful with increasing returns to scale. However, none of the tests are very powerful for values of the scale parameter above 1, especially when the efficiency levels are low as in C3.

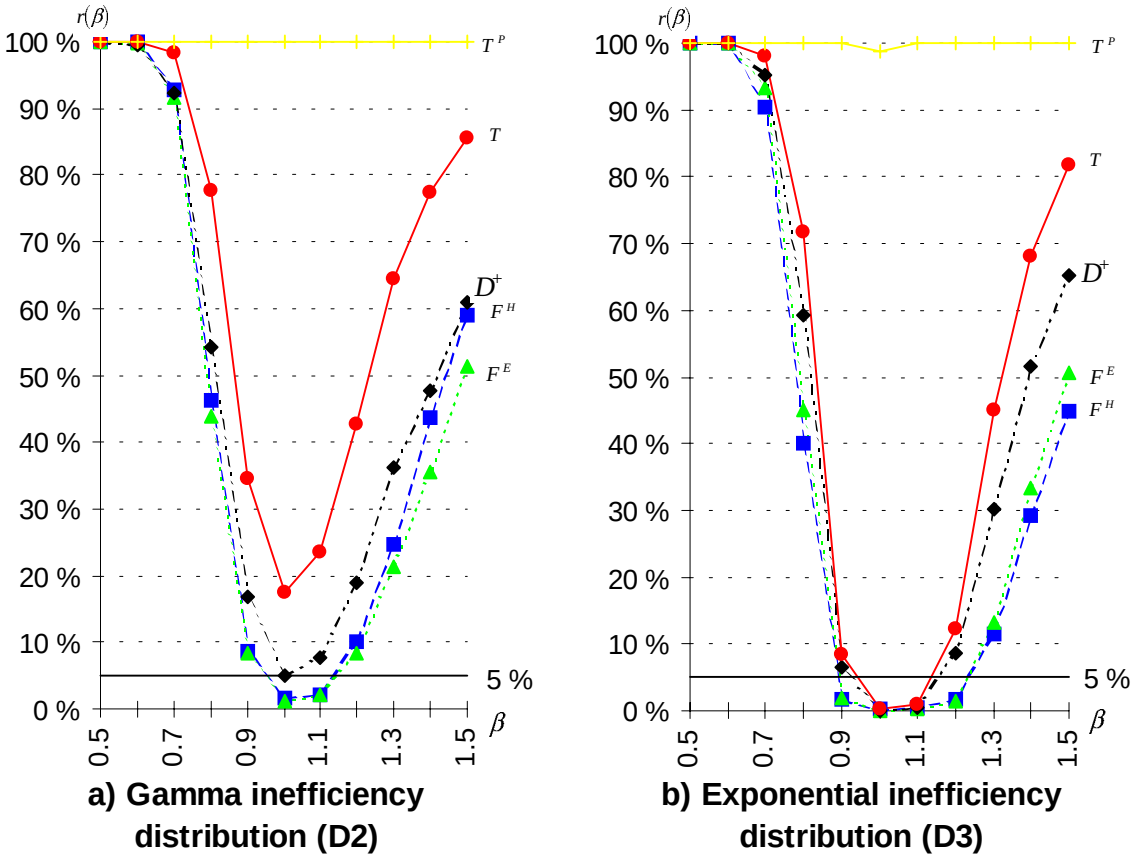



*Figure 8: Power curves for full sample tests in trials D.*

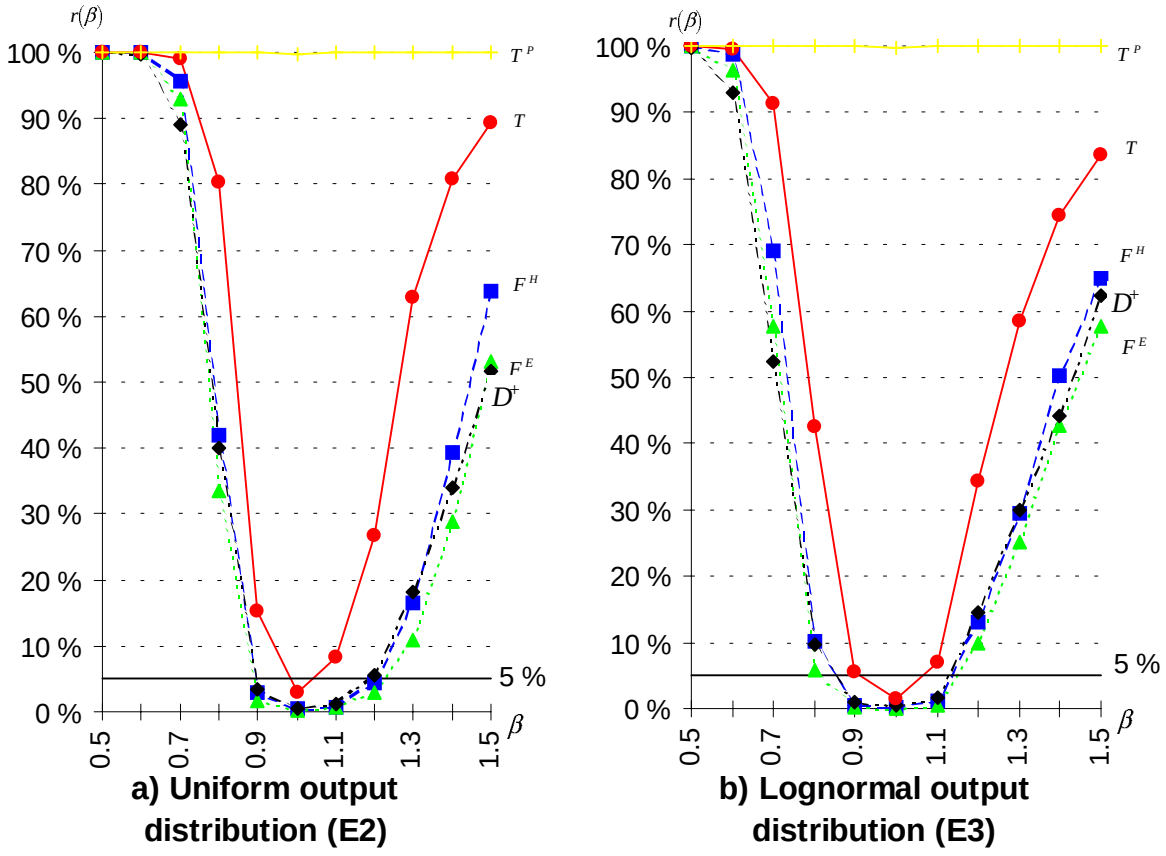



*Figure 9: Power curves for full sample tests in trials E.*

**Table 5. Results from trial D. Different inefficiency distribution functions.**

| Common conditions: | Number of samples s=1000, Sample size n=100, One input/one output, Normal distribution of output with $y \sim N(10,2)$ , Distributions of inefficiencies with E(γ)=0.3989. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Trial | D1 (=A) | | | D2 | | | D3 | | |
| Distribution of inefficiency γ | Halfnormal: \|N(0,0.25)\| | | | Gamma(2,0.1995) | | | Exponential (0.3989) | | |
| Scale parameter, $\beta$ | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 |
| CRS bias, $B^0$ | -0.025 | 0.005 | -0.009 | -0.004 | 0.019 | 0.011 | -0.033 | 0.003 | -0.013 |
| VRS bias, $B^1$ | 0.023 | 0.022 | 0.021 | 0.043 | 0.042 | 0.041 | 0.018 | 0.016 | 0.015 |
| Full sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 47.6 | 0.2 | 5.6 | 46.2 | 1.7 | 10.1 | 40.0 | 0.3 | 1.6 |
| $F^E$ | 42.2 | 0.0 | 3.4 | 43.8 | 1.2 | 8.5 | 45.0 | 0.0 | 1.4 |
| $D^+$ | 46.0 | 0.2 | 7.3 | 54.3 | 5.0 | 19.0 | 59.3 | 0.0 | 8.7 |
| $T$ | 78.8 | 2.3 | 29.5 | 77.7 | 17.6 | 42.6 | 71.8 | 0.2 | 12.2 |
| $T^P$ | 100.0 | 99.4 | 100.0 | 100.0 | 99.9 | 100.0 | 100.0 | 98.8 | 100.0 |
| Split sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 33.3 | 9.6 | 16.6 | 36.1 | 16.3 | 21.7 | 36.2 | 17.9 | 29.2 |
| $F^E$ | 27.7 | 6.2 | 11.0 | 29.7 | 9.7 | 15.1 | 31.0 | 9.5 | 19.2 |
| $D^+$ | 33.1 | 8.0 | 15.8 | 36.4 | 12.4 | 19.9 | 37.9 | 7.0 | 14.5 |
| $T$ | 50.7 | 12.8 | 25.3 | 51.3 | 19.8 | 31.9 | 43.4 | 10.4 | 23.8 |

**Table 6. Results from trial E. Different distributions of output y.**

| Common conditions: | Number of samples s=1000, Sample size n=100, Halfnormal inefficiency distribution $\gamma \sim \lvert N(0,0.25) \rvert$, One input/one output, Distribution of output with E(y)=10, SD(y)= $\sqrt{2}$ . | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Trial | E1 (=A) | | | E2 | | | E3 | | |
| Distribution of output *y* | Normal(10,2) | | | Uniform($10 \pm \sqrt{6}$) | | | Lognormal(10, $\sqrt{2}$ ,10-$\sqrt{6}$ ) | | |
| Scale parameter, $\beta$ | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 |
| CRS bias, $B^0$ | -0.025 | 0.005 | -0.009 | -0.022 | 0.005 | -0.009 | -0.012 | 0.005 | -0.012 |
| VRS bias, $B^1$ | 0.023 | 0.022 | 0.021 | 0.023 | 0.022 | 0.021 | 0.021 | 0.020 | 0.019 |
| Full sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 47.6 | 0.2 | 5.6 | 41.9 | 0.5 | 4.4 | 10.1 | 0.2 | 13.0 |
| $F^E$ | 42.2 | 0.0 | 3.4 | 33.6 | 0.2 | 2.8 | 5.7 | 0.1 | 10.0 |
| $D^+$ | 46.0 | 0.2 | 7.3 | 40.1 | 0.5 | 5.6 | 9.6 | 0.4 | 14.5 |
| $T$ | 78.8 | 2.3 | 29.5 | 80.4 | 3.0 | 26.6 | 42.5 | 1.4 | 34.4 |
| $T^P$ | 100.0 | 99.4 | 100.0 | 100.0 | 99.7 | 100.0 | 100.0 | 99.8 | 100.0 |
| Split sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 33.3 | 9.6 | 16.6 | 31.3 | 10.6 | 19.5 | 21.0 | 9.7 | 20.1 |
| $F^E$ | 27.7 | 6.2 | 11.0 | 24.4 | 5.5 | 13.0 | 14.6 | 5.1 | 12.6 |
| $D^+$ | 33.1 | 8.0 | 15.8 | 29.7 | 7.8 | 14.4 | 17.3 | 7.4 | 17.4 |
| $T$ | 50.7 | 12.8 | 25.3 | 45.3 | 12.4 | 27.2 | 30.5 | 11.8 | 28.0 |

*Inefficiency distribution shape*

