Measuring efficiency of Tunisian schools in the presence of quasi-fixed inputs: A bootstrap data envelopment analysis approach

Essid, Hédi; Ouellette, Pierre and Vigeant, Stéphane
Université des sciences et technologies de Lille

2007

Measuring Efficiency of Tunisian Schools in the Presence of Quasi-Fixed Inputs: A Bootstrap Data Envelopment Analysis Approach

Hédi Essid
Institut des HEC de Carthage
Carthage Présidence-2016 Tunis
Tunisie
E-mail : hedi.essid@ihec.rnu.tn

Pierre Ouellette
Department of Economics, Université du Québec à Montréal
PO Box 8888, Station Centre-Ville
Montréal, Québec
Canada H3C 3P8
E-mail : ouellette.pierre@uqam.ca

and
Stéphane Vigeant (corresponding author)
EQUIPPE (Universités de Lille) and CABREE (University of Alberta)
Faculté des Sciences Économiques et Sociales
Université des Sciences et Technologies de Lille
59655 Villeneuve d'Ascq Cédex,
France
Phone: +33 (0)3 20 33 63 56, Fax: +33 (0)3 20 43 66 55
E-mail: Stephane.Vigeant@univ-lille1.fr

Abstract: The objective of this paper is to measure the efficiency of high schools in Tunisia. We use a statistical Data Envelopment Analysis (DEA)-bootstrap approach with quasi-fixed inputs to estimate the precision of our measure. To do so, we developed a statistical model serving as the foundation of the Data Generation Process (DGP). The DGP is constructed such that we can implement both smooth homogeneous and heterogeneous bootstrap methods. Bootstrap simulations were used to estimate and correct the bias, and to construct confidence intervals for the efficiency measures. The simulation results show that the efficiency measures are subject to sampling variations. The adjusted measure reveals that high schools with residence services would have to give up less than 12.1 percent of their resources on average to be efficient.

JEL classification numbers: D2, I2

Keywords: Educational economics, Efficiency, Productivity, Data Envelopment Analysis, Bootstrap, Quasi-fixed inputs.

December 2007 (Revised February 2009)

1. Introduction

In this paper, we assess the performance of Tunisian high schools at allocating resources to provide education to the population. The Tunisian system being centralized, the resources are distributed by the State to high schools that manage them at the local level. In the absence of a market for secondary level education, efficiency measures can serve as an alternative method to control for performance. Specifically, efficiency measurement allows us to assess how well a decision-maker transforms inputs into outputs. In the Tunisian high school system, principals are responsible for day-to-day decisions, while investment decisions (e.g. school construction) are made centrally. Measuring efficiency as a proportional reduction in inputs implies that every input is under the control of the decision-maker. This is clearly not an appropriate model here and adjustments must be made to take into account the real choice set of the Tunisian principals.

Data Envelopment Analysis (DEA) has proved to be a good tool to measure efficiency of such institutions. DEA estimation requires weak assumptions on the underlying technology and can easily handle quasi-fixed factors such as school size (usually not under the control of the principals) making it a good fit for us. The drawback is its failure to provide confidence intervals for the estimated efficiency measures, ignoring the sensitivity of the results to sampling variations. Seminal contributions by Banker (1993) and Kneip, Park and Simar (1998), among others, introduced the statistical approach into DEA so that frontier and efficiency measures are now understood to be statistical estimators. It has been shown that the probability distribution of a DEA estimator is difficult to identify, in particular in the multivariate case (Simar and Wilson, 2000b). In this case, the bootstrap methodology appears to offer the best solution to approximate the sampling distribution of this estimator. These contributions allowed DEA users to introduce statistical induction into the interpretation of their results.

This paper combines quasi-fixed factors with a statistical approach to DEA estimation of efficiency scores to study the performance of Tunisian high schools and we provide estimates of the precision of the efficiency measures. To do so, we implement a DEA method based on the approach proposed by Banker and Morey (1986) to include the quasi-fixed inputs. The precision of the efficiency

measures is calculated using Simar and Wilson (1998, 2000a) methodology. We define a Data Generating Process (DGP) that allows us to use the smooth bootstrap methods to evaluate the estimator's bias and to construct confidence intervals for the efficiency scores. We conclude with comparisons of the results obtained under the homogeneous and heterogeneous bootstrap.

2. DEA Approach with Quasi-fixed Inputs

There are many methods to handle non discretionary factors in the DEA analysis. These methods can be grouped into two categories: (a) One-stage models: these involve only one DEA analysis in which the non-discretionary factors are directly taken into account. (This approach is based on Banker and Morey (1986).) (b) Multi-stage models: these involve several, DEA and non DEA, sequential stages through which the effect of non-discretionary factors is eliminated from the original efficiency index. Fried and Lovell (1996), Silva Portela and Thanassoulis (2001) and Muñiz (2002), among others, and recently Simar and Wilson (2007) have proposed semi-parametric models where the efficiency (obtained from DEA estimators in the first stage) are regressed on exogenous variables (second stage).[1] The theoretical difference between the two approaches is that a one-stage procedure assumes that non-discretionary factors are part of the technology, while in a multi-stage procedure these factors are assumed exogenous to the production process.

In this paper, we use Banker and Morey's model (i.e. a one-stage procedure) to deal with non-discretionary inputs, which is a common method for this type of analysis. For an input oriented DEA estimator of the frontier, we obtain a variable input requirement set, consistent with the economic intuition. We also adopt the approach developed by Kneip, Park and Simar (1998) to develop a statistical model that includes quasi-fixed inputs and we use this model to characterize the DGP, thus justifying the use of bootstrap methods in DEA analysis.

We suppose that all inputs and outputs are continuous variables and there are two types of inputs: variable (or discretionary) under the direct control of the decision maker and quasi-fixed (non discretionary) not under the control of the manager at decision time.

