On the estimation of educational technical efficiency from sample designs: a new methodology using robust nonparametric models

Juan Aparicio[1,*], Martín González[1], Daniel Santín[2] and Gabriela Sicilia[3]

[1] Center of Operations Research (CIO). Miguel Hernandez University of Elche (UMH), 03202 Elche (Alicante), Spain

[2] Department of Applied Economics, Public Economics and Political Economy, Complutense University of Madrid, Campus de Somosaguas, 28223 Madrid, Spain

[3] Department of Economics and Public Finance, Autonomous University of Madrid, Campus de Cantoblanco 28049 Madrid, Spain

Abstract

Average efficiency is popular in the empirical education literature for comparing the aggregate performance of regions or countries using the efficiency results of their disaggregated decision-making units (DMUs) microdata. The most common approach for calculating average efficiency is to use a set of inputs and outputs from a representative sample of DMUs, typically schools or high schools, in order to characterize the performance of the population in the analysed education system. Regardless of the sampling method, the use of sample weights is standard in statistics and econometrics for approximating population parameters. However, weight information has been disregarded in the literature on production frontier estimation using nonparametric methodologies in education. The aim of this chapter is to propose a preliminary methodological strategy to incorporate sample weight information into the estimation of production frontiers using robust nonparametric models. Our Monte Carlo results suggest that current sample designs are not intended for estimating either population production frontiers or average technical efficiency. Consequently, the use of sample weights does not significantly improve the efficiency estimation of a population with respect to an unweighted sample. In order to enhance future efficiency and productivity estimations of a population using samples, we should define an independent sampling design procedure for the set of DMUs based on the population's production frontier.

Keywords: technical efficiency, sample designs, education sector.

* Corresponding author: Tel.: +34 966658517; Fax: +34 966658715; Email: j.aparicio@umh.es

1. Introduction

Large-scale assessment surveys have played a growing role in educational research over the last three decades. Broadly defined, large-scale assessments are standardized surveys of cognitive skills in different subjects that provide comparable data about many different students, schools and, in short, education systems in one region, country or even several countries around the world. Because both home and school play an important role in how students learn, large-scale surveys also collect extensive information about such background factors at individual, teacher and school level. Some of the best-known international databases worldwide are the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS) or the Progress in International Reading Literacy Study (PIRLS). Additionally, in many developed countries, ministries of education gather similar data for analysing their educational systems.

Researchers can use this information for three important purposes. First, educational databases are useful for making cross-country comparisons of the achievements of different education systems, as well as for introducing the quality of human capital in economic growth regressions (Hanushek and Kimko, 2000; De la Fuente, 2011; Hanushek and Woessmann, 2012). Second, these databases are analysed for disentangling the causal effects of educational policies, law changes and different social factors on educational outcomes through the use of counterfactuals (Strietholt et al., 2014; Cordero et al. 2017). Finally, large-scale assessment surveys are used for measuring technical efficiency through production frontiers in order to benchmark the most successful educational policies (Afonso and St Aubyn, 2005, 2006; De Jorge and Santín, 2010, Agasisti and Zoido, 2018). This latter research line is the focus of this chapter, also addressed in recent related publications by Aparicio et al. (2017a, 2017b), Aparicio et al. (2018) and Aparicio and Santín (2018).

Furthermore, the use of representative samples of a population is an extremely widespread practice in statistics. Multiple methods have been developed for characterizing a population through a sample (see Hedayat and Sinha, 1991 and Särndal et al., 1992). There are some reasons for introducing weight designs in educational databases. First, sampling could oversample or undersample some major school types within the population. For example, the sample could include schools from major, albeit small, territories or regions, which, depending on the sampling method applied, could be either not well represented or overly significant when results are averaged to draw conclusions about the population. Second, school sizes vary across the school population. Therefore, average results at school level hide the fact that

the analysis covers all students at some schools and just a group of students at others. Finally, weighting is used to address non-response issues from some schools.

However, the sample weights that appear in many educational databases have been repeatedly ignored in econometrics. Recently, Lavy (2015) investigated whether instruction time has a positive impact on academic performance across countries using unweighted PISA 2006 data pooled at student level. Jerrim et al. (2017, p54) reanalysed Lavy's data, running the same regression analysis with the PISA final weights to capture the population size of each country. Their results show that the effect of an additional hour of instruction is almost 50% greater in developed countries and 40% smaller in Eastern European countries than Lavy's estimations. As a result, the parameters estimated from a sample might not be representing the population under study.

The same problem could affect production frontiers applied to educational databases when researchers assume that the average efficiency results for an unweighted sample can be straightforwardly identified as a good estimation for the population. So far, extensions have not been developed to incorporate the sample weights when estimating the production frontier and the efficiency scores for comparing the performance of different sets of schools.

Under the production frontier framework, there are basically two potential concerns affecting the estimation of technical efficiency. First, there is a representativeness problem, since only the weighted sample is representative of the population. Thus, sample weights are necessary to make valid estimates and inferences about any population parameter from the sample. Therefore, a straightforward adjustment could be to expand the sample to the population using the sample weights. Basically, this means including these weights to compute the aggregate (average) efficiency of the sector (region, country, etc.) and ensure that the entire population is represented. Second, the DMUs included in the analysis are only one of many possible sampling realizations of the population. Because not all DMUs have the same probability of inclusion in the sample, the omission of best-performers information might affect the shape of the estimated true production frontier. The potential misidentification of the true frontier also impairs the estimation of individual efficiency scores, since they are computed as the relative distance to the estimated frontier.

PISA, TIMSS and PIRLS are based on multi-stage probability proportional to size (PPS) sampling schemes. Basically, the sampling design is composed of two stages. First, schools are randomly selected from different strata taking into account the size of the schools. Second, students are randomly selected within each sampled school. As a result, each school and

each student have different probabilities of being included in the sample, i.e. different sample weights[1]. This makes weighting information crucial for getting unbiased estimates of population characteristics. As stated in the PISA 2015 Technical Report (OECD, 2017, p.116): *"Survey weights are required to analyse PISA data, to calculate appropriate estimates of sampling error and to make valid estimates and inferences of the population"… "While the students included in the final PISA sample for a given country were chosen randomly, the selection probabilities of the students vary. Survey weights must be incorporated into the analysis to ensure that each sampled student appropriately represents the correct number of students in the full PISA population."* For this reason, it could be misleading to extrapolate results from sample to population regardless of weighting. This problem can also arise in other sectors, like health, banking, agriculture, etc., where the use of samples is commonplace too.

How can we deal with weights in production frontiers? Nonparametric methods, and especially data envelopment analysis, have been applied for measuring efficiency much more often than parametric methods in the education literature. Their extensive application is a consequence of their flexibility, as there is no theoretical education production function (Levin, 1974) and few assumptions are needed to envelop the best performers. Nonparametric methods do not explicitly estimate the parameters of a production technology. Instead, they determine an efficiency index reflecting how much use each unit makes of its available resources based on a mathematical model implicitly describing the estimated technology.

Taking insights from the conditional quantile-based approach proposed by Aragon et al. (2005), this paper provides a preliminary methodological strategy to incorporate the information of sample weights into the estimation of the production frontier using robust nonparametric models. The final aim is to enhance the estimation of the technical efficiency of a population of DMUs using a representative sample and its weights as is common practice in education. The reason why we select Aragon et al. (2005), among other possibilities, is that it allows extending the standard frontier analysis to contexts with sample weights[2] in a simple way.

The remainder of the paper is organized as follows. In Section 2, we discuss the main methodological issues related to the estimation of nonparametric production from sample designs. In Section 3 we propose a method to add the sample weights to the estimation of

[1] For a detailed explanation of this sampling design, see Chapter 4 of the PISA 2015 Technical Report.
[2] In the DEA literature, we can find some references that include weights like, for example, Allen et al. (1997) and Färe and Zelenyuk (2003). However, they do not consider sample weights from sample designs.

robust nonparametric frontier models. Section 4 is devoted to check the performance of this method through a Monte Carlo experiment. Finally, Section 5 outlines the main conclusions.

2. Methodological issues

In this section, we briefly review the main nonparametric frontier estimators, their robust estimation through partial frontiers, and some key notions about finite population sampling in statistics before extending Aragon et al.'s approach (2005) to the context where information on the sampling design is available. See also Daouia and Simar (2007) and Daraio and Simar (2007).

2.1. Nonparametric frontier models

In frontier analysis, most of the nonparametric approaches —free disposal hull (FDH) and data envelopment analysis (DEA)— are based upon enveloping the set of observations from above to let the data speak for themselves, as well as requiring certain properties (like monotonicity). According to economic theory (Koopmans, 1951; Debreu, 1951; Shephard, 1953), the production set, where the activity is described by a set of m inputs $x \in R_+^m$ used to produce a set of p outputs $y \in R_+^p$, is defined as the set of all physically producible activities given a certain knowledge (x,y): $\Psi = \{(x,y) \in R_+^{m+p} : x \text{ can produce } y\}$ (see also Pastor et al., 2012).

In this paper, we assume that Ψ is a subset of R_+^{m+p} that satisfies the following postulates (see Färe et al., 1985).

(P1) $\Psi \neq \varnothing$;

(P2) $\Psi(x) := \{(u,y) \in \Psi : u \leq x\}$ is bounded $\forall x \in R_+^m$;

(P3) $(x,y) \in \Psi, (x,-y) \leq (x',-y') \Rightarrow (x',y') \in \Psi$, i.e., inputs and outputs are freely disposable;

(P4) Ψ is a closed set;

(P5) Ψ is a convex set.