The functional form of the distribution of the inefficiency term is varied in the D trials reported in table 5 and figure 8. These trials are calibrated with a common expected inefficiency term $\gamma$, but through the transformation in (10), this does not give rise to a common mean efficiency level $E$. The exponential distribution has the greatest density at full efficiency, while the gamma distribution has a mode below 0.9 and a zero density at 1 (Johnson & Kotz, 1970a). This means that the Gamma distribution does not belong to the class that make the DEA estimators maximum likelihood estimators, but since it has a positive density arbitrarily close to the frontier, it still satisfies assumption A4) necessary for consistency.

This naturally gives rise to a greater average bias in the gamma-distributed case, and the least bias with the exponential distribution. The variation is quite large, and it does not seem possible to construct any bias-correction measure without knowing the true distributional form. Again, the tests show much the same pattern. Interestingly, in the exponential case the F-test that assumes the correct distribution does only slightly better than the halfnormal F-test, even though both have less power than the T-test and the $D^+$ test. The T-test is most affected by the higher bias of the Gamma-distribution, resulting in failure the size criteria. For this distribution the Kolmogorov-Smirnov test is best.

*Output distribution shape*

Table 6 and figure 9 report the results of trial E where the generated distribution of the output *y* is varied. These distributions are calibrated to have the same mean and standard deviation. The Lognormal distribution, which has three parameters, is in addition constructed to have the same lower bound as the uniform distribution, but has like the normal distribution no upper bound.

The CRS estimates are nearly identical for the three trials. In the VRS case the bias is slightly lower when the output is lognormally distributed. This does not lead to noticeable differences in the tests, and in all cases the T-test is best, with the F-tests and $D^+$ test also meeting the size criteria.

**Table 7. Results from trial F. Different number of inputs.**

| Common conditions: | Number of samples s=1000, Sample size n=100, Halfnormal inefficiency distribution $\gamma \sim \lvert N(0,0.25) \rvert$, Normal distribution of output, Input mix as ratio of normally distributed numbers, all N(10,2), Cobb-Douglas production function with equal share parameters. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Trial | F1 (=A) | | | F2 | | | F3 | | |
| Number of inputs, $L$ | 1 | | | 2 | | | 3 | | |
| Scale parameter, $\beta$ | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 | 0.8 | 1 | 1.2 |
| CRS bias, $B^0$ | -0.025 | 0.005 | -0.009 | 0.001 | 0.019 | 0.011 | 0.021 | 0.036 | 0.030 |
| VRS bias, $B^1$ | 0.023 | 0.022 | 0.021 | 0.046 | 0.044 | 0.042 | 0.065 | 0.063 | 0.061 |
| Full sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 47.6 | 0.2 | 5.6 | 44.2 | 2.0 | 6.2 | 47.1 | 3.7 | 11.5 |
| $F^E$ | 42.2 | 0.0 | 3.4 | 42.9 | 1.9 | 4.7 | 50.3 | 4.8 | 10.6 |
| $D^+$ | 46.0 | 0.2 | 7.3 | 37.6 | 2.9 | 9.0 | 39.2 | 5.1 | 8.6 |
| $T$ | 78.8 | 2.3 | 29.5 | 75.9 | 10.9 | 28.6 | 74.6 | 15.6 | 30.7 |
| $T^P$ | 100.0 | 99.4 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Split sample rejection rates in percent (5% tests) | | | | | | | | | |
| $F^H$ | 33.3 | 9.6 | 16.6 | 32.7 | 19.3 | 23.6 | 34.7 | 18.9 | 26.3 |
| $F^E$ | 27.7 | 6.2 | 11.0 | 28.4 | 14.4 | 16.7 | 31.7 | 16.0 | 20.7 |
| $D^+$ | 33.1 | 8.0 | 15.8 | 28.7 | 13.5 | 17.5 | 28.0 | 15.6 | 18.7 |
| $T$ | 50.7 | 12.8 | 25.3 | 44.9 | 23.8 | 26.9 | 42.0 | 22.3 | 29.1 |