2.1 The Frontier Model with Quasi-fixed Inputs

Consider a production process using variable inputs $x=\{x_i, i=1,\ldots,m_1\}$ and quasi-fixed inputs $z=\{z_j, j=1,\ldots,m_2\}$ to produce an output vector $y=\{y_r, r=1,\ldots,s\}$. The production possibility set is given by:

$$\Psi=\left\{(x,z,y)\in\mathbb{R}_+^{m_1+m_2+s} \middle| (x,z,y) \text{ is feasible}\right\}. \quad (1)$$

The efficiency of a DMU is measured by the distance between the observed input-output mix from the optimal mix located on the frontier of Ψ. By choosing a direction to approach the frontier, the true input oriented efficiency measure in the sense of Farrell (1957) is defined to be a triplet $(x,z,y)\in\Psi$ satisfying:

$$\theta(x,z,y)=\min\left\{\theta \middle| (\theta x,z,y)\in\Psi\right\}. \quad (2)$$

The scalar θ is interpreted as the maximal proportion by which the input vector *x* that produces *y* can be shrank so that it still produces *y* given the quasi-fixed input vector *z*. Therefore, the efficient input level is $x^{\partial}(z,y)=\theta(x,z,y)x$ and the efficient quantity of inputs is a proportion of the observed input quantities.

2.2 The DEA estimator with quasi-fixed inputs

Consider a sample (of DMUs) of size *n* defined as $\{(x_j,z_j,y_j), j=1,\ldots,n\}\subseteq\Psi$. The DEA estimator of Ψ is given by:

$$\hat{\Psi}_{DEA}=\left\{(x,z,y)\in\mathbb{R}_+^{m_1+m_2+s} \middle| x\geq\sum_{j=1}^{j=n}\lambda_j x_j, z\geq\sum_{j=1}^{j=n}\lambda_j z_j, y\leq\sum_{j=1}^{j=n}\lambda_j y_j, \sum_{j=1}^{j=n}\lambda_j=1\right\}, \quad (3)$$

The estimated input oriented efficiency measure of the triplet (x,z,y) is $\hat{\theta}(x,z,y)=\min\left\{\theta \middle| (\theta x,z,y)\in\hat{\Psi}_{DEA}\right\}$ with $0<\hat{\theta}(x,z,y)\leq 1, \forall(x,z,y)\in\Psi$ and the efficient variable input bundle is $\hat{x}^{\partial}(z,y)=\hat{\theta}(x,z,y)x$. By construction, we have that, $\hat{\Psi}\subseteq\Psi$ and $\theta(x,z,y)\leq\hat{\theta}(x,z,y),\ \forall(x,z,y)\in\Psi$.

3. The Statistical Model and the Bootstrap Method

3.1 The Data Generating Process (DGP)

Given an output level, a stock of quasi-fixed inputs and an efficiency parameter, the stochastic content of the production process is completely characterized by identifying the variable inputs to a vector of random variables. When the production process is not efficient, *x* is not on the frontier of the variable input set. The specification of the DGP makes this explicit: for a given true frontier, the vector *x* is a random variable along the ray through the origin defined by $\left\{\theta x \middle| (\theta x, z, y) \in \Psi\right\}$. For this DGP, any particular combination (x_j, z_j, y_j) can be generated. That is, for decision making unit j we have $(x_j, z_j, y_j) = (x^{\partial}(z_j, y_j)/\theta_j, z_j, y_j)$ where the efficient variable input level, $x^{\partial}(z_j, y_j)$, is unknown but can be interpreted as a "parameter" to be estimated.

Now, suppose instead that the efficiency measure, $\theta_j \in]0,1]$, is a random variable with a probability measure admitting a density $f(\bullet)$, then the DGP $\Im_j$ generating x_j, conditionally on the output y_j, the quasi-fixed inputs z_j, and a proportion of input observed is equivalently characterized by $x^{\partial}(z_j, y_j)$ and *f*. That is, $\Im_j = (x^{\partial}(z_j, y_j), f), j = 1, \ldots, n$ or $\Im = (\Psi, f)$.

3.2 A consistent estimator of the DGP

To find a consistent estimator of the DGP $\Im = (\Psi, f)$ is equivalent to find a consistent estimator of its components: the production set Ψ and the density *f*. Based on Kneip, Park & Simar (1998), Essid, Ouellette and Vigeant (2007) have shown that Banker and Morey's estimator, given by equation (3), is a consistent estimator of the production set. Thus, we only need to obtain a consistent estimator of the density of the θs. In this paper, we use two approaches to estimate the density: the homogenous and heterogeneous bootstrap methods.

To estimate the density when the efficiency structure is homogenous, that is $f(\theta|\eta, z, y) = f(\theta)$, we smooth the probability density *f* with a kernel as in Simar and Wilson (1998).[2] A smooth estimator can be obtained from the Gaussian kernel.

Under this simple form, it can be shown that the estimator of the density is not consistent in the neighborhood of one. To correct the bias, we follow the suggestion in Simar and Wilson (1998) and we use the reflection method developed by Schuster (1985) and Silverman (1986). The method consists in reflecting each estimate of the efficiency measure $\hat{\theta}_j \leq 1$ with its image, given by $2-\hat{\theta}_j \geq 1$. The kernel estimator is then evaluated on the basis of *2n* observations and is defined as follows:

$$\hat{f}^c(t) = \begin{cases} 2\hat{g}(t) & \text{if } t \leq 1 \\ 0 & \text{otherwise} \end{cases}, \text{ where } \hat{g}(t) = \frac{1}{2nh}\sum_{j=1}^{j=n}\left[\phi\left(\frac{t-\hat{\theta}_j}{h}\right)+\phi\left(\frac{t-2+\hat{\theta}_j}{h}\right)\right] \tag{4}$$

The bandwidth *h* is set following the normal reference rule (Silverman (1986)).