A certain activity (observation) is considered to be technically inefficient if it is possible to either expand its output bundle y without requiring any increase in its inputs x or contract its input bundle without requiring a reduction in its outputs. The capacity for expanding the output bundle reflects output-oriented inefficiency. Likewise, potential input savings indicate input-oriented inefficiency. Exactly which of these two orientations is selected depends on the analysed empirical framework. On the one hand, it is assumed, in the case of input-oriented contexts, that the output bundle (like the number of patients to be treated at a hospital) is fixed or given by the demand side. Hence, it is reasonable to save on the use of inputs to contain costs. In this case, determining input-oriented technical efficiency measurements by scaling down x (the frontier of Ψ) as far as possible is the most rational first step. On the other hand, when the input bundle is predetermined (like land at a farm), output-oriented technical efficiency measurements would appear to be a better option.

For simplicity's sake, we assume in this paper that firms, schools if we refer to the education sector, cannot change their inputs in the short run or that they are given. Consequently, output-orientation is the best choice, and we will evaluate their performance based on the assessment of the production of outputs from a certain level of inputs. In this context, it is common practice to work with the notion of requirement set. The requirement set, denoted as $Y(x)$, is the set of all outputs that a firm can produce using $x \in R_+^m$ as inputs. Mathematically speaking, $Y(x) = \left\{ y \in R_+^p : (x,y) \in \Psi \right\}$.

Assumptions on the data generating process (DGP) encompass the statistical model, which defines how the observations in Ψ are generated. There are many alternatives. However, since nonparametric methods for estimating frontiers have no need of parametric assumptions about the DGP, we will simply assume that the production process, which generates the set of observations $\Theta_n = \left\{ (x_i, y_i) : i = 1,...,n \right\}$, is defined by the joint distribution of the random vector $(X,Y) \in R_+^m \times R_+^p$, where X represents the random inputs and Y represents the random outputs. Where Ψ is equal to the support of the distribution of (X,Y) and $p = 1$, another way to define the production frontier is through the notion of production function. The production function, denoted as φ, is characterized for a given level of inputs $x \in R_+^m$ by the upper boundary of the support of the conditional distribution of the univariate Y given $X \leq x$

, i.e., $\varphi(x) = \sup\{y \in R_+ : F(y \mid x) < 1\}$, where $F(y \mid x)$ is the conditional distribution function of Y given $X \leq x$. The inequality $X \leq x$ should be interpreted componentwise. We owe this formulation of the production function to Cazals et al. (2002), and it is useful for expressing the customary notion of production function by distribution functions.

Regarding the practical determination of the technology from a data sample, economists before Farrell (1957) used to parametrically specify the corresponding production functions (e.g. a Cobb-Douglas function) and apply ordinary least squares (OLS) regression analysis to estimate an 'average' production function, assuming that disturbance terms had zero mean. However, the notion of production function moves away from the concept of average. In this respect, Farrell (1957) was the first author to show how to estimate an isoquant enveloping all the observations and, therefore, was the first to econometrically implement the idea of production frontier.

The line of research initiated by Farrell in 1957 was later taken up by Charnes et al. (1978) and Banker et al. (1984), resulting in the development of the data envelopment analysis (DEA) approach, where the determination of the frontier is only constrained by its axiomatic foundation and the property of convexity plays a major role. Additionally, Deprins et al. (1984) introduced a more general version of the DEA estimator, depending exclusively upon the free disposability assumption of inputs and outputs and neglecting convexity. Indeed, the two main nonparametric frontier techniques in the literature nowadays are: DEA and FDH. In the case of DEA, the frontier estimator is, as already mentioned, constructed as the smallest polyhedral set that contains the observations and satisfies free disposability, whereas FDH makes fewer assumptions than DEA. Graphically, the convex hull of the FDH estimate is the same as the DEA estimate of the production technology.

Aigner and Chu (1968) reported a more natural follow-on from previous research by econometricians than DEA and FDH. They showed how to apply a technique based on mathematical programming to yield an envelope 'parametric' Cobb-Douglas production function by controlling the sign of the disturbance terms and, consequently, following the standard definition of production function. A more general parametric approach is the stochastic frontier analysis (SFA) by Aigner, Lovell and Schmidt (1977) and Meeusen and Van den Broeck (1977).

Generally speaking, two different approaches have been introduced in the literature: deterministic frontier models, like DEA and FDH, which assume with probability one that all the observations in Θ_n belong to Ψ, and stochastic frontier models, like SFA, where, due to random noise, some observations may be outside of Ψ.

2.2. Nonparametric robust estimators: partial frontiers

Nonparametric deterministic frontier models, like DEA and FDH, are very attractive because they depend on very few assumptions. However, by definition, they are very sensitive to extreme values. To solve this problem, Cazals et al. (2002) and Aragon et al. (2005) proposed robust nonparametric frontier techniques. In this section, we briefly review the main features of these two approaches.

Cazals et al. (2002) introduced the notion of expected maximal output frontier of order $m \in \mathrm{N}^*$, where N^* denotes the set of all integers $m \geq 1$. It is defined as the expected maximum achievable level of output across m units drawn from the population using less than a given level of inputs. Formally, for a fixed integer $m \in \mathrm{N}^*$ and a given level of inputs $x \in R_+^m$, the order-m frontier is defined as

$$\varphi_m(x) = E\left[\max\left(Y^1,...,Y^m\right)\right] = \int_0^\infty \left(1 - \left[F(y \mid x)\right]^m\right) dy, \qquad (1)$$

where $\left(Y^1,...,Y^m\right)$ are m independent identically distributed random variables generated by the distribution of Y, given $X \leq x$. Its nonparametric estimator is defined by $\hat{\varphi}_m(x) = \int_0^\infty \left(1 - \left[\hat{F}(y \mid x)\right]^m\right) dy$, which is based upon the estimation of the distribution function. In particular, $\hat{F}(y \mid x) = \hat{F}(x,y) / \hat{F}(x)$ is the empirical version of the conditional distribution function of Y given $X \leq x$, with $\hat{F}(x,y) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x, Y_i \leq y)$ and

$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$. Cazals et al. (2002) were also able to rewrite the FDH estimator of the production function in terms of the conditional distribution function as $\hat{\varphi}_{FDH}(x) = \sup\left\{y \in R_+ : \hat{F}(y \mid x) < 1\right\} = \max_{i:X_i \leq x}\{Y_i\}$.

By definition, the order-m frontier does not envelop all the observations in the sample. Consequently, it is more robust to extreme values and outliers than the standard FDH estimator. Additionally, using an appropriate selection of m as a function of the sample size, $\hat{\varphi}_m(x)$ estimates the production function $\varphi(x)$ while, at the same time, retaining the asymptotic properties of the FDH estimator.

Later, Aragon et al. (2005) proposed a nonparametric estimator of the production function that is, as they demonstrated, more robust to extreme values than the standard DEA and FDH estimators and the nonparametric order-m frontier by Cazals et al. (2002). This model is based upon quantiles of the conditional distribution of Y given $X \leq x$. These conditional quantiles define a natural notion of a partial production frontier in place of the order-m frontier. Moreover, Aragon et al. (2005) proved that their estimators satisfy most of the good properties of the order-m estimator.

In particular, the quantile production function of order α, $\alpha \in [0,1]$, given a certain level of inputs $x \in R_+^m$, can be defined as

$$q_\alpha(x) = F^{-1}(\alpha \mid x) = \inf\left\{y \in R_+ : F(y \mid x) \geq \alpha\right\}. \tag{2}$$

This conditional quantile is the production threshold exceeded by $100(1-\alpha)\%$ of units that use less than the level x of inputs. Notice that, by the traditional production function definition, $\varphi(x)$ coincides with the order-one quantile production function, i.e. $\varphi(x) = q_1(x)$.

The natural way of estimating $q_\alpha(x)$ is to substitute the conditional distribution function by its empirical estimation $\hat{F}(\cdot \mid x)$:

$$\hat{q}_\alpha(x) = \hat{F}^{-1}(\alpha \mid x) = \inf\left\{y \in R_+ : \hat{F}(y \mid x) \geq \alpha\right\}. \tag{3}$$

This estimator may be computed explicitly as follows (see Aragon et al., 2005). Let $s_x = \left\{i_1,...,i_{n_x}\right\}$ be the subset of observations in the data sample such that $X_i \leq x$, where $n_x = \sum_{i=1}^{n} 1(X_i \leq x)$, i.e. the number of elements in s_x. Hence, $Y_{i_1},...,Y_{i_{n_x}}$ corresponds to the outputs observed in s_x, while $Y_{(i_1)},...,Y_{(i_{n_x})}$ are their ordered values. Additionally, it is assumed that the labels $i_1,...,i_{n_x}$ contain no information as to the ordering of the values of $Y_{i_1},...,Y_{i_{n_x}}$. For example, it is not necessarily true that $Y_{i_j} < Y_{i_k}$ for $j < k$. However, $Y_{(i_j)} \leq Y_{(i_k)}$ for all $j \leq k$. We also assume that $n_x \neq 0$.