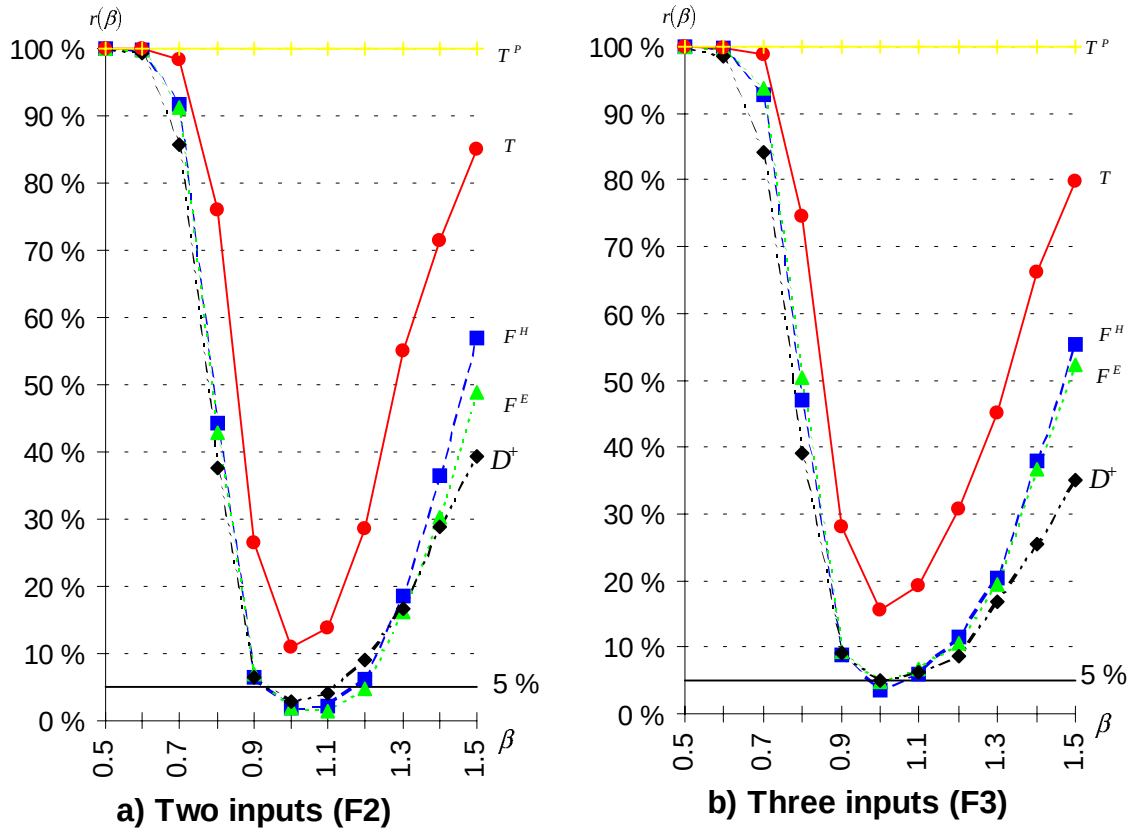



*Figure 10: Power curves for full sample tests in trials F.*

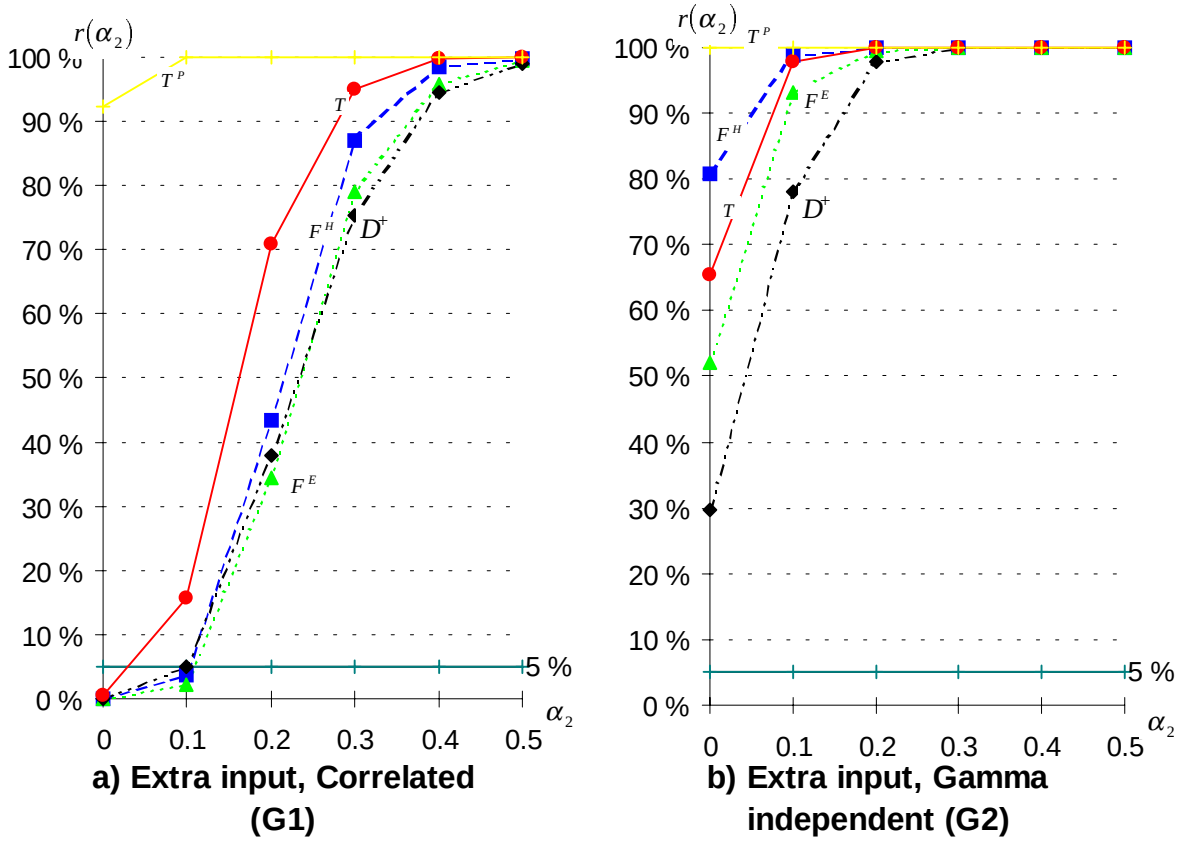



*Figure 11: Power curves for full sample tests in trials G.*

*Dimensionality*

All previous trials have had the same dimensionality with one input and one output, but in the F trials reported in table 7 the number of inputs is varied. The production frontier is no longer linear, but is a Cobb-Douglas core with equal factor share parameters $\alpha_l = 1/L$ in the frontier function in (4). This represents two types of shift in the analysis. Firstly the computational burden of the simulations is greatly increased. Secondly, Kneip, Park and Simar (1996) notes that there is a qualitative step in the deterioration of the rate of convergence of the efficiency estimators.