The homogeneity assumption might be too restrictive, so we use the heterogeneous bootstrap to handle the possibility that the efficiency score θ and (η, z, y) are not independent, as we assumed above. The simulations use a multivariate Gaussian kernel to estimate the density $f(\theta, \eta, z, y)$. It is assumed that the bandwidth matrix is diagonal with only one parameter, *h*. The support of $f(\theta, \eta, z, y)$, given by $\Omega =]0,1] \times [0, \pi/2]^{m_1 - 1} \times \mathbb{R}_+^{m_2} \times \mathbb{R}_+^{s}$, is bounded, and so the estimator $\hat{f}(\theta, \eta, z, y)$ is not consistent in the neighborhood of the boundaries. Schuster-Silverman reflection method is used to correct the bias.[3] The generalization of the procedure proposed by Simar and Wilson (2000a) to the case of quasi-fixed inputs is as follows:

Let $P = \left[y_j \quad z_j \quad \eta_j \quad \hat{\theta}_j \right]$ where $j = 1, \ldots, n$, be the matrix of observations. The j^{th} line of *P* contains observations written in polar coordinates on the j^{th} DMU. Let $P_R = \left[y_j \quad z_j \quad \eta_j \quad 2-\hat{\theta}_j \right]$ be the matrix of the points reflected in the neighborhood of one. In that case, the $2n \times (m_1 + m_2 + s)$ data matrix is given by:

$$\tilde{P} = \begin{bmatrix} P \\ P_R \end{bmatrix} \tag{5}$$

We use the matrix $\tilde{P}$ to construct a bias corrected estimator of *f*. Let $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ be estimators of the covariance matrix of *P* and P_R, respectively and partition them as follows:

$$\hat{\Sigma}_1 = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \text{ and } \hat{\Sigma}_2 = \begin{bmatrix} S_{11} & -S_{12} \\ -S_{21} & S_{22} \end{bmatrix} \tag{6}$$

where S_{11} is the covariance matrix of (y, z, η), S_{22} is the variance of $\hat{\theta}$ and $S_{12} = S_{21}^T$ is the vector of the covariance between (y, z, η) and $\hat{\theta}$ ((y, z, η) and $2-\hat{\theta}$ for $\hat{\Sigma}_2$). As in Simar and Wilson (2000a), we use Campbell's M-estimator method (Campbell, 1980) to obtain $\hat{\Sigma}_1$ and then $\hat{\Sigma}_2$. The estimator of the density, *f*, is defined as follows:

$$\hat{f}^c(u) = \begin{cases} 2\hat{f}(u) & \text{if } u \in \Omega \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where $\hat{f}(u) = \frac{1}{2nh^{m_1+m_2+s}} \sum_{j=1}^{j=n} \left[K_1\left(\frac{u-u_j}{h}\right) + K_2\left(\frac{u-u_{R_j}}{h}\right) \right]$ with $u_j = (\theta_j, \eta_j, z_j, y_j)$ and $u_{R_j} = (2-\theta_j, \eta_j, z_j, y_j)$, and $K_l(\bullet)$ is the probability density of a normal vector with zero mean and variance-covariance matrix $\hat{\Sigma}_l$, $l = 1, 2$. To calculate the bandwidth parameter we use once again Silverman's normal rule. The complete algorithms for the smooth homogeneous bootstrap and smooth heterogeneous bootstrap methods with quasi-fixed inputs are presented in Appendix A.

Let $\hat{\Im} = (\hat{\Psi}, \hat{f})$ be a consistent estimator of DGP $\Im$, generated as above. The estimator $\hat{\theta}(x_j, z_j, y_j)$ of $\theta(x_j, z_j, y_j)$ obtained from the original sample generated by $\Im$, has an unknown sampling distribution, but we can implement a bootstrap procedure to find an approximation of this distribution. This is done by generating *B* samples, Ψ_b^*, $b = 1, \ldots, B$ of size *n* and using the DEA method to obtain *B* pseudo-estimators $\left\{\hat{\theta}_b^*(x_0, z_0, y_0),\ b = 1, \ldots, B\right\}$ for all $(x_0, z_0, y_0) = (x_j, z_j, y_j), j = 1, \ldots, n$. Then, the empirical distribution of those pseudo-values

provides a Monte Carlo approximation of the sampling distribution of $\hat{\theta}^*\left(x_0, z_0, y_0\right)$ given the estimator $\hat{\Im}$.

3.3 Bootstrap and bias corrections

Even though the DEA estimator is consistent, it is also biased. The bootstrap bias estimator is defined by $\widehat{bias_B}\left(\hat{\theta}_j\right) = \overline{\hat{\theta}_j^*} - \hat{\theta}_j \quad \forall j = 1, \ldots, n$, where $\overline{\hat{\theta}_j^*} = (1/B)\sum_{b=1}^{b=B} \hat{\theta}_{bj}^*$. A bias corrected DEA estimator is $\hat{\hat{\theta}}_j = \hat{\theta}_j - bias\left(\hat{\theta}_j\right) = 2\hat{\theta}_j - \overline{\hat{\theta}}_j^* \quad \forall j = 1, \ldots, n$. However, this correction introduce a new noise (Efron and Tibshirani, 1993) leading to the possibility that the standard error of the corrected estimator $\hat{\hat{\theta}}_j$ be larger than the standard error of the original estimator $\hat{\theta}_j$. Consequently, the correction is applied only when $r_j = \left(bias_B\left(\hat{\theta}_j^*\right)\right)^2 / 3\hat{\sigma}_{\hat{\theta}_j^*}^2 > 1$ for all $j = 1, \ldots, n$, where $\hat{\sigma}_{\hat{\theta}_j^*}^2 = (1/B)\sum_{b=1}^{b=B}\left(\hat{\theta}_{bj}^* - \overline{\hat{\theta}_j^*}\right)^2$.