The estimation of the conditional distribution function is

$$\hat{F}(y \mid x) = \frac{\sum_{i:X_i \leq x} 1(Y_i \leq y)}{n_x} = \frac{\sum_{j=1}^{n_x} 1\left(Y_{(i_j)} \leq y\right)}{n_x}. \tag{4}$$

Hence,

$$\hat{F}(y \mid x) = \begin{cases} 0, & \text{if } y < Y_{(i_1)} \\ k/n_x, & \text{if } Y_{(i_k)} \leq \text{y } < Y_{(i_{k+1})}, 1 \leq k \leq n_x - 1 \\ 1, & \text{if } y \geq Y_{(i_{n_x})} \end{cases} \tag{5}$$

Consequently, for any $\alpha > 0$, we have that

$$\hat{q}_\alpha(x) = \begin{cases} Y_{(i_{\alpha n_x})}, & \text{if } \alpha n_x \in \mathrm{N}^* \\ Y_{([\alpha n_x]+1)}, & \text{otherwise} \end{cases}, \tag{6}$$

where $[\alpha n_x]$ is the largest integer less than or equal to αn_x. Consequently, the conditional empirical quantile $\hat{q}_\alpha(x)$ is computed as the simple empirical quantile of $Y_{i_1},...,Y_{i_{n_x}}$. Additionally, note that $\varphi(x) = q_1(x) = Y_{(i_{n_x})} = \max_{i:X_i \le x}\{Y_i\}$, i.e. it is equal to the FDH estimation.

In this research, we extend Aragon et al.'s notion of α-quantile production function (Aragon et al., 2005) in order to deal with situations where the data sample is the result of applying a particular sampling design on a finite population.

2.3. Sampling designs on a finite population

Let us now assume that the first stage of the production process generates a finite population of production units $\Theta_N = \{(X_i, Y_i) : i = 1,...,N\}$. For simplicity's sake, let the i-th element be represented by its label i. Thus, we denote the finite population as $U = \{1,...,N\}$. Additionally, we assume that we are an observer outside the production process. Hence, the values X_i and Y_i, $i = 1,...,N$, are unknown to us. Let us also suppose that, in a second stage, we select a subset of the population, called a sample and denoted as $s \subseteq U$, with the aim of estimating some parameters associated with the population. Specifically, we are able to observe and collect the values of X_i and Y_i for all $i \in s$. In particular, as is very common in social science, we consider samples that are realized by a probabilistic (randomized) selection scheme.

Given a sample selection scheme, it is possible, although not always simple, to establish the probability of selecting a specified sample s. We shall use the notation $p(s)$ for this probability. In this way, we assume that there is a function $p(\cdot)$ such that $p(s)$ gives the probability of selecting s under the scheme in use. The function $p(\cdot)$ is usually called the sampling design in finite population sampling theory. This notion plays a central role because it determines the essential statistical properties (sampling distribution, expected value and variance) of random quantities calculated from the data sample (estimators) in order to estimate certain population parameters or functions of parameters.

For a given sampling design $p(\cdot)$, we can regard any sample s as the outcome of a set-valued random variable S, whose probability distribution is specified by the function $p(\cdot)$. Let Γ be the set of all samples. Thus the cardinal of Γ is 2^N if we consider the empty set as well as U itself. Then we have that $\Pr(S = s) = p(s)$ for any $s \in S$. Because $p(\cdot)$ is a probability distribution on the set Γ, we have (i) $p(s) \geq 0$, $s \in S$, and (ii) $\sum_{s \in \Gamma} p(s) = 1$. Note that the probability of some (usually many) of the 2^N samples contained in Γ is equal to zero. The subset of Γ composed of any samples for which $p(s)$ is strictly positive constitutes the set of possible samples. They are the only ones that can be drawn given the specified design.

The sample size, denoted as n_s, is the number of elements in s. Note that n_s depends on the sample and is not, therefore, necessarily the same for all possible samples. If, in fact, all possible samples have the same size, then the sample size is denoted, as usual, as n. For example, Bernoulli sampling can generate different sample sizes, while simple random sampling without replacement always yields the same sample size (Särndal et al., 1992). For simplicity's sake, we will assume hereafter that all possible samples have the same size n.

An interesting feature of a finite population of N units is that each unit can be given different probabilities of inclusion in the sample. The sampling statistician often takes advantage of the identifiability of the population unit by deliberately attaching different inclusion probabilities to the various elements. This is one way to get more accurate estimates, for example, by using strata, clusters or some known auxiliary variable related to the size of the population units.

Given a sampling design $p(\cdot)$, the probability that unit k was included in a sample, denoted π_k, is obtained from the given design $p(\cdot)$ as $\pi_k = \Pr(k \in S) = \sum_{s:k \in s} p(s)$.

One very usual parameter to be estimated in these contexts is the total of a population, defined for a response variable Z as $t_z = \sum_{i \in U} Z_i$. An unbiased estimator of t_z, under any sampling design, is the so-called π estimator, which resorts to the use of the inclusion probabilities of the units belonging to the data sample. In particular, it is expressed as follows:

$$\hat{t}_{\pi z} = \sum_{i \in s} \frac{Z_i}{\pi_i}. \tag{7}$$

The π estimator expands the values collected in the sample by increasing the importance of the observed population units. Because the sample contains fewer elements than the original population, an expansion is required to reach the level for the total population. The i-th unit, when present in the sample, will represent $1/\pi_i$ population units. As it is unbiased, the π estimator is the cornerstone of the main estimators in finite population sampling theory. Formulations of the variance and estimations of the variance of the π estimator can be found in many textbooks (see, for example, Särndal et al. 1992 and Hedayat and Sinha, 1991). Horvitz and Thompson (1952) were the first authors to use this expansion principle to estimate the total of a population, on which ground the π estimator is also called the Horvitz-Thompson (HT) estimator in the literature.

3. An adaptation of the order-α quantile-type frontier for dealing with sampling designs

In this framework, we adapt the estimation of the order-α quantile production function, $\alpha \in [0,1]$, to work with a data sample derived from a sampling design $p(\cdot)$. The results reported in this section are completely new.

The conditional distribution function of the survey variable Y given $X \le x$ in a finite population of size N is defined as follows:

$$F_U(y \mid x) = \frac{\sum_{i=1}^{N} 1(X_i \le x, Y_i \le y)}{N_x}, \tag{8}$$

where $N_x = \sum_{i=1}^{N} 1(X_i \le x)$ represents the number of units in the population such that $X \le x$.

Notice that, from the point of view of finite population sampling, $F_U(y \mid x)$ is a population parameter since it is defined through the unknown values of survey variables X and Y for all the population units U. At the same time, $F_U(y \mid x)$ could be considered the empirical estimation of $F(y \mid x)$ (see Section 2) for the original production process from the N generated observations, which are, as already pointed out, unknown to us.

Note also that the estimation process linked to the quantile production function described above could be applied to U instead of the observed s in order to determine an estimation for $q_\alpha(x)$. In this case, $\hat{F}(y \mid x)$ should be substituted by $F_U(y \mid x)$.

In our framework, however, we are an observer outside the production process. Consequently, the values X_i and Y_i, $i = 1,...,N$, are unknown to us. It implies that we cannot apply the estimation process by Aragon et al. (2005) for the quantile production function directly on U. Instead, we select a subset of the population $s \subseteq U$ and try to accurately estimate some population parameters of interest, like $F_U(y \mid x)$.

To do this, note first that $t_{I(x,y)} = \sum_{i=1}^{N} 1(X_i \le x, Y_i \le y)$ is really the total of the population for the binary membership indicator variable $I(x,y)$ defined as:

$$I_i(x,y)=\begin{cases}1, & \text{if } X_i \le x \text{ and } Y_i \le y\\ 0, & \text{otherwise}\end{cases}. \tag{9}$$

Additionally, $N_x=\sum_{i=1}^{N}1(X_i \le x)$ is the total of the population for the binary membership indicator variable $I(x)$ defined as:

$$I_i(x)=\begin{cases}1, & \text{if } X_i \le x\\ 0, & \text{otherwise}\end{cases}. \tag{10}$$

Then, $F_U(y \mid x)=t_{I(x,y)}/N_x$ is the ratio of two population totals. The HT estimator can estimate each total without bias. Hence, we propose the following estimator for $F_U(y \mid x)$:

$$\hat{F}_U(y \mid x)=\frac{\sum_{i\in s}1(X_i \le x, Y_i \le y)/\pi_i}{\sum_{i\in s}1(X_i \le x)/\pi_i}=\frac{\sum_{j=1}^{n_x}1\left(Y_{(i_j)} \le y\right)/\pi_{(i_j)}}{\sum_{j=1}^{n_x}1/\pi_{i_j}}. \tag{11}$$

Using (first-order) Taylor linearization of the ratio $\hat{F}_U(y \mid x)$, we get

$$\hat{F}_U(y \mid x) \approx F_U(y \mid x)+\frac{1}{N_x}\sum_{i\in s}\frac{I_i(x,y)-F_U(y \mid x)I_i(x)}{\pi_i}. \tag{12}$$

This implies that $E\left[\hat{F}_U(y \mid x)\right] \approx E\left[F_U(y \mid x)+\frac{1}{N_x}\sum_{i\in s}\frac{I_i(x,y)-F_U(y \mid x)I_i(x)}{\pi_i}\right]=$

$F_U(y \mid x)+\frac{1}{N_x}E\left[\sum_{i\in s}\frac{I_i(x,y)-F_U(y \mid x)I_i(x)}{\pi_i}\right]$. And $E\left[\sum_{i\in s}\frac{I_i(x,y)-F_U(y \mid x)I_i(x)}{\pi_i}\right]=$

$$E\left[\sum_{i\in s}\frac{I_i(x,y)}{\pi_i}-F_U(y\mid x)\sum_{i\in s}\frac{I_i(x)}{\pi_i}\right]=E\left[\sum_{i\in s}\frac{I_i(x,y)}{\pi_i}\right]-F_U(y\mid x)E\left[\sum_{i\in s}\frac{I_i(x)}{\pi_i}\right]=$$

$t_{I(x,y)}-F_U(y\mid x)N_x=t_{I(x,y)}-\frac{t_{I(x,y)}}{N_x}N_x=0$, which means that $E\left[\hat{F}_U(y\mid x)\right]\approx F_U(y\mid x)$.