Table 7 shows indeed that increasing dimensionality leads to dramatic increases in the bias of the estimators. In the true CRS case bias increases by a factor of four from the one-input to the two-input case, doubling again to the three-input case. The VRS estimate biases increase more slowly, by approximately 100% and 50% respectively, so that the differences between the CRS and VRS estimators are in fact only moderately increased from 0.017 to 0.027.

This is reflected in the tests, whose rejection rates generally increases somewhat. Even though bias increases markedly with increased dimensionality, since this happens for both estimators, the sizes of the tests are only moderately affected. The full sample T-test fails the size criteria when there are multiple inputs, so that the full sample F and $D^+$ tests contend for best place. There is not much difference between them, but none are particularly powerful if there are increasing returns to scale in the true technology.

### 4.3 Testing for Variable Inclusion

The final set of trials differs from the others in that the null hypothesis is not that the production function exhibits constant returns to scale, but rather that there is only one relevant input. The trials have observations that are generated under the same assumptions that underlie the basic model, with halfnormal inefficiency, a sample size of 100, a CRS technology assumption, one input and one output. In these trials an extra

variable $x_2$ is generated that has the same random normal distribution as the output, but is irrelevant to the production technology.

The null hypothesis is the true one that the extra variable is not relevant to the production technology, while the alternative hypothesis is that $x_2$ should be included in the analysis as an input. In both cases the DEA model is solved as a CRS problem, but in model 1 there is an added restriction due to the extra variable; there is in other words an extra dimension. By proposition 6 in the appendix the one-input model is nested within the two-input model in the same way as the CRS model is nested within the VRS model. Under the true null hypothesis there is the same strict ordering of estimators as in (18), but when the null hypothesis is false the 1-input bias may be negative.

The data are generated with the factor share parameter $\alpha_2$ varying from 0 to 0.5. The difference between the trials is that in G1, $x_2$ is generated in the specified data generating process, with the result that it is highly correlated with $x_1$, while in the other trials the two potential inputs are less correlated. In empirical work the first case would usually be more realistic.

For the first two trials, both input levels depend on the output. With $\beta = 1, \alpha_2 = 0$, i.e. when the null hypothesis is true, the frontier level of $x_1^*$ is determined directly from the output level $y$. In trials G1 and G2, the frontier mapping of the second input is then determined by the ratio of two normal numbers, as specified in (7). In G1 both the relevant and the irrelevant inputs are multiplied by the same inefficiency term in (8). The result is a set of power curves that are one-sided, but otherwise remarkably similar to those in the base trial A. The full sample test minus the paired T-test and plus the split sample exponential F-test meet the size criteria, but the T-test is again best.

In the second trial G2, the specified DGP is violated by introducing an input-specific inefficiency term in (8), since there could often be reason to believe that slacks will vary for the different inputs in real production activities. This reduces substantially the correlation between the input levels, at the same time increasing the bias of the two-input estimates. The result for the tests can be seen from figure 11b) to be an upward shift of

**Table 8. Results from trial G. Testing for relevance of second input.**

| Common conditions: | Number of samples s=1000, Sample size n=100, Halfnormal inefficiency distribution γ~\|N(0,0.25)\|, Normal distribution of output, Input mix as ratio of normally distributed numbers, all N(10,2), Cobb-Douglas production functions, CRS models $\beta = 1$. | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Trial | G1 | | | G2 | | | G3 | G4 |
| Second input $x_2$ dependent on output $y$ | Yes | | | Yes | | | No | No |
| Common inefficiency $\gamma$ | Yes | | | No | | | Yes | No |
| Second input share $\alpha_2$ | 0 | 0.2 | 0.4 | 0 | 0.2 | 0.4 | 0 | 0 |
| Input correlation Mean$_j\left(\hat{\rho}_i\left(x_1^{ij}, x_2^{ij}\right)\right)$ | 0.770 | 0.735 | 0.715 | 0.227 | 0.159 | 0.124 | 0.687 | 0.002 |
| 1-input estimate bias, $B^0$ | 0.005 | -0.027 | -0.073 | 0.005 | -0.027 | -0.074 | 0.005 | 0.005 |
| 2-input estimate bias, $B^1$ | 0.012 | 0.018 | 0.019 | 0.043 | 0.059 | 0.064 | 0.012 | 0.043 |
| Full sample rejection rates in percent (5% tests) | | | | | | | | |
| $F^H$ | 0.1 | 43.4 | 98.4 | 80.7 | 100.0 | 100.0 | 0.0 | 79.4 |
| $F^E$ | 0.0 | 34.4 | 95.7 | 52.0 | 99.6 | 100.0 | 0.0 | 51.6 |
| $D^+$ | 0.0 | 38.0 | 94.4 | 29.8 | 97.9 | 100.0 | 0.0 | 31.1 |
| $T$ | 0.4 | 70.7 | 99.7 | 65.4 | 99.9 | 100.0 | 0.1 | 64.4 |
| $T^P$ | 92.2 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 94.4 | 99.9 |
| Split sample rejection rates in percent (5% tests) | | | | | | | | |
| $F^H$ | 7.5 | 33.3 | 73.4 | 52.0 | 92.7 | 99.9 | 7.0 | 50.4 |
| $F^E$ | 3.2 | 23.6 | 67.7 | 31.5 | 82.9 | 99.5 | 3.2 | 29.9 |
| $D^+$ | 5.3 | 27.3 | 72.7 | 26.4 | 76.6 | 98.0 | 4.8 | 26.9 |
| $T$ | 8.5 | 43.4 | 84.6 | 40.7 | 89.6 | 99.8 | 7.1 | 42.1 |

The definition of mean is given in table 1 and correlation in table 2.

all power curves, so that no test now meet the size criteria. However, the Kolmogorov-Smirnov test, and this time the split-sample variant, comes closest by having the least proportion of Type I errors.

The final two trials violate the DGP further, by removing the dependence of the second input on the output level, instead assuming an independent "optimal" input drawn from the same distribution. The generated $x_2^*$ are thereby not on the isoquant, and generating observations with positive output shares would be meaningless. By varying the inefficiency assumptions two different correlations with the first input are achieved. The results, both for bias when the null hypothesis is true, and for the size of the tests, are quite similar to the first two G trials. In deciding the variable specification of a model, it is clearly important to measure the extent of correlation. If correlation is low, a conservative test is advisable.

These simulations are based on comparing 1-input and 2-input estimates, which, as noted earlier, implies a qualitative step in the increase in bias. One would therefore expect the relative increase to be less when comparing L-input to L+1-input estimates when L>1, and rejection rates would therefore generally be lowered. This would make the full sample tests usable for considerably lower input correlations than 0.7.

Finally, it should be noted that variable aggregation implies very similar model changes as does variable exclusion. By proposition 6 in the appendix, an aggregated model is nested in a disaggregated one, and has therfore equal or lower efficiency estimates. Although I offer no simulations to support it, this would imply that the tests should be usable for approximately the same ranges of sample sizes and variable correlations.