3.4 Bootstrap confidence intervals

To construct the confidence intervals of the efficiency scores, we start from a procedure based on the estimation of the bias. Once again, this introduces an additional noise in the confidence interval estimation. To take this additional noise into account, Simar and Wilson (2000a, b) proposed a method that consists in finding values, a_α and b_α such that:

$$\Pr\left(-b_\alpha \le \hat{\theta}\left(x_0, z_0, y_0\right) - \theta\left(x_0, z_0, y_0\right) \le -a_\alpha\right) = 1 - \alpha . \tag{8}$$

The estimators of the bounds a_α and b_α are obtained from the empirical bootstrap distribution of the pseudo-estimators $\left\{\hat{\theta}_{jb}^*, b = 1, \ldots, B\right\}$ satisfying

$$\Pr\left(-\hat{b}_\alpha \le \hat{\theta}^*\left(x_0, z_0, y_0\right) - \hat{\theta}\left(x_0, z_0, y_0\right) \le -\hat{a}_\alpha \middle| \hat{\Im}\right) = 1 - \alpha , \tag{9}$$

where $1-\alpha$ is the size of the confidence interval. To obtain $\hat{a}_\alpha$ and $\hat{b}_\alpha$ we sort $\left(\hat{\theta}_b^*\left(x_0, z_0, y_0\right) - \hat{\theta}\left(x_0, z_0, y_0\right)\right), b = 1, \ldots, B$ in ascending order and then we eliminate $\left(0.5\alpha \times 100\right)$

percent of the elements on the right and the left of the sorted list. The values $\hat{a}_\alpha$ and $\hat{b}_\alpha$ correspond to the left and right limits of the truncated series, with $\hat{a}_\alpha \leq \hat{b}_\alpha$. Consequently, the bootstrap approximation of (8) is:

$$\Pr\left(-\hat{b}_\alpha \leq \hat{\theta}(x_0, z_0, y_0) - \theta(x_0, z_0, y_0) \leq -\hat{a}_\alpha\right) \approx 1-\alpha . \tag{10}$$

The estimated $(1-\alpha)$-percent confidence interval is then

$$\hat{\theta}(x_0, z_0, y_0) + \hat{a}_\alpha \leq \theta(x_0, z_0, y_0) \leq \hat{\theta}(x_0, z_0, y_0) + \hat{b}_\alpha . \tag{11}$$

4. Data

In Tunisia, schooling is divided into two steps: basic learning for the first nine years (576,088 students in 2004-05), then secondary education for four years (508,790 students in 2004-05). At the end of these thirteen years, each student writes the *baccalaureat* exam. (The *baccalauréat* is the grade obtained at the end of high school; it is the equivalent of a high school diploma. In recent years, the average success rate has been about 70% of the 65,000 student taking the exam).[4] Recently, enrolment at the secondary level has substantially increased, 33% between 2000/01 and 2004/05. This trend has been accompanied by a larger involvement of the State to satisfy human and material needs (new schools were built and hiring has increased).

The administration of secondary teaching is centralized at the level of the Department of education. The Department of education creates the programs and determines the pedagogical content of these programs, it hires the teachers and administrative staff and dispatch them based on the estimated needs of the schools and finally it allocates the operating budget between the different institutions according to some general planning established by the government. Note that the larger share of high schools' budget comes from State subsidies (85% of the budget). The private sector, local government and households account for the remaining share. As a consequence, schools are very vulnerable to budget changes orchestrated by the State. The high schools implement the general rules emanating from the Department of education on matters related to the programs and their content. The internal management requires, at the beginning of every academic year, that each

school determines their human resource needs (teaching and administrative staffs), provides an estimate of their operating budget requirement, and estimates the number of classrooms and laboratories needed for the year. The Department of education then tries to satisfy those requirements within the limits of its own budget.

In this study, we consider that a high school is a multi-output firm and each output is associated to a service to be evaluated. We are trying to identify and evaluate how the material and physical means are used to produce the different services generated by the school.

4.1 Measurement of the outputs

The output of the learning activity is the result of the standard exams at the end of the last year of high school (RESBAC). It is a very rough measure because it does not take into account the admission conditions for some specific programs. However, since the data concerning the students' performance at the beginning and at the end of their program are not available and since we cannot address the problems related to the identification of the shares of the learning attributable to the family and to the external environment in the school grades, we merely say that these standard exam results are an approximation of the value added to the student in the schooling system.

The number of students enrolled in the school (STUDENTS) is used as a second output and serves two purposes. First, it is an indicator of the volume of the high school's activity. Second, the number of students enrolled shows also that some value is added to the student independently from the fact that he or she may not graduate. Thus, it is also an indicator of the value added to students not completing their degree.[5]

High schools also supply complementary services. In Tunisia, the residence service cannot be separated from the teaching activities, mostly for high schools located in rural regions. This activity is measured using the number of beds (BEDS) and meals served (MEALS).

4.2 Measurement of the inputs

The production of the outputs is done using human resources and materials. Factors that are not related directly to human resources can be difficult to measure quantitatively. The variable inputs

used in this study are the number of teachers (TEACHERS), the administrative and supporting staff (ADM), the technical staff and janitors (BLUECOL), and an index representing the material and office supplies (desk, stationery, equipment, furniture, etc.). The latter variable is not directly observed. To obtain the quantity of input effectively used by the school we construct a quantity index defined as the ratio of the budget for these categories and the consumer price index, to which we add the food expenditures to capture the food and accommodation service inputs.[6] We use this new variable to proxy the materials used (F&MAT). Since we do not have data on the building used for the residences, we have used only two quasi-fixed inputs: the number of general classrooms (GENROOM) and the number of specialized classrooms (SPECROOM).[7]

The data used come from two sources. The data for the academic year 2003/04, and for almost all high schools are from the "*Bureau des études, de la planification et de la programmation*" of the Tunisian Department of Education. The National Statistical Institute of Tunisia has provided the consumer price index (CPI) for the year 2004 (base year 2000). We have 166 institutions in our database for the academic year 2003-2004. Descriptive statistics are found in Table 1.