In other words, the estimator $\hat{F}_U(y\mid x)$ is approximately unbiased for $F_U(y\mid x)$, which is, at the same time, the estimator that Aragon et al. (2005) would use for approximating $F(y\mid x)$.

Moreover, $\hat{F}_U(y\mid x)$ may be expressed as

$$\hat{F}_U(y\mid x)=\begin{cases}0, & \text{if } y<Y_{(i_1)}\\ \dfrac{\sum_{j=1}^{k}1/\pi_{(i_j)}}{\sum_{j=1}^{n_x}1/\pi_{i_j}}, & \text{if } Y_{(i_k)}\le y<Y_{(i_{k+1})}, 1\le k\le n_x-1.\\ 1, & \text{if } y\ge Y_{(i_{n_x})}\end{cases} \tag{13}$$

Let us now introduce some new notation. Let $W_k=\sum_{j=1}^{k}1/\pi_{(i_j)}$. Additionally, let $q_{\alpha,U}(x)$ be the empirical quantile of Y calculated from Y_i, $i\in U$ such that $X_i\le x$.

Then, following Aragon et al.'s (2005) approach, an estimator of $q_{\alpha,U}(x)$ would be $\hat{q}_{\alpha,U}(x)=Y_{(i_k)}$, where k, $1\le k\le n_x$, is the smallest index such that $W_k\ge\alpha W_{n_x}$, for $\alpha>0$.

In the extreme case of $\alpha=1$, i.e. when the quantile to be estimated is equal to the maximum, note that $\hat{q}_{1,U}(x)=Y_{(i_{n_x})}=\max_{i:X_i\le x}\{Y_i\}$. This means that the estimation of the traditional

production function constructed from the N population units is equal to the standard FDH estimation calculated from n observations, regardless of the sampling design.

The following proposition establishes that any sampling design that generates identical inclusion probabilities for all the elements of the population produces the same quantile production function estimator as Aragon et al.'s (2005) approach applied directly to the n observations, i.e. without using the information contained in π_i, $i \in U$.

Proposition 1. Let $p(\cdot)$ such that $\pi_i = \pi \ \ \forall i \in U$, then $\hat{q}_{\alpha,U}(x) = \hat{q}_{\alpha}(x)$ for any $\alpha > 0$.

Proof. From (11), $\hat{F}_U(y \mid x) = \dfrac{\sum_{i \in s} 1(X_i \le x, Y_i \le y)/\pi_i}{\sum_{i \in s} 1(X_i \le x)/\pi_i} \overset{\text{by hypothesis}}{=} \dfrac{\sum_{i \in s} 1(X_i \le x, Y_i \le y)/\pi}{\sum_{i \in s} 1(X_i \le x)/\pi} =$

$\dfrac{\sum_{i \in s} 1(X_i \le x, Y_i \le y)}{\sum_{i \in s} 1(X_i \le x)} = \dfrac{\sum_{i:X_i \le x} 1(Y_i \le y)}{n_x}$, since $n_x = \sum_{i=1}^{n} 1(X_i \le x)$. By expression (4), we have that $\hat{F}_U(y \mid x)$ is equal to $\hat{F}(y \mid x)$ in Aragon et al. (2005). Consequently, $\hat{q}_{\alpha,U}(x) = \hat{q}_{\alpha}(x)$ for any $\alpha > 0$. ■

Several well-known sampling designs satisfy the hypothesis in Proposition 1, including Bernoulli sampling, simple random sampling without replacement and systematic sampling. The following result establishes that, under these designs, our approach generates the same estimations as the approach by Aragon et al. (2005).

Corollary 1. Applying Bernoulli sampling (BE), simple random sampling without replacement (SRS) and systematic sampling (SS), $\hat{q}_{\alpha,U}(x) = \hat{q}_{\alpha}(x)$ for any $\alpha > 0$.

As a consequence of Corollary 1, BE, SRS and SS generate the same estimation for the quantile production function of order $\alpha > 0$ as by directly applying Aragon et al.'s approach (2005) without taking into account that the sample has been drawn from a finite population U. Hence, if a researcher's database is built from the above sampling design types, then it suffices, as suggested by Aragon et al. (2005), to determine the empirical quantile of observations in the sample such that $X \leq x$. The problem arises when the data used in the empirical study come from a sampling design with non-equal inclusion probabilities. For example, the sampling statisticians in the famous PISA report resort to random schemes based on inclusion probabilities proportional to a positive and known auxiliary variable, such as the number of students in each school. This means that the inclusion probabilities vary across the population units and, therefore, the hypothesis stated in Proposition 1 does not hold. The effect of this deviation on the estimation of the quantiles is something that warrants detailed investigation due to the importance of reports like PISA. Indeed, the PISA technical report states that "*While the students included in the final PISA sample for a given country were chosen randomly, the selection probabilities of the students vary. Survey weights must therefore be incorporated into the analysis to ensure that each sampled student represents the appropriate number of students in the full PISA population*" (PISA 2012 Technical Report, p. 132).

The above process of estimation is able to generate a point estimation for the population quantile $q_{\alpha,U}(x)$. However, a confidence interval of this parameter sometimes has to be used to make other inference types. Next, we propose an approximate confidence interval for the population quantile $q_{\alpha,U}(x)$.

Our approach is inspired by Woodruff (1952). This method was used by Woodruff (1952) for confidence intervals of medians, although it can be generalized to other quantiles. In our context, the approach requires computing a confidence interval for $F_U(y \mid x)$. Assuming that these values are F_l and F_u, the confidence interval for $q_{\alpha,U}(x)$, namely (q_l, q_u), is implicitly defined by the equations: $\hat{F}_U(q_l \mid x) = F_l$ and $\hat{F}_U(q_u \mid x) = F_u$.

In order to determine a confidence interval for the population parameter $F_U(y \mid x)$, we first need to propose an estimation of the variance of the estimator $\hat{F}_U(y \mid x)$. Following Särndal et al. (1992), an approximate variance of the π estimator of the ratio of two totals is

$$V\left(\hat{F}_U(y \mid x)\right) \approx \frac{1}{N_x^2} \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \frac{I_i(x,y) - F_U(y \mid x) I_i(x)}{\pi_i} \frac{I_j(x,y) - F_U(y \mid x) I_j(x)}{\pi_j}, \tag{14}$$

which can be estimated through

$$\hat{V}\left(\hat{F}_U(y \mid x)\right) \approx \frac{1}{\left(\sum_{j=1}^{n_x} 1/\pi_{i_j}\right)^2} \sum_{i \in s} \sum_{j \in s} \tilde{\Delta}_{ij} \frac{I_i(x,y) - \hat{F}_U(y \mid x) I_i(x)}{\pi_i} \frac{I_j(x,y) - \hat{F}_U(y \mid x) I_j(x)}{\pi_j}, \tag{15}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, $\tilde{\Delta}_{ij} = \Delta_{ij} / \pi_{ij}$ and $\pi_{ij} = \Pr(i \in S, j \in S)$.

Then, a confidence interval for $F_U(y \mid x)$ at the approximate level $1 - \beta$ can be computed as

$$\hat{F}_U(y \mid x) \pm z_{1-\beta/2} \left[\hat{V}\left(\hat{F}_U(y \mid x)\right)\right]^{1/2}, \tag{16}$$

where $z_{1-\beta/2}$ is the constant exceeded with probability $\beta/2$ by the $N(0,1)$ random variable.

Let us define the following two elements, which we will use to define the confidence interval:

$$F_l = \hat{F}_U(y \mid x) - z_{1-\beta/2} \left[\hat{V}\left(\hat{F}_U(y \mid x)\right)\right]^{1/2} \tag{17}$$

and

$$F_u = \hat{F}_U(y \mid x) + z_{1-\beta/2}\left[\hat{V}\left(\hat{F}_U(y \mid x)\right)\right]^{1/2}. \quad (18)$$

Then, a confidence interval for $q_{\alpha,U}(x)$ at the approximate level $1-\beta$ is (q_l, q_u) with $q_l = Y_{(i_l)}$, where l, $1 \le l \le n_x$, is the largest index such that $W_l \le F_l W_{n_x}$, and $q_u = Y_{(i_u)}$, where u, $1 \le u \le n_x$, is the largest index such that $W_u \le F_u W_{n_x}$. The problem in this case is that we need to know not only π_i but also π_{ij} to determine the approximate confidence interval. Unfortunately, the database owner (e.g. the OECD for PISA) does not always provide this information.

4. Monte Carlo experiment

In order to test the performance of the proposed method, we perform a Monte Carlo experiment applied to three different scenarios assuming different sample designs. As discussed in Section 3, it is common in the educational context to observe complex sample designs where the probabilities of inclusion in the sample are not equal across the population units. Particularly, most large-scale international educational assessments (e.g. PISA, TIMSS, PIRLS, etc.) are based on a probability proportional to size (PPS) design, where the inclusion probabilities are proportional to a positive and known auxiliary variable (e.g. the number of students in each school).

4.1. Experimental design

To carry out the experiment, we replicate Aragon et al.'s Example 1 (2005). Thus, the data generation process is rooted in a Cobb-Douglas log-linear single-input single-output model given by $Y = X^{0.5}\exp^{-U}$, where the input X is uniformly distributed between (0,1) and the efficiency component U is exponentially distributed with mean $\left(\frac{1}{3}\right)$. Finally, the true frontier is defined by $\varphi(x) = x^{0.5}$.