## 5. Conclusion

The simulations show that bias is important, and that it varies systematically, increasing with dimensionality, and decreasing with sample size, average efficiency and a high density of observations near the frontier. The size and power of the suggested tests are both increased by this bias, offsetting the reduced size and power stemming from the

dependence of the efficiency estimators. Although the last set of trials show that the tests suggested so far in the literature are in no way perfect, there are some substantive findings that should give grounds for conclusions in empirical work.

Firstly, the full sample tests generally do better than the split sample tests, the latter being not very powerful. In fact correcting for dependency does not seem helpful unless one can also correct for bias. Secondly, the T-test seems less affected by the dependence in the full samples, so that if bias is low due to e.g. large samples, low dimensionally, and inefficiency distributions which are dense near the frontier, the T-test is quite useful. If, however, bias is somewhat higher, due to medium sample size, higher dimensionality and inefficiency distributions that have their mode away from the frontier, the F-tests and the Kolmogorov-Smirnov tests seem best. There is not much difference between these. Finally, if bias is expected to be very high due to sample size less than about 100, none of these test will be very good.

Empirical analysis could therefore be done using the full sample tests, if proper consideration is taken of the inaccuracies reported above, notably that the tests are approximations and that the sample size should not be to small.

At this point it would have been possible to tabulate corrected critical values for the different tests in each trial, by sorting the 1000 Monte Carlo values of the null hypothesis run, and reporting e.g. the $10^{th}$ and $50^{th}$ value. For example, for the full sample halfnormal F test, the theoretical critical value in the A trials is 1.392, but the Monte Carlo estimate for the critical value is 1.230. Using corrected critical values for all the full sample tests in trials A ensures correct size and maximises their power. The problem is that these critical values depend on those factors which determine the relative bias, and these factors are often quite specific to each application. It would require quite a large set of tables to have corrected critical values for all categories, unless the systematic variations could be approximated by simple formulas. These would still only be approximations, since as pointed out by Gijbels et al. (1996) the true value of the bias depends on the density of the efficiencies near the frontier and the curvature of the frontier at each point.

The results show that some of the evaluated tests are useful, but at the same time better tests and less biased efficiency estimators are needed for cases with small samples and large dimensionality. Possibly, the only way forward is through bootstrapping along the lines of Simar & Wilson (1997), but further theoretical work building on Korostelev et al. (1995a,1995b) and/or further experimental work could result in bias correction formulas that give better estimators of efficiency, and at the same time gives access to better testing tools.

**References**

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

Aigner, D.J., C.A.K. Lovell and P. Schmidt, 1977, Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics* 6, 21-37.

Banker, R.D., 1993, Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation, *Management Science* 39, 1265-1273.

Banker, R.D., 1996, Hypothesis Tests Using Data Envelopment Analysis, *Journal of Productivity Analysis* 7, 139-159.

Banker, R.D., A. Charnes and W.W. Cooper, 1984, Some models for estimating technical and scale inefficiencies, *Management Science* 30, 1078-1092.

Bauer, P.W., 1990, Recent developments in the econometric estimation of frontiers, *Journal of Econometrics* 46, 39-56.

Bhattacharyya, G.K. and R.A.Johnson, 1977, *Statistical concepts and methods* (John Wiley & Sons, New York).

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

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

Charnes,A., W.W.Cooper and E.Rhodes, 1978, Measuring the efficiency of Decision Making Units, *European Journal of Operational Research* 2, 429-444.

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

Deprins, D., L.Simar and H.Tulkens, 1984, Measuring labour-efficiency in post offices, ind M.Marchand, P.Pestiau and H.Tulkens (eds.), *The Performance of public enterprises* (North-Holland, Amsterdam).

Diewert, W.E. and M.N.F.Mendoza, 1996, The Le Chatelier Principle in Data Envelopment Analysis, Paper presented at the 2nd Georgia Productivity Workshop, University of Georgia, November 1996.

Engle, R.F., 1984, Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics, in Z.Grilliches and M.D.Intriligator, *Handbook in Econometrics* (North-Holland, Amsterdam).

Farrell, M.J., 1957, The measurement of productive efficiency, *Journal of the Royal Statistical Society* 120, 449-460.

Fried, H.O., C.A.K. Lovell and S.S. Schmidt, 1993, *The measurement of productive efficiency: Techniques and Applications*, (Oxford University Press, Oxford).

Färe, R., S.Grosskopf and C.A.K.Lovell, 1985, *The measurement of efficiency of production* (Kluwer-Nijhof, Boston).

Färe, R. and C.A.K.Lovell, 1978, Measuring the technical efficiency of production, *Journal of Economic Theory* 19, 150-162.

Färe, R. and D.Primont, 1987, Efficiency measures for multiplant firms with limited data, in: Eichhorn, ed., *Measurement in Economics* (Physica-Verlag, Heidelberg).

Førsund, F.R. and L. Hjalmarsson, 1974, On the measurement of productive efficiency, *Swedish Journal of Economics* 76, 141-54.

Førsund, F.R. and L. Hjalmarsson, 1987, *Analyses of industrial structure: A putty-clay approach* (Almqvist & Wiksell International, Stockholm).

Gijbels, I., E.Mammen, B.U.Park and L.Simar, 1996, On estimation of monotone and concave frontier functions, Discussion Paper 9611, Institut de Statistique, Université Catholique de Louvain.

Greene, W.H., 1993a, *Econometric Analysis*, 2nd edition (Prentice-Hall, London).

Greene, W.H., 1993b, The econometric approach to efficiency analysis, in Fried, Lovell and Schmidt (1993).

Grosskopf, S., 1996, Statistical inference and nonparametric efficiency: A selective survey, *Journal of Productivity Analysis, 7, 161-176.*

Johnson, N.L. and S.Kotz, 1970a, *Distributions in statistics: Continous univariate distributions-1* (Houghton Mifflin, Boston).

Johnson, N.L. and S.Kotz, 1970b, *Distributions in statistics: Continous univariate distributions-2* (Houghton Mifflin, Boston).

Kneip, A., B.U.Park and L.Simar, 1996, A note on the convergence of nonparametric DEA efficiency measures, Discussion Paper 9603, Institut de Statistique, Université Catholique de Louvain.

Korostelev, A.P., L.Simar and A.B.Tsybakov, 1995a, Efficient estimation of monotone boundries, *Annals of Statistic*s 23(2), 476-489.

Korostelev, A.P., L.Simar and A.B.Tsybakov, 1995b, On estimation of monotone and convex boundries, *Publications de l'Institut de statistique de l'Université de Paris* 39:, 3-18.

Koutsoyiannis, A., 1977, *Theory of Econometrics* (Macmillan, London).

Magnussen, J., 1996, Efficiency measurement and the operationalization of hospital production, *Health Services Research* 31, 21-37.

Meeusen, W., and J. van den Broeck, 1977, Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review* 18, 435-444.

Olesen, O.B., and N.C.Petersen, 1991, Collinearity in Data Envelopment Analysis: An extended facet approach, Publications from Department of Management, Odense University, No. 1/1991.

Olesen, O.B., and N.C.Petersen, 1996, Indicators of ill-conditioned data sets and model misspecification in Data Envelopment Analysis: An extended facet approach, *Management Science* 42, 205-219.