INSERT TABLE 1 HERE

5. Results

The simulation results are summarized in Tables 2, 3, 4 and 5, while the complete results are available in Appendix B (). These results are obtained using a SAS algorithm with 1000 replications.[8] A look at Table 2 shows that before the bootstrap, almost 45% (75) of the DMUs are efficient, ($\hat{\theta} = 1$) and are therefore located on the estimated frontier. The practices of those efficient units are thus the reference for the high schools not considered efficient. In other words, the inefficient institutions use too much material and/or staff when compared to similar institutions located on the frontier. The scores of the inefficient high schools range between 0.716 and 0.998. The high school with the worst performance must give up almost 29% of the resources it uses to reach an efficient point. The least inefficient high school is not far from the best practices, however. This sums up the traditional interpretation of the results of an efficiency study using DEA.

However, as we will now show, such an interpretation can be misleading for the decision maker that allocates the resources at the Department of Education. We find that these results show a strong sensitivity to sampling variations and this tells us that we cannot compare the initial DEA results between DMUs freely.

We also present in Table 2 a summary of both bootstrap simulations. The simulations did not change the proportion of efficient units; it is still equal to 45%. After correction for the bias, the distribution of the scores is larger; the average is lower by one percentage point for both bootstrap procedures, while the standard deviation increased to 0.098 from 0.069 in the case of the homogenous bootstrap and to 0.101 in the heterogeneous bootstrap case. It is also noticeable that the results obtained with both bootstrap procedures are very similar, showing the robustness of the analysis.

[INSERT TABLE 2 HERE]

Table 3 presents the distribution of the bias. It is non-negative in 117 cases for the homogenous bootstrap with an average value equal to 0.0288. Thirty five schools are efficient in all simulations and have a null bias. These results would confirm that the DEA estimator tends to over-estimate the real efficiency score. As noted above the correction is not made in all cases to avoid that the quadratic error of the corrected estimates become larger than the one of the original estimates. The correction is made for 12 high schools only. In most cases, the correction is not trivial and can be as large as 0.224. In the case of the heterogeneous bootstrap, the bias is corrected for 17 high schools and in some cases the correction is large, as it reaches 0.236. Contrary to what was expected, the average bias is negative and equal to -0.016 for the heterogeneous bootstrap.

[INSERT TABLE 3 HERE]

We present confidence interval estimation in Table 4. Confidence interval estimation increases the proportion of efficient units. In the case of the homogenous bootstrap, 107 schools have observed scores not significantly different than one at size equal to 95%. This number goes up to 119 with the

heterogeneous bootstrap. Consequently, based on the confidence interval inference, in both sets of simulations, high schools must give up 12.1% of their resources on average to be efficient.

[INSERT TABLE 4 HERE]

Specific examples are reported in Table 5 to shed some light on the contribution of the bootstrap procedure to the statistical content of the efficiency scores. For example, the difference between the initial DEA score of L1461 and L1462 is 0.019, a fairly small magnitude. A comparison of both high school bias corrected scores reveals a substantial difference under the homogenous bootstrap (0.146) and smaller but significant one under the heterogeneous bootstrap (0.043).

The bootstrap simulations allow us to parallel the standard results of the classical theory on confidence intervals. For example, L42108 and L42114 are not efficient even after the bias correction. However, their respective confidence interval contains the value one (under both bootstrap methods). This means that the observed inefficiency is very likely due to sampling variations and not real. Therefore, the efficiency is said to be perverted by sampling variations. This parallels the standard t-test for an estimated parameter for which the value is tested to be equal to one under the null hypothesis. We are not able to reject the hypothesis that the DMUs are efficient. We are also conducted to revise the resources an institution must give up to become efficient.

Another consequence of the sampling variation is that we have to compare institutions using a statistical reasoning. Looking at our results, we observe that the confidence intervals of many institutions overlap. In other words, it is often possible that two initial scores that appeared to differ are in fact in the same confidence region. For example, high schools L1461 and L1462 are in this situation. The initial efficiency scores are not identical but the difference is not statistically significant, based on the confidence interval. This shows that a comparison of the initial efficiency scores for these high schools is not appropriate and may lead us to wrong conclusions.

[INSERT TABLE 5 HERE]

An analysis of the precision of the estimators of the efficiency scores based on box plots (see Appendix C()) reveals two things. More than half of the efficient high schools keep generating

efficiency scores equal to one throughout the simulation process. The bias for these institutions is simply equal to zero. These schools can be interpreted as dominant. The efficiency measure of these institutions is characterized by a very high precision. The length of the boxes allows us to compare the dispersion of the values of the efficiency parameter generated by the simulation process. We note that the dispersion is not constant between schools leading to the conclusion that the precision of the efficiency rate estimator is not homogenous across schools.

Finally, since we cannot calculate the optimal bandwidth in most cases, it is important to assess the sensitivity of the results to the choice of bandwidth. To do this, we have repeated the simulations with bandwidth taking values $0.5h$ and $1.5h$ and we have recalculated all confidence intervals for each simulation experiment.[9] Comparisons of the results show that there is only a marginal difference between the confidence bounds, confirming the robustness of our results to the value of the bandwidth.