All scenarios in this Monte Carlo experiment are based on a PPS design with a population size (N) equal to 1,000. First, we compute a scenario assuming that the sample is drawn using a PPS design and the auxiliary variable T_j is not correlated with the efficiency level (referred to hereinafter as the *non-informative design* scenario). The second scenario is generated drawing the sample from a PPS design and assuming that the auxiliary variable T_j is highly correlated with the level of efficiency by $T_j = (\exp^{-U})^4$ (referred to hereinafter as the *informative design* scenario). Finally, the third scenario is simulated using a two-stage sampling design. In this scenario, half of the sample is drawn using an *informative design* and the second half of the sample is drawn using a simple random sample (SRS) design (referred to hereinafter as the *two-stage design* scenario). In this scenario, we use the first half of the sample only to estimate the frontier and the second half of the sample only to estimate the average efficiency.

We replicate each scenario for different sample sizes (50, 100, 300 and 500), i.e. we simulate four different sampling fractions $f = \frac{n}{N}$. In large-scale international assessment, we usually observe sample sizes of around 50 schools. This is usually no more than 10% of the population at country level. However, there are some exceptions where f can be very large (even equal to 1), for example, when some countries expand the sample at regional level. In this vein, we aim to simulate different sample sizes to dimension the problem according to the sampling fraction.

For each dataset, we estimate the population quantile frontier $\hat{q}_{\alpha,U}(x)$ and the individual efficiency score $\hat{\theta}_j$ for each observation included in the sample $j = 1, 2, ..., n$ by running the order-α quantile-type frontier model proposed by Aragon et al. (2005) (referred to hereinafter as the order-α model) and our proposed adaptation of this model to include the sample weights (referred to hereinafter as the AGSS model) for $\alpha = 0.8$, $\alpha = 0.9$ and $\alpha = 1$. Finally, for each dataset, we estimate the average population efficiency μ from the sample, both omitting (which is the standard practice) and accounting for inclusion probabilities. Thus, we define the following estimators:

$$\hat{\mu}^{order-\alpha} = \frac{1}{n}\sum_{j=1}^{n} \hat{\theta}_j^{order-\alpha} \qquad (19)$$

$$\hat{\mu}^{order-\alpha,\pi} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\pi_j}\hat{\theta}_j^{order-\alpha} \qquad (20)$$

$$\hat{\mu}^{AGSS} = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_j^{AGSS} \qquad (21)$$

$$\hat{\mu}^{AGSS,\pi} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\pi_j}\hat{\theta}_j^{AGSS} \qquad (22)$$

Note that we use only the half of the sample drawn from a SRS design to estimate the population average efficiency in the *two-stage* sampling design. This means that the probabilities of inclusion are identical for all observations, and, consequently, it is not necessary to take this information into account. Thus, for this sampling design, we only provide the estimators $\hat{\mu}^{order-\alpha}$ and $\hat{\mu}^{AGSS}$.

In summary, we simulate 36 scenarios (three sample designs, four sampling fractions and three levels of α). In order to make the results more reliable, we undertook a Monte Carlo experiment, where B, the number of replicates, is 100. Therefore, all measures were computed in each replication and then averaged to get the results reported in the next section.

4.2. Results

4.2.1. Results on the population quantile-type frontier estimation

In order to dimension the effect of taking into account the sample weights to estimate the order-α quantile-type frontier in finite population samples, we compare the results from both the *order-*α and *AGSS* models. To do this, we compute the mean square error (MSE) for each model:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{q}_{\alpha,U}(x)_j - q_{\alpha,U}(x)_j\right)^2, \qquad (23)$$

where $q_{\alpha,U}(x)_j$ is the population quantile-type frontier of order α evaluated at unit *j* and $\hat{q}_{\alpha,U}(x)_j$ is the estimation of this order-α frontier at the same point. Note that, for $\alpha = 1$, the quantile production $q_1(x)$ coincides with the production function $\varphi(x)$. Results from this analysis are shown in Table 1. To illustrate the above ideas, we report the results for the population production frontier estimation from one particular simulation. Figures 1, 2 and 3

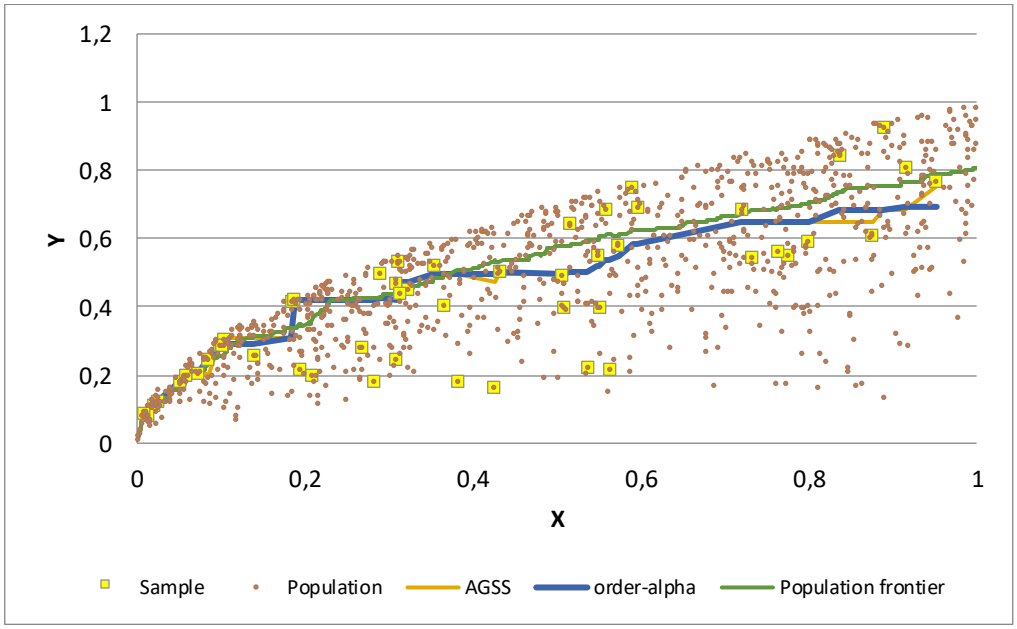
show these results for the *non-informative* design, *informative* design and *two-stage* design ($\alpha = 0.9$ and n=50, n=300, respectively). Note that these results are plotted merely for illustrative purposes, since they represent only one simulation. To properly compare model performance, we also compare the MSE of the Monte Carlo simulation.

The first remarkable result from the Monte Carlo experiment is that, in the extreme case of $\alpha = 1$, i.e. when the quantile to be estimated is equal to the maximum (last two columns of Table 1), we obtain the same estimation of the population production function with both models, regardless of the sample design. Consequently, from now on, we will focus on the results for $\alpha < 1$.

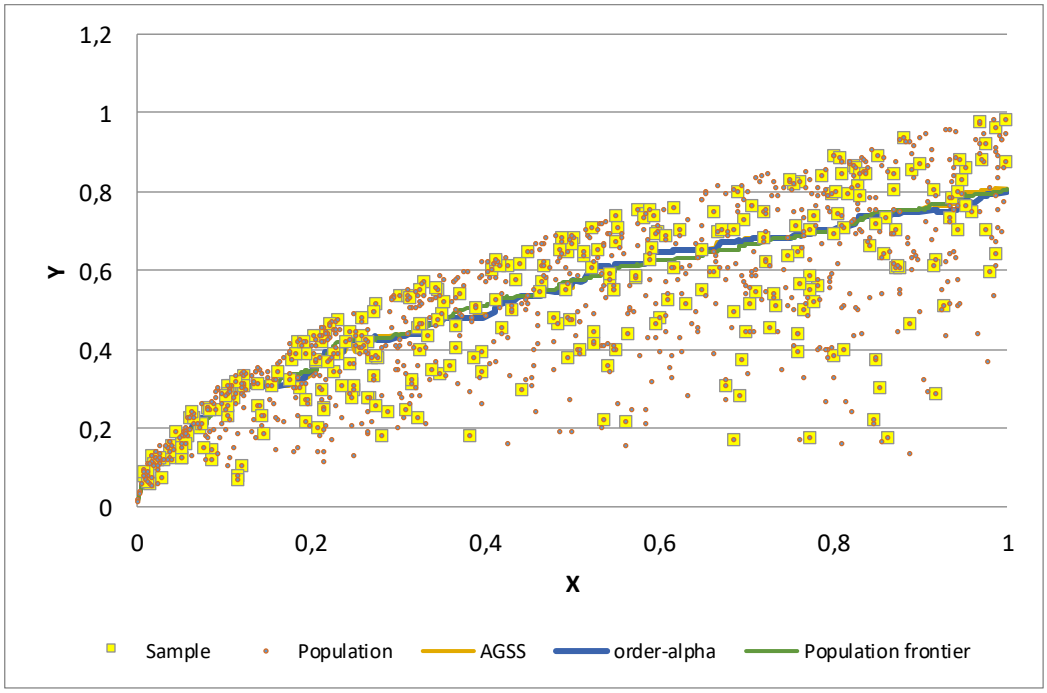
In the *non-informative* scenario, the results demonstrate that the order-α model proposed by Aragon et al (2005) performs reasonably well. In other words, the omission of different probabilities of inclusion does not, in this case, pose a problem in terms of population frontier identification provided that they are independent of the efficiency. Moreover, the inclusion of the sample weights through the adaptation of the order-α model leads to larger MSE values.

Conversely, when the auxiliary variable in the PPS design is informative, i.e. it is correlated with the efficiency of the units in the population (e.g. larger schools are more efficient than smaller ones), failure to include the probability of inclusion information in the model significantly impairs the estimation of the population frontier for all levels of α and sample sizes. Figures 2 and 3 illustrate this result. Note that if we compare both PPS informative designs, the most pronounced improvements of considering the sample weights are observed in the *informative* scenario, because all the units included in the sample are used to estimate the frontier in this case. However, only half of the sample is used to identify the population frontier in the *two-stage* dataset.