Petersen, N.C. and O. Olesen, 1995, Chance constrained efficiency evaluation, *Management Science,* 41, 442-457.

Press, W.H., B.P.Flannery, S.A.Teukolsky and W.T.Vetterling, 1989, *Numerical recipes in Pascal* (Cambridge University Press, Cambridge).

Seiford, L.M. and R.M. Thrall, 1990, Recent developments in DEA - The mathematical programming approach to frontier analysis, *Journal of Econometrics*, 46, 7-38.

Shephard, R.W., 1970, *Theory of cost and production functions*, 2nd ed. (Princeton University Press, Princeton).

Simar, L., 1996, Aspects of statistical analysis in DEA-type frontier models, *Journal of Productivity Analysis, 7, 177-185.*

Simar, L. and P.W.Wilson, 1995, Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science,* forthcoming. Available as Discussion Paper 9503, Institut de Statistique, Université Catholique de Louvain.

Simar, L. and P.W.Wilson, 1997, Nonparametric tests of returns to scale. Paper presented at the 5th European Workshop on Efficiency and Productivity Measurement, The Royal Veterinary and Agricultural University, Copenhagen.

Tulkens, H. and P. Vanden Eeckaut, 1991, Non-frontier measures of efficiency, progress and regress, CORE discussion paper no. 9155, Centre for Operations Research and Econometrics, Université Catholique de Louvain.

Valdmanis, V., 1992, Sensitivity analysis for DEA models. An empirical example using public vs. NFP hospitals. *Journal of Public Economics*, 48, 185-205.

## *Appendix A. Ranked and nested DEA models*

The purpose of this appendix is to prove that certain DEA models are *nested*, in that the efficiency measures of one model as a special case can be equal to those of the other, and that these models also are *ranked*, in that one knows the sign of the difference if the measures are not equal.

Let $\{1,\ldots,K\}$ be the set of all possible outputs and $\{1,\ldots,L\}$ be the set of all possible inputs. Let further $P \subset \Re_+^{K+L}$ be a set of the non-negative quantities $(y,x) = (y_1,\ldots,y_K,x_1,\ldots,x_L) \geq 0$ which defines either the true production technology $\{(y,x) | y \text{ can be produced from } x\}$, or an estimate of this set. The technology $P$ is assumed to satisfy certain regularity conditions, notably that the input requirement set $L(y) = \{x | (y,x) \in P\}$ is closed and convex for all output vectors $y \in \Re_+^K$.

***Definition 1:*** *The* Farrell (1957) Technical Input Efficiency *measure of an output-input vector* $(y,x)$ *with respect to the technology set* $P$ *is*

$$E(y,x,P) = \begin{cases} \underset{\theta}{Min}\{\theta | (y,\theta x) \in P\} \theta\{ | (\theta y, x) \in P\} \neq \varnothing \\ \infty, \text{ otherwise} \end{cases} \qquad \text{(A.1)}$$

*where the second part includes the case* $L(y) = \varnothing$ *and the axis of essential inputs.*

As noted by many authors (see e.g. Färe and Lovell, 1978), the first part of this definition is the inverse of the Shephard (1970) input distance function, and has therefore corresponding properties to those derived for the input distance function. Among these are that $E(y,x,P)$ is homogenous of degree -1 in inputs, and that it provides an equivalent characterisation of technology $P = \{(y,x) | E(y,x,P) \leq 1\}$ (Shephard, 1970, p. 68, 71).

***Definition 2****: Let* $P^0$ *be the technology set under the hypothesis* $H^0$*, and* $P^1$ *the technology set under* $H^1$*. The hypothesis* $H^0$ *is* nested *within* $H^1$ *if and only if there*

*exists a* $P^{1*}$ *such that* $E\left(y,x,P_0^0\right)=E\left(y,x,P^{1*}\right)$ *for all admissable* $P_0^0$ *when* $H^0$ *is true and for all observed points* $(y,x)$.

The parametric statistical literature uses the term *nested* loosely in the sense that model 0 is a special case of model 1, i.e. that model 0 can be obtained from model 1 by imposition of restrictions on the parameters, implying a loss of fit (see e.g. Green, 1990). Pesaran (1987) suggests a formal definition of nestedness in the parametric case, claiming that there previously had not existed any satisfactory formal definition. His definition uses the Kullback-Leibler information criterion (KLIC) for the distance between the two models in the range of variation of the data, and defines $H^0$ as nested in $H^1$ if for any true parameter vector in $H^0$, the infinum over the parameters in $H^1$ of the KLIC distance is zero. At the same time he notes that *"the classification of the hypothesis of interest into the nested and the non-nested categories does not depend on the particular distance measure used"* (p.70). Since the Farrell efficiency measure (as the reciprocal of the Shephard distance function) is a measure of distance, this could be used instead.

As an example, take a one-input one-output linear parametric production function as the frontier of the production technology, with $H^0: P^0=\left\{(y,x)\middle| y-ax\le 0\right\}$ with one parameter *a* and $H^1: P^1=\left\{(y,x)\middle| y-b-cx\le 0\right\}$ with two parameters *b,c*. Now for any true *a*, e.g. $a^0$, there are a set of parameters *b,c*, i.e. $b^*=0$, $c^*=a^0$ that ensures that $E\left(y,x,P_0^0\right)=E\left(y,x,P^{1*}\right)$ for all possible observations of $(y,x)$. In the estimation of parametric functions, this could be expressed as a restriction of the parameter *b*.

In non-parametric methods there are no direct parameters to be restricted, but as will be seen the notion of loss of fit carries over in that the distance from inefficient observations to the frontier is greater or equal in model 0. Paradoxically, this entails a less restrictive definition of the estimator of the technology set P. Definition 2 above is chosen since it closely follows the concepts behind Pesaran's definition, requiring that the set $P^{1*}$ is one that minimizes the distance $\left|E\left(y,x,P_0^0\right)-E\left(y,x,P^{1*}\right)\right|$ between the sets, and that this distance then is zero. The definition gives rise to the same categorisation of models as in the parametric literature, i.e. a constant returns model is nested within a variable returns

model, a model excluding one variable is nested within the model including this variable, etc. Following Greene (1990), the terms model and hypothesis will be used equivalently.

***Definition 3**: Model 0 is* ranked *within model 1 if* $E(y,x,P^0) \le E(y,x,P^1)$ *for all points for all* $(y,x) \in \Re_+^{K+L}$.

***Definition 4**: Model i assumes* feasibility *if all observed points* $(y^0,x^0) \in P^i$ .

***Proposition 1:** If model 0 is ranked within model 1 and the latter model is an estimator* $\hat{P}^1$ *that assumes feasibility, then model 0 is nested within model 1.*

Proof: When model 0 is true, for any true $P_0^0$, there is a possiblity of observing the point $\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o\right)$ corresponding to every other observation $(y^o,x^o)$, i.e. the true frontier mapping of each observation is also observed. By the homogeneity of –1 of the Farrell input measure with respect to inputs $E\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o, P_0^0\right)=1$. When this point is observed, it is in the model 1 estimate by the feasibility assumption, and then we know that $E\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o, \hat{P}^1\right)\le 1$. Combining this with the ranking assumption $E\left(y,x,P^0\right) \le E\left(y,x,\hat{P}^1\right)$, we get

$1 = E\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o, P_0^0\right) \le E\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o, \hat{P}^1\right)\le 1$. This can only be true if $E\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o, P_0^0\right) = E\left(y^o, E\left(y^o,x^o,P_0^0\right)x^o, \hat{P}^1\right)=1$, which by homogeneity implies $E\left(y^o,x^o,P_0^0\right) = E\left(y^o,x^o,\hat{P}^1\right)$. There is thus some set of observations that make the Farrell efficiency measures equal for all observations for any true $P_0^0$.