6. Conclusion

We have evaluated the statistical precision of efficiency measures for Tunisian high schools calculated with the DEA method. The paper shows how to use the bootstrap to estimate the bias of the efficiency measure estimator, how to estimate the sampling distribution and how to calculate the confidence intervals for each school. These results allow the decision maker to check the reliability and robustness of the efficiency measures. This also shows that sound statistical inference on the performance of schools with limited data, as it is often the case in developing countries, is possible.

Our results prove the sensitivity of the standard DEA estimation to sampling variations and also confirm the proposition that DEA estimators tend to overestimate the real rate of efficiency in the case of the homogeneous bootstrap but it is not as clear cut in the case of the heterogeneous bootstrap. In both sets of simulations we have to conclude that on average the high school with a residence service must give up less then 12.1% of their resource to reach the frontier. In other words, it is possible to consider that this type of high school is fairly efficient overall.

References

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

Banker, R.D. and Morey, R.C. (1986). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. *Operations Research*, 34, 513-521.

Campbell N.A. (1980). Robust Procedures in Multivariate Analysis I: Robust Covariance. *Applied Statistics*, 29, 231-237.

Cohn, E. Millman, S.D. and Chew, I.K. (1975). *Input-Output Analysis in Public Education*. Ballinger, Cambridge, MA.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Essid, H., Ouellette, P. and Vigeant, S. (2007). *A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores with Quasi-Fixed Inputs*. mimeo, Institut des HEC de Carthage.

Farrell, M.J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society* A 120, 253-290.

Fried, H.O. and Lovell, C.A.K. (1996). Searching the Zeds, *Working Paper presented at II Georgia Productivity Workshop*.

Hanushek, E.A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24, 1141-1176.

Johnes, J., (2006). Data Envelopment Analysis and its Application to the Measurement of Efficiency in Higher Education. *Economics of Education Review*, 25, 273–288.

Kneip, A., Park, B.U. and Simar, L. (1998). A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores. *Econometric Theory*, 14, 783-793.

Muñiz, M.A. (2002). Separating Managerial Inefficiency and External Conditions in Data Envelopment Analysis. *European Journal of Operational Research*, 143, 625-643.

Ouellette, P. and Vierstraete, V. (2005). Technological Change and Efficiency in the Presence of Quasi-Fixed Inputs: A DEA Application to the Hospital Sector. *European Journal of Operational Research*, 154 (3), 755-763.

Schuster, E.F., (1985). Incorporating Support Constraints into Nonparametric Estimators of Densities. *Communications in Statistics-Theory and Method*, 14, 1123-1136.

Silva Portela and Thanassoulis E. (2001). Decomposing School and School-type Efficiency. *European Journal of Operational Research*, 132, 357-373.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Simar, L. and Wilson, P.W. (2007). Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes. *Journal of Econometrics*, 136, 31-64.

Simar, L. and Wilson, P.W. (2000a). A General Methodology for Bootstrapping in Non-Parametric Frontier Models. *Journal of Applied Statistics*, 27, 779-802.

Simar, L. and Wilson, P.W. (2000b). Statistical Inference in Nonparametric Frontier Models: The State of the Art. *Journal of Productivity Analysis*, 13, 49-78.

Simar, L. and Wilson, P.W. (1998). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management Science*, 44, 49-61.

Appendix A: Algorithm for the smooth bootstrap with quasi-fixed inputs

A.1 Homogenous bootstrap

Step 1: Compute $\hat{\theta}_j = \hat{\theta}_j\left(x_j, z_j, y_j\right) \ \forall\, j = 1,\ldots,n$.

Step 2: Generate smoothed resampled pseudo-efficiencies as follows. First generate $\left\{\rho_j^*, j = 1,\ldots,n\right\}$ by resampling with replacement a sample of size *n*, from the empirical distribution $\left\{\hat{\theta}_j, j = 1,\ldots,n\right\}$. Then generate the sequence $\left\{\tilde{\rho}_j^*, j = 1,\ldots,n\right\}$ as follows:

$$\tilde{\rho}_j^* = \begin{cases} \rho_j^* + h\varepsilon_j^* & \text{if } \left(\rho_j^* + h\varepsilon_j^*\right) \le 1 \\ 2 - \left(\rho_j^* + h\varepsilon_j^*\right) & \text{otherwise} \end{cases}, \text{ where } \varepsilon_j^* \sim N(0,1).$$

Then, generate the pseudo-efficiencies γ_j^* for all $j=1,\ldots,n$ using $\gamma_j^* = \bar{\rho}^* + \left(\tilde{\rho}_j^* - \bar{\rho}^*\right)/\sqrt{1+h^2/\hat{\sigma}_{\hat{\theta}}^2}$, where $\bar{\rho}^* = (1/n)\sum_{j=1}^{n}\rho_j^*$.

Step 3: Compute the pseudo variable inputs, $x_j^* = (1/\gamma_j^*)\hat{\theta}_j x_j \;\; j=1,\ldots,n$

Step 4: Compute the bootstrapped efficiency measures $\hat{\theta}_j^*, j=1,\ldots,n$ using the pseudo variable inputs based on the following program:

$$\hat{\theta}^*\left(x_0,z_0,y_0\right) = \min\left\{\theta | \theta x_0 \geq \sum_{j=1}^{j=n}\lambda_j x_j^*,\; z_0 \geq \sum_{j=1}^{j=n}\lambda_j z_j, y_0 \leq \sum_{j=1}^{j=n}\lambda_j y_j, \sum_{j=1}^{j=n}\lambda_j = 1,\; \lambda_j \geq 0\right\}.$$

Step 5: Repeat steps 2-5 *B* times to obtain *B* efficiency measures for every DMU *j*, $\left\{\hat{\theta}_{bj}^*, j=1,\ldots,n, b=1,\ldots,B\right\}$.

A.2 Heterogeneous bootstrap

Step 1: Compute $\hat{\theta}_j = \hat{\theta}_j\left(x_j,z_j,y_j\right) \;\forall\, j=1,\ldots,n$.