Figure 1 Estimation of the population production frontier using the order-α and AGSS models, *non-informative* design

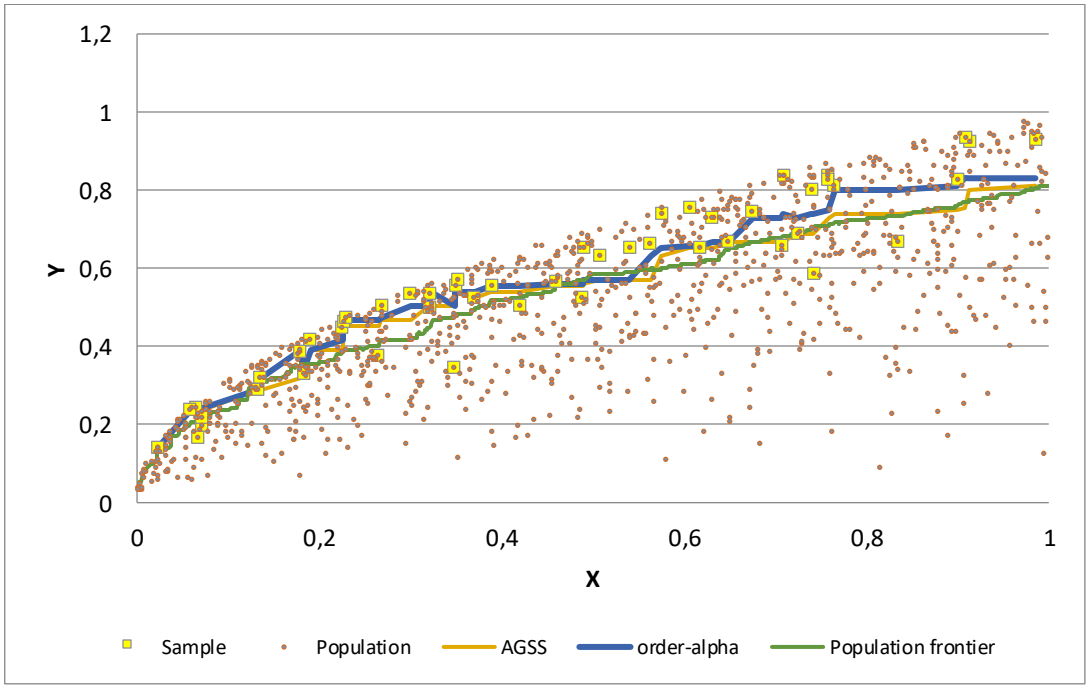


Panel (a) Non-informative design n=50 *and* $\alpha = 0.9$

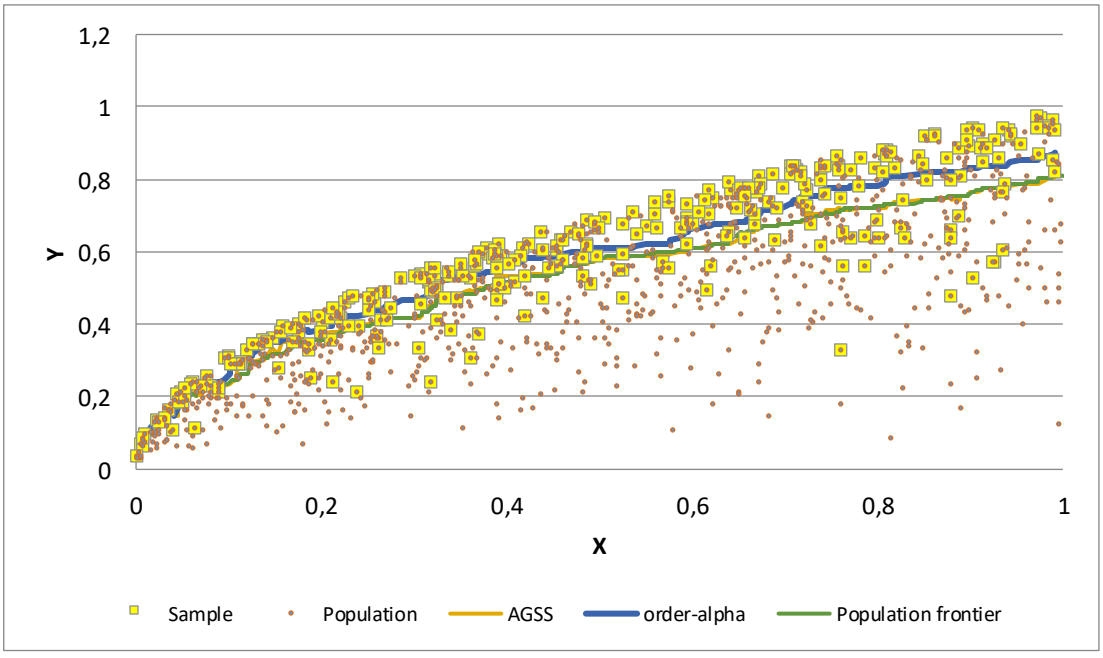


Panel (b) Non-informative design n=300 *and* $\alpha = 0.9$

Figure 2 Estimation of the population production frontier using the order-α and AGSS models, *informative* design

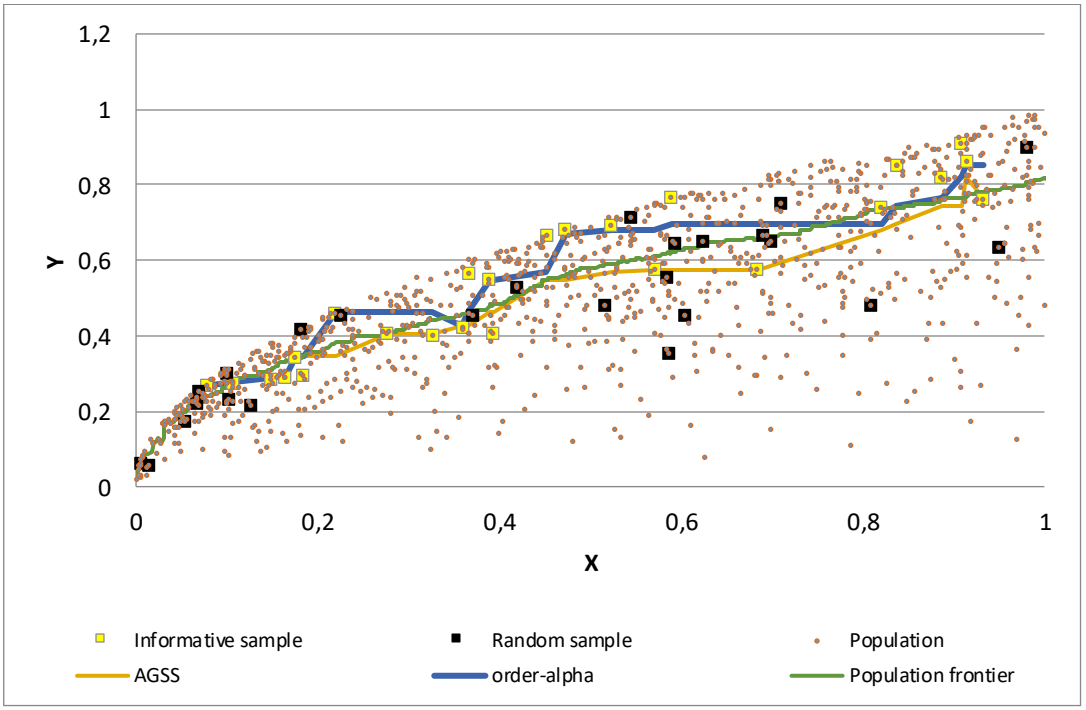


Panel (a) Informative design n=50 *and* $\alpha = 0.9$

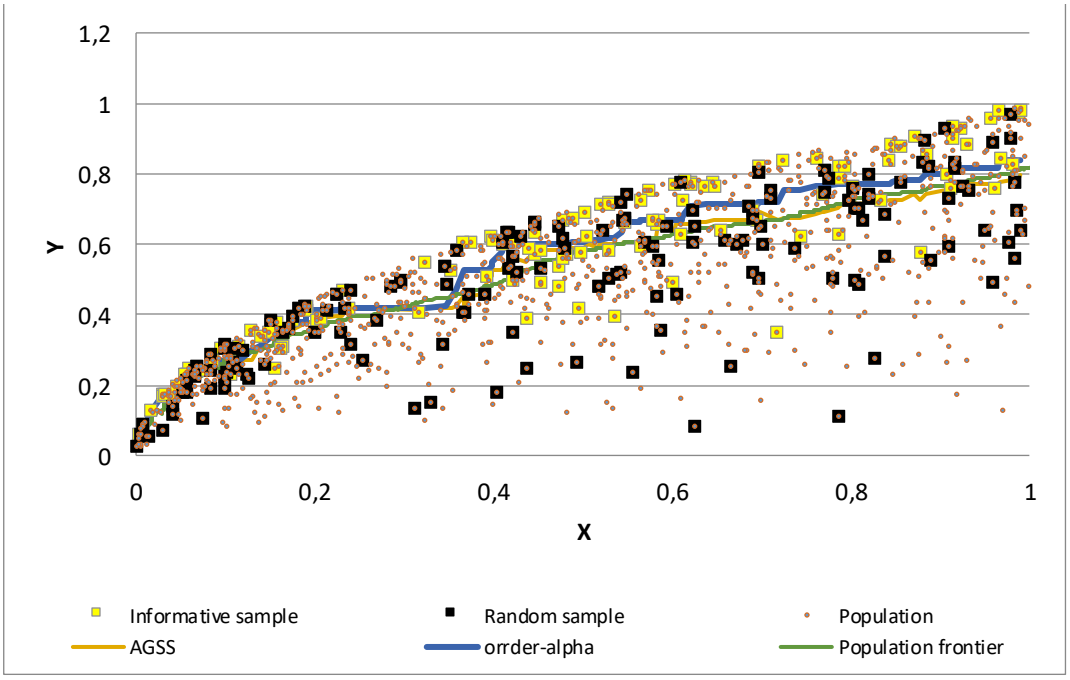


Panel (b) Informative design n=300 *and* $\alpha = 0.9$

Figure 3 Estimation of the population production frontier using the order-α and AGSS models, *two-stage* design