Note that the converse is not true. Nestedness and feasibility are not sufficient to ensure ranking, as can be seen from the example of linear production frontiers described previously.

***Proposition 2:** Model 0 is ranked within model 1, if and only if* $P^1 \subseteq P^0$.

Proof: Consider first a point $(y,x)$ where $\left\{\theta\middle|(y,\theta x)\in P^1\right\}=\varnothing$. Then by definition (A.1) $E\left(y,x,P^1\right)=\infty$ so that any value of $E\left(y,x,P^0\right)\le E\left(y,x,P^1\right)$. Consider next the point $\left(y,E\left(y,x,P^1\right)x\right)$ where $\left\{\theta\middle|(y,\theta x)\in P^1\right\}\neq\varnothing$. Since the efficiency measure is then the optimal value of $\theta$ in (A.1), this point will be on the boundary of $P^1$, and therefore $\left(y,E\left(y,x,P^1\right)x\right)\in P^1$. When $P^1\subseteq P^0$ this implies that $\left(y,E\left(y,x,P^1\right)x\right)\in P^0$. This can equivalently be written as $E\left(y,E\left(y,x,P^1\right)x,P^0\right)\le 1$, and since the efficiency measure is homogenous in degree -1 in inputs, this implies ranking $E\left(y,x,P^0\right)\le E\left(y,x,P^1\right)$. Conversely consider any point $(y,x)\in P^1$, which equivalently can be expressed as $E\left(y,x,P^1\right)\le 1$. When the models are ranked $E\left(y,x,P^0\right)\le E\left(y,x,P^1\right)$, we must have $E\left(y,x,P^0\right)\le 1$, which by the same equivalence implies $(y,x)\in P^0$, so that $P^1\subseteq P^0$.

***Proposition 3:*** *Let* $P^0$ *and* $P^1$ *be sets defined by a set of restrictions on the inputs and outputs* $P^0=\left\{(y,x)\in\Re_+^{K+L}\middle|g_i(y,x)\le 0,i=1,\ldots,I\right\}$ *and* $P^1=\left\{(y,x)\in\Re_+^{K+L}\middle|g_i(y,x)\le 0,i=1,\ldots,I,I+1,\ldots,I+J\right\}$ *i.e. with J added restrictions, then model 0 is ranked within model 1.*

Proof: Consider a point $(y,x)$ which satisfies all the restrictions $g_i(y,x)\le 0,i=1,\ldots,I,\ldots,I+J$ so that $(y,x)\in P^1$, then that point will also satisfy the first $I$ of these restrictions, implying $(y,x)\in P^0$. Hence all points in $P^1$ will also be in $P^0$, so that $P^1\subseteq P^0$ and by proposition 2, $E\left(y,x,P^0\right)\le E\left(y,x,P^1\right)$. Note that this does not imply nestedness unless one also assumes feasibility in model 1.

***Definition 5:*** *Given N observations of input-output vectors* $(y^n,x^n),n=1,\ldots,N$ *the DEA Constant Returns to Scale (CRS) estimator of the technology set defined over outputs* $1,\ldots,K_0$ *and inputs* $1,\ldots,L_0$ *is*

$$\hat{P}^{CRS} = \left\{ (y,x) \middle| \begin{array}{l} y_k - \sum_{n=1}^{N} \lambda_n y_k^n \leq 0, \\ \sum_{n=1}^{N} \lambda_n x_l^n - x_l \leq 0, \\ \lambda_n \geq 0, n = 1,\ldots,N, \\ k = 1,\ldots,K_0, l = 1,\ldots,L_0 \end{array} \right\} \quad \text{(A.2)}$$

***Definition 6:*** *Given N observations of input-output vectors* $(y^n, x^n), n = 1,\ldots,N$ *the DEA Variable Returns to Scale (VRS) estimator of the technology set defined over* $K_0$ *outputs and* $L_0$ *inputs is*

$$\hat{P}^{VRS} = \left\{ (y,x) \middle| \begin{array}{l} y_k - \sum_{n=1}^{N} \lambda_n y_k^n \leq 0, \\ \sum_{n=1}^{N} \lambda_n x_l^n - x_l \leq 0, \\ \sum_{n=1}^{N} \lambda_n = 1, \\ \lambda_n \geq 0, n = 1,\ldots,N, \\ k = 1,\ldots,K_0, l = 1,\ldots,L_0 \end{array} \right\} \quad \text{(A.3)}$$

Definition 5 corresponds to the original Farrell (1957) input measure as formulated by Charnes et al. (1985), while definition 6 corresponds to the formulation of Banker, Charnes and Cooper (1984). Both definititons assume feasibility, convexity and free disposal. Note that not all *possible* inputs and outputs need to be restricted in these definitions; if e.g. $L_0 < L$ then some potential inputs are assumed not to matter for this technology, and are free non-negative variables. The models may be misspecified in that the input-output list used does not correspond to the true input-output list.

***Proposition 4:*** *The DEA CRS estimator of technology* $\hat{P}^{CRS}$ *is ranked and nested within the DEA VRS estimator* $\hat{P}^{VRS}$ *when defined over the same inputs and outputs.*

Proof: Note that the restriction $\sum_{n=1}^{N}\lambda_n = 1$ can be written as $g_{I+1}(y,x) = \sum_{n=1}^{N}\lambda_n - 1 \le 0$ and $g_{I+2}(y,x) = -\sum_{n=1}^{N}\lambda_n + 1 \le 0$, so that $\hat{P}^{VRS}$ is $\hat{P}^{CRS}$ with the addition of these two restrictions. Therefore by proposition 3 $\hat{P}^{VRS} \subseteq \hat{P}^{CRS}$ and by proposition 2 $E\left(y,x,\hat{P}^{CRS}\right) \le E\left(y,x,\hat{P}^{VRS}\right)$. Since all DEA models assume the feasibility of observed points, by proposition 1, $\hat{P}^{CRS}$ is nested within $\hat{P}^{VRS}$.

***Proposition 5:*** *The DEA CRS estimator of technology $\hat{P}_{L_0}^{CRS}$ with $L_0$ inputs is ranked and nested within the DEA CRS estimator $\hat{P}_{L_1}^{CRS}$ with $L_1 > L_0$ inputs when defined over the same outputs, and the DEA VRS estimator of technology $\hat{P}_{L_0}^{VRS}$ with $L_0$ inputs is ranked and nested within the DEA VRS estimator $\hat{P}_{L_1}^{VRS}$ with $L_1 > L_0$ inputs when defined over the same outputs.*

Proof: Writing the restrictions for the extra $L_1 - L_0$ inputs as $g_{I+l-L_0}(y,x) = \sum_{n=1}^{N}\lambda_n x_l^n - x_l \le 0,\ l = L_0+1,\ldots,L_1$, then $\hat{P}_{L_1}^{CRS}$ is $\hat{P}_{L_0}^{CRS}$ with the addition of these extra restrictions. Therefore by proposition 3, $\hat{P}_{L_1}^{CRS} \subseteq \hat{P}_{L_0}^{CRS}$, and by proposition 2, $E\left(y,x,\hat{P}_{L_0}^{CRS}\right) \le E\left(y,x,\hat{P}_{L_1}^{CRS}\right)$. Finally, since the CRS estimator assumes feasibility of observations, by proposition 1, the smaller model with $L_0$ inputs is nested in the larger one with $L_1$ inputs. By the same reasoning, $\hat{P}_{L_1}^{VRS} \subseteq \hat{P}_{L_0}^{VRS}$, $E\left(y,x,\hat{P}_{L_0}^{VRS}\right) \le E\left(y,x,\hat{P}_{L_1}^{VRS}\right)$, and the smaller model is nested in the larger one.