Step 2: Transform the variable inputs $x_j \;\forall j=1,\ldots,n$, expressed in Cartesian coordinates into polar coordinates and build the matrix $\tilde{P}$ as in (5).

Step 3: Compute the estimated variance-covariance matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ as in (6). Then, calculate L_1 and L_2 such that $\hat{\Sigma}_1 = L_1 L_1^T$ and $\hat{\Sigma}_2 = L_2 L_2^T$ using a Cholesky decomposition.

Step 4: Draw with replacement *n* lines from the matrix $\tilde{P}$. Denote this new matrix $\tilde{P}^*$ and compute the mean of each column of this matrix. The result is the $1\times\left(s+m_1+m_2\right)$ vector $\bar{P}^*$.

Step 5: Use a pseudo-random number generator to generate a $n\times\left(s+m_1+m_2\right)$ matrix ε from a standard *i.i.d.* normal distribution. Let ε_j be the j^{th} line of this matrix. Compute the $n\times\left(s+m_1+m_2\right)$ matrix ε^* where the j^{th} line is the vector $\varepsilon_j^* = L_l\varepsilon_j$ with *l*=1,2 and such that if the j^{th} line of the matrix $\tilde{P}^*$ is among the lines of the matrix *P*, then $\varepsilon_j^* = L_1\varepsilon_j$ and if the j^{th} line of the matrix $\tilde{P}^*$ is among lines of the matrix P_R, then $\varepsilon_j^* = L_2\varepsilon_j$.

Step 6: Compute the $n\times(s+m_1+m_2)$ matrix $\Gamma=\left(1+h^2\right)^{-1/2}\left(M\tilde{P}^*+h\varepsilon^*\right)+i_n\otimes\overline{P}^*$, where $M=I_n-(1/n)i_ni_n^T$, $i_n^T=[1\ldots1]$ and $\otimes$ is the Kronecker product.

Step 7: Partition the matrix Γ into four blocks, $\underset{n\times(s+m_1+m_2)}{\Gamma}=\left[\underset{n\times s}{\Gamma1}\quad\underset{n\times m_2}{\Gamma2}\quad\underset{n\times(m_1-1)}{\Gamma3}\quad\underset{n\times1}{\Gamma4}\right]$, with $\Gamma1=\left(\gamma_j^1\right)$, $\Gamma2=\left(\gamma_j^2\right)$, $\Gamma3=\left(\gamma_j^3\right)$, and $\Gamma4=\left(\gamma_j^4\right)$ for all $j=1,\ldots,n$. Then define the j^{th} line of the pseudo-value bootstrap matrix T^*, of size $n\times(s+m_1+m_2)$, as follows:

$$t_j^*=\begin{cases}\left(\gamma_j^1,\gamma_j^2,\gamma_j^3,\gamma_j^4\right) & \text{if } \gamma_j^4\le1\\ \left(\gamma_j^1,\gamma_j^2,\gamma_j^3,2-\gamma_j^4\right) & \text{otherwise}\end{cases}.$$

Step 8: Convert the polar coordinates of T^* back into Cartesian coordinates as follows: Use the j^{th} line of $\Gamma3$ to construct the matrix $\tilde{X}_{n\times m}=\left(x_{1j},x_{1j}tg\gamma_{1j}^3,x_{1j}tg\gamma_{2j}^3,\ldots,x_{1j}tg\gamma_{(m_1-1)j}^3\right)$ for all $j=1,\ldots,n$, then calculate $\tilde{\theta}\left(\tilde{x}_0,\gamma_0^2,\gamma_0^1\right)$ for all $\left(\tilde{x}_0,\gamma_0^2,\gamma_0^1\right)=\left(\tilde{x}_j,\gamma_j^2,\gamma_j^1\right),j=1,\ldots,n$ using the following program:[10]

$$\tilde{\theta}\left(\tilde{x}_0,\gamma_0^2,\gamma_0^1\right)=\min\left\{\theta|\theta\tilde{x}_0\ge\sum_{j=1}^{j=n}\lambda_j\tilde{x}_j,\gamma_0^2\ge\sum_{j=1}^{j=n}\lambda_j\gamma_j^2,\gamma_0^1\le\sum_{j=1}^{j=n}\lambda_j\gamma_j^1,\sum_{j=1}^{j=n}\lambda_j=1,\lambda_j\ge0\right\}.$$

This step implies solving *n* optimization programs. If the program has no immediate solution, repeat steps 4-7 until it works. Then, the pseudo-variable input vector is given by $x_j^*=\left(\tilde{\theta}\left(\tilde{x}_j,\gamma_j^2,\gamma_j^1\right)/\theta_j^*\right)\tilde{x}_j,\ j=1,\ldots,n$.

Step 9: Given $\left(x_j^*,\gamma_j^2,\gamma_j^1\right)$, compute $\hat{\theta}^*\left(x_0^*,\gamma_0^2,\gamma_0^1\right)$ for all $\left(x_0^*,\gamma_0^2,\gamma_0^1\right)=\left(x_j^*,\gamma_j^2,\gamma_j^1\right)\ j=1,\ldots,n$, using the following program for $j=1,\ldots,n$:

$$\hat{\theta}^*\left(x_0^*,\gamma_0^2,\gamma_0^1\right)=\min\left\{\theta|\theta x_0^*\ge\sum_{j=1}^{j=n}\lambda_jx_j^*,\ \gamma_0^2\ge\sum_{j=1}^{j=n}\lambda_j\gamma_j^2,\gamma_0^1\le\sum_{j=1}^{j=n}\lambda_j\gamma_j^1,\sum_{j=1}^{j=n}\lambda_j=1,\lambda_j\ge0\right\}.$$

Step 10: Repeat steps 4-9 *B* times to obtain the bootstrapped estimators of the efficiency measures for each DMU *j*: $\left\{\hat{\theta}_{bj}^*,j=1,\ldots,n,b=1,\ldots,B\right\}$.