Panel (a) Two-stage design n=50 *and* $\alpha = 0.9$



Panel (b) Two-stage design n=300 *and* $\alpha = 0.9$

Finally, an interesting finding in both *informative* designs for $\alpha < 1$ is that the MSE also increases as the sample size increases. This means that the negative effect of omitting information about different sample weights across the sample intensifies in these contexts as the sample size increases. The accurate estimation of the population production frontier is extremely important when we set out to measure technological change over time or compare the performance between different sectors or groups of units. In these contexts, if there is any previous evidence about the potential correlation between the auxiliary variable and population efficiency, it would be recommendable to use the AGSS model instead of overlooking the probabilities of inclusion. In fact, since there is not a substantial difference in the MSE between both models in the *non-informative* scenario, it might be preferable, if there is any inkling, even if there is no robust evidence, of such a correlation, to include, rather than omit, the sample weights.

Table 1 Mean square error for the estimation of the population frontier from PPS sample designs

	α=0.8		α=0.9		α=1	
	order-α	AGSS	order-α	AGSS	order-α	AGSS
Non-informative						
n=50	0.095	0.103	0.089	0.095	0.168	0.168
n=100	0.093	0.102	0.088	0.096	0.152	0.152
n=300	0.079	0.087	0.074	0.082	0.102	0.102
n=500	0.060	0.070	0.055	0.063	0.063	0.063
Informative						
n=50	0.250	0.144	0.156	0.102	0.052	0.052
n=100	0.400	0.159	0.241	0.099	0.044	0.044
n=300	1.028	0.200	0.579	0.104	0.015	0.015
n=500	1.626	0.226	0.910	0.093	0.000	0.000
Two-stage						
n=50	0.161	0.133	0.104	0.083	0.054	0.054
n=100	0.245	0.151	0.151	0.090	0.053	0.053
n=300	0.549	0.192	0.322	0.094	0.035	0.035
n=500	0.838	0.169	0.481	0.089	0.022	0.022

Note: Mean values after 100 replications.

4.2.2. Results on the population average efficiency

We are also interested in exploring the effect of taking into account the existence of different probabilities of inclusion π_j when we aggregate the individual efficiencies to estimate the population average efficiency μ. To do this, we compute, after the 100 replications, the mean bias relative to the true population average efficiency μ for each estimator $\hat{\mu}$ (Equations 19 to 22):

$$Bias = \frac{\sum_{b=1}^{100} \hat{\mu}_b - \mu_b}{100}$$

and the MSE relative to the true population efficiency μ:

$$MSE = \frac{\sum_{b=1}^{100} \left(\hat{\mu}_b - \mu_b\right)^2}{100}. \qquad (24)$$

Note that, for $\alpha < 1$, some observations will be located above the quantile frontier $q_\alpha(x)$, i.e. the efficiency level of these units will be lower than 1. Then, the parameter μ and the estimators $\hat{\mu}$ could also take values smaller than 1, leading to a positive or a negative bias. Results for the bias and the MSE are shown in Tables 2 and 3, respectively.

As in the previous case, for $\alpha = 1$, the results from both models are equal. Thus, we will focus on $\alpha < 1$. In the first scenario, the *non-informative* design, both the bias and the MSE results show that omission of the sample weights is, in this case, the best strategy for estimating the population average efficiency for all sample sizes and levels of α.

In the *informative* design, the bias and MSE lead to different conclusions. In terms of bias, it appears to be better to take into account the information on the probabilities of inclusion to estimate the population frontier and then aggregate the individual efficiencies to estimate the population average efficiency. This result holds for all sample sizes and both levels of $\alpha = 0.8$ and $\alpha = 0.9$. However, if we focus on the MSE, the conclusion is the exact opposite. In this case, the estimator $\hat{\mu}^{order-\alpha}$ performs best for all sample sizes and levels of α, which means that it is more accurate to ignore this information. Note that, in this scenario, there is a trade-off between population frontier estimation accuracy and population average efficiency. With a view to population frontier estimation accuracy, it would be necessary to include the sample weights (i.e. using the AGSS model). However, this implies a considerable deterioration in terms of the MSE in the estimation of the population average efficiency.

Finally, the *two-stage* design addresses this trade-off. In this context, it is also more accurate to include the probabilities of inclusion in the model for estimating the population average efficiency for all sample sizes and levels of α, and there are no contradictory results between

bias and MSE. Moreover, the estimation of the population average efficiency μ from this sampling design is much more accurate than the *informative* design, regardless of the estimator that we use. This finding is notable in terms of public policy design, since large-scale assessment surveys are usually used to measure technical efficiency through production frontiers. However, current sample designs (PPS) are not designed for estimating either population production frontiers or average technical efficiency. If there is any previous evidence (e.g. earlier studies) indicating that there is any correlation between the auxiliary variable (e.g. school size) and the efficiency of the schools in the population, it would be advisable to define a *two-stage* sampling design instead of a standard PPS to enhance future population efficiency and productivity estimations using samples. This issue is even more important when the aim of the analysis is to compare school performance over time (i.e. technological change) or different educational sectors (e.g. public and private schools).

Table 2 Bias for the estimation of the population average efficiency

	Non-informative				***Informative***				***Two-stage***			
	order-α	order-α (π)	AGSS	AGSS(π)	order-α	order-α (π)	AGSS	AGSS(π)	order-α	order-α (π)	AGSS	AGSS(π)
α=0.8												
n = 50	-0.002	-0.008	0.006	0.002	-0.107	0.135	-0.194	-0.028	0.054	---	-0.032	---
n = 100	-0.017	-0.012	-0.015	-0.009	-0.115	0.084	-0.208	-0.003	0.087	---	0.005	---
n = 300	-0.011	-0.009	-0.012	-0.010	-0.120	0.086	-0.208	-0.041	0.101	---	-0.008	---
n = 500	-0.010	-0.009	-0.012	-0.010	-0.121	0.068	-0.211	-0.054	0.108	---	0.001	---
α=0.9												
n = 50	-0.020	-0.026	-0.017	-0.021	-0.151	0.108	-0.215	-0.017	0.004	---	-0.054	---
n = 100	-0.032	-0.026	-0.029	-0.023	-0.158	0.054	-0.226	0.017	0.045	---	-0.017	---
n = 300	-0.022	-0.020	-0.022	-0.020	-0.163	0.062	-0.227	-0.031	0.065	---	-0.014	---
n = 500	-0.017	-0.016	-0.022	-0.020	-0.163	0.041	-0.229	-0.048	0.078	---	-0.004	---
α=1												
n = 50	-0.151	-0.157	-0.151	-0.157	-0.313	-0.036	-0.313	-0.036	-0.189	---	-0.189	---
n = 100	-0.122	-0.115	-0.122	-0.115	-0.299	-0.067	-0.299	-0.067	-0.114	---	-0.114	---
n = 300	-0.069	-0.067	-0.069	-0.067	-0.280	-0.028	-0.280	-0.028	-0.053	---	-0.053	---
n = 500	-0.046	-0.045	-0.046	-0.045	-0.274	-0.044	-0.274	-0.044	-0.024	---	-0.024	---

Note: Mean values after 100 replications.

Table 3 Mean square error for the estimation of the population average efficiency

	Non-informative				***Informative***				***Two-stage***			
	order-α	order-α (π)	AGSS	AGSS(π)	order-α	order-α (π)	AGSS	AGSS(π)	order-α	order-α (π)	AGSS	AGSS(π)
α=0.8												
n = 50	0.005	0.007	0.006	0.008	0.000	0.274	0.047	0.087	0.030	---	0.027	---
n = 100	0.003	0.004	0.003	0.004	0.018	0.155	0.053	0.437	0.017	---	0.011	---
n = 300	0.001	0.001	0.001	0.001	0.019	0.090	0.051	0.049	0.013	---	0.004	---
n = 500	0.000	0.000	0.001	0.001	0.019	0.040	0.052	0.022	0.013	---	0.003	---
α=0.9												
n = 50	0.006	0.009	0.006	0.010	0.032	0.313	0.061	0.145	0.028	---	0.030	---
n = 100	0.003	0.004	0.003	0.004	0.035	0.173	0.066	0.490	0.012	---	0.011	---
n = 300	0.001	0.001	0.001	0.001	0.036	0.105	0.065	0.074	0.007	---	0.003	---
n = 500	0.000	0.001	0.001	0.001	0.036	0.044	0.066	0.032	0.007	---	0.002	---
α=1												
n = 50	0.022	0.027	0.022	0.027	0.123	0.342	0.123	0.342	0.065	---	0.065	---
n = 100	0.012	0.012	0.012	0.012	0.113	0.220	0.113	0.220	0.024	---	0.024	---
n = 300	0.003	0.003	0.003	0.003	0.101	0.137	0.101	0.137	0.006	---	0.006	---
n = 500	0.001	0.001	0.001	0.001	0.097	0.061	0.097	0.061	0.002	---	0.002	---

Note: Mean values after 100 replications.