***Proposition 6:*** *The DEA CRS estimator of technology $\hat{P}_s^{CRS}$ with input s defined as a linear aggregation of inputs t and u, $x_s = \alpha x_t + \beta x_u,\ \beta \ge 0,\ \alpha > 0$ is ranked and nested within the DEA CRS estimator $\hat{P}_{t,u}^{CRS}$ when defined over the separate inputs t and u entered directly, and the same outputs and other inputs, and the DEA VRS estimator of technology $\hat{P}_s^{VRS}$ with input s defined as a linear aggregation of inputs t and u, $x_s = \alpha x_t + \beta x_u,\ \beta \ge 0,\ \alpha > 0$ is ranked and nested within the DEA VRS estimator*

$\hat{P}^{VRS}_{t,u}$ *when defined over the separate inputs t and u entered directly, and the same outputs and other inputs.*

Proof: This is also proved by Färe and Primont (1987). Note that in the aggregated model restriction I can be written as $g_I(y,x) = \sum_{n=1}^{N} \lambda_n(\alpha x_t^n \beta + x_u^n \gamma) - (\beta x_t + x_u) \leq 0$, while in the disaggregated the two restrictions $g_{I+1}(y,x) = \sum_{n=1}^{N} \lambda_n x_t^n - x_t \leq 0$ and $g_{I+2}(y,x) = \sum_{n=1}^{N} \lambda_n x_u^n - x_u \leq 0$ are included instead. But since a positive linear aggregation of the two latter restrictions implies the aggregated restriction, $\alpha g_{I+1}(y,x) + \beta g_{I+2}(y,x) = g_I(y,x)$, the aggregated restriction can be included in the disaggregated model definition without changing the extent of the set. Then the disaggregated model $\hat{P}^{CRS}_{t,u}$ is $\hat{P}^{CRS}_{s}$ with the addition of two extra restrictions, and therefore Propositions 1,2 and 3 apply. The same reasoning holds for the VRS case.

***Proposition 7:*** *The true technology* $P^{CRS}$ *is ranked and nested within the DEA CRS estimator* $\hat{P}^{CRS}$ *if a) the latter is defined over the same true set of relevant outputs and inputs, b) the true technology has free disposability of inputs and outputs, is convex and linearly homogenous, and c) inputs and outputs are observed without error, and further that the true technology* $P^{VRS}$ *is ranked and nested within the DEA VRS estimator* $\hat{P}^{VRS}$ *if a) the latter is defined over the same true set of relevant outputs and inputs, b) the true technology has free disposability of inputs and output and is convex, and c) inputs and outputs are observed without error.*

Proof: Note that any point in $\hat{P}^{VRS}$ can be written as $\left( \sum_{n=1}^{N} \lambda_n y^n - s^y, \sum_{n=1}^{N} \lambda_n x_l^n + s^x \right)$, $\sum_{n=1}^{N} \lambda_n = 1, \lambda_n \geq 0$, where $s_k^y \geq 0, s_l^x \geq 0$ are the slack variables form the optimal solution of (A.3). Firstly, $\left(y^n, x^n\right) \in P^{VRS}$ since by the assumption of no measurement error, all observed points are feasible. Secondly, $\left( \sum_{n=1}^{N} \lambda_n y^n, \sum_{n=1}^{N} \lambda_n x_l^n \right) \in P^{VRS}$ since this is a convex

combination of the observations that are in the true technology set. Finally, $\left(\sum_{n=1}^{N}\lambda_n y^n - s^y, \sum_{n=1}^{N}\lambda_n x_l^n + s^x\right) \in P^{VRS}$ by the assumption of free disposability of inputs and outputs. Therefore $\hat{P}^{VRS} \subseteq P^{VRS}$ and $E\left(y,x,P^{VRS}\right) \le E\left(y,x,\hat{P}^{VRS}\right)$, and the true CRS technology is ranked and nested within the DEA CRS estimator.

In the same manner, any point in $\hat{P}^{CRS}$ can be written as $\left(\sum_{n=1}^{N}\lambda_n y^n - s^y, \sum_{n=1}^{N}\lambda_n x_l^n + s^x\right)$, $\lambda_n \ge 0,$ which can be reformulated as a proportionally scaled point $\Lambda\cdot\left(\sum_{n=1}^{N}\mu_n y^n - \frac{1}{\Lambda}s^y, \sum_{n=1}^{N}\mu_n x_l^n + \frac{1}{\Lambda}s^x\right),\ \sum_{n=1}^{N}\mu_n = 1, \mu_n \ge 0$ by dividing and multiplying by $\Lambda = \sum_{n=1}^{N}\lambda_n$. Since $\left(\sum_{n=1}^{N}\mu_n y^n - \frac{1}{\Lambda}s^y, \sum_{n=1}^{N}\mu_n x_l^n + \frac{1}{\Lambda}s^x\right) \in P^{CRS}$ by the assumptions of feasibility of observations, convexity and free disposal, the original points $\Lambda\cdot\left(\sum_{n=1}^{N}\mu_n y^n - \frac{1}{\Lambda}s^y, \sum_{n=1}^{N}\mu_n x_l^n + \frac{1}{\Lambda}s^x\right) \in P^{CRS}$ by the assumption of linear homogeneity. Therefore $\hat{P}^{CRS} \subseteq P^{CRS}$ and $E\left(y,x,P^{CRS}\right) \le E\left(y,x,\hat{P}^{CRS}\right)$, and the true CRS technology is ranked and nested within the DEA CRS estimator.

**References:**


Banker, R.D., A. Charnes and W.W. Cooper, 1984, Some models for estimating technical and scale inefficiencies, *Management Science* 30, 1078-1092.

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

Farrell, M.J., 1957, The measurement of productive efficiency, *Journal of the Royal Statistical Society* 120, 449-460.

Färe, R. and C.A.K.Lovell, 1978, Measuring the technical efficiency of production, *Journal of Economic Theory* 19, 150-162.

Färe, R. and D.Primont, 1987, Efficiency measures for multiplant firms with limited data, in: Eichhorn, ed., *Measurement in Economics* (Physica-Verlag, Heidelberg).

Green, W.H., 1990, *Econometric Analysis*, 2nd ed. (Prentice Hall, Englewood Cliffs).

Pesaran, M.H., 1987, Global and partial non-nested hypothesis and asymptotic local power, *Econometric Theory* 3, 69-97.

Shephard, R.W., 1970, *Theory of cost and production functions*, 2nd ed. (Princeton University Press, Princeton).