Table 1: Descriptive Statistics:
High School with Residence in 2003-04 (166 High Schools)

Variable	Average	Standard Error	Minimum	Maximum
STUDENTS	1293.35	470.61	346	2769
BEDS	247.3	200.92	0	931
MEALS	346.71	191.37	16	931
TEACHERS	72.4	23.82	26	145
ADM	11.47	4.85	2	28
BLUECOL	17.78	7.62	5	48
F&MAT	916.78	335.58	322.86	1983.86
GENROOM	26.84	8.74	11	59
SPECROOM	10,71	4.08	3	24
RESBAC	172.03	88.24	39	526

Table 2: Distribution of the Efficiency Scores Before and After the Bootstrap

	Before the bootstrap	After the bootstrap	
		Homogenous	Heterogeneous
Number of efficient units	75	75	75
Total units			
Mean	0.939	0.928	0.925
Standard deviation	0.069	0.098	0.101
Min	0.716	0.491	0.480
Max	1	1	1
Inefficient units			
Mean	0.889	0.870	0.864
Standard deviation	0.058	0.101	0.101
Min	0.716	0.491	0.480
Max	0.998	0.998	0.998

Table 3: Distribution of the Bias

	Homogenous bootstrap	Heterogeneous bootstrap
Number of units with positive bias	117	65
With bias equal to 0	35	0
Number of corrected scores	12	17
Mean	0.028	-0.016
Standard deviation	0.052	0.089
Min	-0.040	-0.202
Max	0.224	0.236

Table 4: 95% Confidence Interval Estimation

	Homogenous bootstrap		Heterogeneous bootstrap	
Number of efficient units	107		119	
	Lower limit	Upper limit	Lower limit	Upper limit
Mean	0.879	0.993	0.879	1.103
Standard deviation	0.139	0.098	0.139	0.167
Min	0.433	0.621	0.433	0.600
Max	1	1.177	1	1.415

Table 5: Specific Examples of Confidence intervals

School	Original Score	Homogenous bootstrap			Heterogeneous bootstrap.		
		Corrected Score	95 % conf. int.		Corrected Score	95 % conf. int.	
L1461	0.817	0.817	0.634	0.830	0.689	0.634	0.822
L1462	0.798	0.671	0.596	0.801	0.646	0.596	0.829
L42108	0.949	0.949	0.899	1.073	0.949	0.899	1.130
L42114	0.976	0.976	0.952	1.117	0.976	0.952	1.124

Acknowledgments: We would like to thank Frédéric Brousseau for considerable help with the SAS programs used in for the estimations in this paper. Participants at EWEPA-X gave us useful feedbacks on an earlier version of the paper.

[1] We think that this type of models is more adapted to explain efficiency than to estimate it. For this reason we will not consider them in this study.

[2] Because Farrell's measure is radial, we are allowed to write the input vector x in polar coordinates. That is, the modulus of x is $\omega = \omega(x) = \|x\| = \sqrt{x^T x}$ and the angle is $\eta = \eta(x) \in [0, \pi/2]^{m_1 - 1}$ where $\eta = (\eta_1 \ldots \eta_i \ldots \eta_{m_1 - 1})$. This allows us to write the density as $f(x, z, y) = f(\omega, \eta, z, y)$ and the correspondence between the modulus and θ follows because the efficiency measure is radial.

[3] As in Simar & Wilson (2000a) we restrict the reflection of θ to the values in the neighbourhood of one. This amount to reflect $\eta_{i_1} \in [0, \pi/2]$ $i_1 = 1, \ldots, m_1 - 1$ in the neighborhood of zero and $\pi/2$, $z_{i_2} \geq 0$ $i_2 = 1, \ldots, m_2$ in the neighborhood of zero and $y_r \geq 0$ $r = 1, \ldots, s$ in the neighborhood of zero, as well.

[4] The number of student is not evenly distributed over the four year cycle and is significantly smaller in the last year. This is often explained by large number of students repeating a year or quitting before completing the degree (the rates are respectively 15.4 and 13.2% the first year, 16.5% and 12.9% the second year, 9.8% and 6.9% the third year and finally 23.6% and 8.9% the last year).

[5] Articles in the literature have often used student cohorts with information on academic results, socio-economic conditions to measure the outputs (e.g. Silva Portela & Thanassoulis (2001) and Muñiz (2002)). Others (e.g. Ouellette & Vierstraete, 2005) have used enrolment to measure outputs, a standard procedure. Here we have census data with information on specific high schools (number of students, success rates, number of teachers, and so on) and information on a standard results. So, we can say that our approach is midway between the standard approach and the value added models.

[6] This implicitly supposes that the price index is equal to one over the sample period.

[7] Note that for some authors, the quasi-fixed inputs are the socio-economic conditions of the students. For example, Cohn, Millman & Chew (1975), Hanushek (1986), Muñiz (2002) among others show that these conditions influence directly the students' performances. These socio-economic conditions seem to explain the efficiency results of the schools, more than some other factor entering directly in the production process. However, this is a drift from the optimal allocation of the resources.

[8] We have also conducted the same experiment for the Tunisian schools without residence. The results are qualitatively very close to the results for the first model. For this reason, we chose not to report the results here. An appendix containing the results for this model simulations and the corresponding interpretation are available from the authors upon request.

[9] The results of these simulations are available from the authors upon request.

[10] Simar et Wilson (2000a) have proposed to put $\tilde{x}^*_{1j} = 1$ in the transformation matrix $\tilde{X}$, but this leads to $\tilde{\theta}_j = 1$ $\forall j$.