5. Concluding remarks

Nowadays, it is quite common to find educational databases based on complex sampling designs used to minimize survey costs and improve the precision of the estimates of some parameters of interest for the population. However, the use of the information provided by sample weights has been repeatedly overlooked in the literature on production frontier estimation, leading to estimations (from sample data) that are not representative of the population under study. In this research, we develop an extension of robust nonparametric order-α frontier methods to incorporate sample weight information into the estimation of the population production frontier. Monte Carlo results show that when the auxiliary variable in the PPS sample design contains information about the level of efficiency in the population, the estimation of the population frontier can be improved if the nonparametric model accounts for information on sample weights. In this context, however, the PPS sample design should be transformed into a *two-stage* sampling design in order to properly estimate the average educational efficiency for the target population.

This research should be regarded as a foundation stone for addressing the issue of incorporating sample weight information into the estimation of technical efficiency. More research is needed in several directions to explore other potential solutions for improving the accuracy of nonparametric estimations. Probably, the most straightforward and intuitive alternative is to explore the potential of incorporating sample weight information into the conventional bootstrap methodology (Simar and Wilson, 1998). In particular, it is important to test its validity and performance, since the basic assumption of this method (i.e. observed data in the sample come from independent and identically distributed random variables) does not hold in the case of complex sampling designs in finite populations. Another fruitful line of research would be to address this issue in the parametric framework, for example, by incorporating sample weights into the corrected ordinary least square (COLS) model.

Acknowledgments

The authors are greatly indebted to Fundación Ramon Areces for supporting and encouraging mutual collaboration on productivity analysis in education as part of projects *La medición de la eficiencia de la educacion primaria y de sus determinantes en España y en la Union Europea: un análisis con TIMSS-PIRLS 2011* (D. Santín) and *Evaluación de la eficiencia en la producción educativa a partir de diseños muestrales* (J. Aparicio, M. González and G. Sicilia).

References

Afonso, A., & Aubyn, M. S. (2005). Non-parametric approaches to education and health efficiency in OECD countries. *Journal of Applied Economics*, 8(2), 227-246.

Afonso, A., & Aubyn, M. S. (2006). Cross-country efficiency of secondary education provision: A semi-parametric analysis with non-discretionary inputs. *Economic Modelling*, 23(3), 476-491.

Agasisti, T., and Zoido, P. (2018). "Comparing the Efficiency of Schools through International Benchmarking: Results from an Empirical Analysis of OECD PISA 2012 Data". *Educational Researcher*, 47(6), 352-362.

Aigner, D.J. and Chu, S.F. (1968). "On Estimating the Industry Production Function". *American Economic Review*, 58, 826–839.

Aigner, D.J., Lovell, C.A.K., and Schmidt, P. (1977). "Formulation and estimation of stochastic frontier production functions". *Journal of Econometrics*, 6, 21–37.

Allen, R., Athanassopoulos, A., Dyson, R.G. and Thanassoulis, E. (1997). "Weights restrictions and value judgements in Data Envelopment Analysis: Evolution, development and future directions". *Annals of Operations Research*, 73, 13-34.

Aparicio, J., Cordero, J. M., and Pastor, J. T. (2017a). "The determination of the least distance to the strongly efficient frontier in data envelopment analysis oriented models: modelling and computational aspects". *Omega*, 71, 1-10.

Aparicio, J., Crespo-Cebada, E., Pedraja-Chaparro, F., and Santín, D. (2017b). "Comparing school ownership performance using a pseudo-panel database: A Malmquist-type index approach". *European Journal of Operational Research*, 256(2), 533-542.

Aparicio, J., Cordero, J. M., Gonzalez, M., and Lopez-Espin, J.J. (2018). "Using non-radial DEA to assess school efficiency in a cross-country perspective: An empirical analysis of OECD countries". *Omega*, 79, 9-20.

Aparicio, J., and Santin, D. (2018). "A note on measuring group performance over time with pseudo-panels". *European Journal of Operational Research*, 267(1), 227-235.

Aragon, Y., Daouia, A., and Thomas-Agnan, C. (2005). "Nonparametric frontier estimation: a conditional quantile-based approach". *Econometric Theory*, 21(2), 358–389.

Banker, R.D., Charnes, A., and Cooper, W.W. (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis". *Management Science* 30(9), 1078–92.

Barro, R.J. (2001). “Human Capital and Growth”. *The American Economic Review*, 91(2), 12-17.

Barro, R.J., and Lee, J.-W. (2012) “A New Data Set of Educational Attainment in the World, 1950–2010”. *Journal of Development Economics*, 104, 184–98.

Cazals, C., Florens, J. P., and Simar, L. (2002). “Nonparametric frontier estimation: a robust approach”. *Journal of Econometrics*, 106(1), 1–25.

Charnes, A., Cooper, W.W., and Rhodes, E. (1978). “Measuring the Efficiency of Decision Making Units”. *European Journal of Operational Research*, 2(6), 429–44.

Charnes, A., Cooper, W.W., and Rhodes, E. (1981). “Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through”. *Management Science*, 27(6), 668–97.

Coelli, T.J., Rao, D.S.P., O'Donnell, C.J., and Battese, G.E. (2005). *An Introduction to Efficiency and Productivity Analysis*. Springer, New York.

Cordero, J. M., Cristobal, V., and Santín, D. (2018). “Causal inference on education policies: a survey of empirical studies using PISA, TIMSS and PIRLS”. *Journal of Economic Surveys*, 32(3), 878-915.

Daouia, A. and Simar, L. (2007). “Nonparametric efficiency analysis: A multivariate conditional Quantile”. *Journal of Econometrics*, 140, 375-400.

Daraio, C. and Simar, L. (2007). “Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach”. *Journal of Productivity Analysis*, 28: 13-32.

De Jorge, J., and Santín. (2010). “Determinantes de la eficiencia educativa en la Unión Europea”. *Hacienda Pública Española*, 193, 131-155.

De La Fuente, A. (2011). “Human capital and productivity”. *Nordic Economic Policy Review*, 2(2), 103-132.

Debreu, G. (1951). “The Coefficient of Resource Utilization”, *Econometrica*, 19(3), 273–92.

Deprins, D., L. Simar, and H. Tulkens (1984). “Measuring Labor Inefficiency in Post Offices”. In The Performance of Public Enterprises: Concepts and Measurements, ed. M. Marchand, P. Pestieau, and H. Tulkens, 243–267. Amsterdam: North-Holland.

De Witte, K., and Lopez-Torres, L. (2017). “Efficiency in education: a review of literature and a way forward”. *Journal of the Operational Research Society*, 68(4), 339–363.

Färe, R., Grosskopf, S., and Lovell, C.A.K. (1985). *The Measurement of Efficiency of Production*. Kluwer Nijhof Publishers, Boston, Massachusetts.

Färe, R. and Zelenyuk, V. (2003). "On aggregate Farrell efficiencies". *European Journal of Operational Research*, 146, 615-620.

Farrell, M.J. (1957). "The Measurement of Productive Efficiency". *Journal of the Royal Statistical Society. Series A (General)*, 120(3), 253–90.

Hanushek, E.A. (1979). "Conceptual and Empirical Issues in the Estimation of Educational Production Functions". *Journal of human Resources*, 351–88.

Hanushek, E.A., and Kimko, D.D. (2000). "Schooling, Labor-Force Quality, and the Growth of Nations". *American Economic Review*, 90(5), 1184–208.

Hanushek, E.A., and Woessmann, L. (2012). "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation". *Journal of Economic Growth*, 17(4), 267–321.

Hedayat, A.S., and Sinha, B.K. (1991). *Design and inference in finite population sampling*. Wiley.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., and Shure, N. (2017). "What happens when econometrics and psychometrics collide? An example using the PISA data". *Economics of Education Review*, 61, 51-58.

Johnes, J. (2015). "Operational Research in Education". *European Journal of Operational Research*, 243(3), 683–696.

Koopmans, T.C. (1951). "Analysis of Production as an Efficient Combination of Activities". In T.C. Koopmans (Ed.), Activity Analysis of Production and Allocation, New York, John Wiley, 33–97.

Lavy, V. (2015). "Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries". *The Economic Journal*, 125, 397- 424

Levin, H.M. (1974). "Measuring Efficiency in Educational Production". *Public Finance Quarterly*, 2(1), 3–24.

Meeusen, W. and van den Broeck, J. (1977). "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error". *International Economic Review*, 18, 435–444.

OECD (2017) PISA 2015 Technical Report, OECD.

Pastor, J. T., Lovell, C. K., and Aparicio, J. (2012). “Families of linear efficiency programs based on Debreu’s loss function”. *Journal of Productivity Analysis*, 38(2), 109-120.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Science and Business Media.

Shephard, R. W. (1953). *Cost and Production functions*. Princeton University Press.

Simar, L., and Wilson, P.W. (1998). “Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models”. *Management science*, 44(1), 49–61.

Strietholt, R., Gustafsson, J.E., Rosen, M. and Bos, W. (2014). “Outcomes and causal inference in international comparative assessments”. In R. Stietholt, W. Bos, J.E. Gustafsson and M. Rosen (eds.), *Educational Policy Evaluation through International Comparative Assessments*. New York: Waxman, Münster.

Worthington, A.C. (2001). “An Empirical Survey of Frontier Efficiency Measurement Techniques in Education”. *Education Economics*, 9(3), 245–68.