mathematics

Article

Efficiency Analysis with Educational Data: How to Deal with Plausible Values from International Large-Scale Assessments

Juan Aparicio [1], Jose M. Cordero [2,*] and Lidia Ortiz [1]

[1] Center of Operations Research (CIO), University Miguel Hernandez of Elche (UMH), 03202 Elche, Spain; j.aparicio@umh.es (J.A.); lidia.ortiz@umh.es (L.O.)

[2] Department of Economics, University of Extremadura (UEX), 06006 Badajoz, Spain

* Correspondence: jmcordero@unex.es

Abstract: International large-scale assessments (ILSAs) provide several measures as a representation of educational outcomes, the so-called plausible values, which are frequently interpreted as a representation of the ability range of students. In this paper, we focus on how this information should be incorporated into the estimation of efficiency measures of student or school performance using data envelopment analysis (DEA). Thus far, previous studies that have adopted this approach using data from ILSAs have used only one of the available plausible values or an average of all of them. We propose an approach based on the fuzzy DEA, which allows us to consider the whole distribution of results as a proxy of student abilities. To assess the extent to which our proposal offers similar results to those obtained in previous studies, we provide an empirical example using PISA data from 2015. Our results suggest that the performance measures estimated using the fuzzy DEA approach are strongly correlated with measures calculated using just one plausible value or an average measure. Therefore, we conclude that the studies that decide upon using one of these options do not seem to be making a significant error in their estimates.

Keywords: data envelopment analysis; fuzzy; PISA; plausible values

check for updates

Citation: Aparicio, J.; Cordero, J.M.; Ortiz, L. Efficiency Analysis with Educational Data: How to Deal with Plausible Values from International Large-Scale Assessments. *Mathematics* **2021**, *9*, 1579. https://doi.org/10.3390/math9131579

Academic Editors: Massimiliano Ferrara, Ana M. Aguilera, Sara M. González-Betancor and Carmen Pérez-Esparrells

Received: 21 May 2021
Accepted: 30 June 2021
Published: 5 July 2021

1. Introduction

Since the pioneering work carried out in the early days of educational economics, the exploration of the educational and societal factors that affect educational attainment has attracted the interest of researchers working in this field. Several authors have named this educational effectiveness research [1,2], which in recent years has been growing rapidly due to the development of new learning methods and the use of information technologies [3,4].

In this context, the rich and extensive information provided by large-scale international assessments (hereafter ILSAs) has become a very useful tool for analyzing the performance of education systems around the world and promoting reforms of national education policies [5,6]. Although the Program for International Student Assessment (PISA) conducted by the OECD is undoubtedly the best known, several other studies can also be included within this group, including the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), both coordinated by the International Association for the Evaluation of Educational Achievement (IEA).

The technical complexities of these large-scale assessments often create multiple challenges for applied researchers when analyzing their data [7] (Sjøberg [8] described the data generated by PISA as a playground for psychometricians). In this paper, we focus on the construction of measures representing the competences of students, a very relevant issue that has been generating enormous interest in the recent literature [9]. On this issue, it is widely assumed that the competences of students cannot be measured with a simple score derived from the answers provided by students in standardized tests, but by several values randomly obtained from the distribution function of test results, known in the psychometric literature as plausible values. These values should be interpreted as the representation of the ability range for each student [10], which is very difficult to measure in a precise way.

The procedure for handling these values in econometric analyses (where they are frequently included as the dependent variable) is well known for users of these databases, since technical reports and user manuals of international databases describe it in detail [11–13]. As we explain in Section 2.2, this procedure requires estimating the model several times: once using each of the plausible values, and then by computing the average of the estimates. However, this issue has been overlooked in studies using these data to estimate efficiency measures of performance using frontier methods, such as data envelopment analysis (DEA) (this nonparametric approach is the most used in these types of studies due to its greater flexibility [14]), which allows the best practices that can serve as benchmarks for the rest to be identified. Such studies have become increasingly common due to restrictions on increasing public spending in education in most countries; thus, policy makers and researchers are very interested in identifying the references that can provide them with some guidelines to improve student and school outcomes without spending more money in educational resources.

In this framework, plausible values are often used as the output measures representing educational outcomes. Nevertheless, researchers often forget that, when selecting a single plausible value or the average of all available plausible values, some of the available information about students' abilities is overlooked. This is because, in contrast to econometric analysis, when applying traditional frontier methods, it is not possible to account for the fact that the plausible values represent a distribution of results and, therefore, out measures are imprecise.

In this paper, we intend to bridge this knowledge gap by proposing an innovative way of dealing with imprecise data in output measures when applying frontier methods to data from ILSAs. This approach is based on the notion of fuzziness [15] and, more precisely, we rely on the so-called fuzzy DEA (FDEA), which allows for the consideration of the whole distribution of test scores into the estimation of efficiency measures. Among all the many existing approaches to implement FDEA [16], we apply the methodology proposed by Kao and Liu [17], which basically transforms the fuzzy "radial" DEA model to several conventional DEA models by applying the so-called α-cut procedure.

We apply this technique to assess the performance of Spanish students participating in PISA 2015. In our model, we include three outputs (test scores in mathematics, reading and sciences) that are treated as fuzzy in an FDEA model since they are represented by ten different plausible values. Since the estimated efficiency measures are represented using membership functions, it is possible to provide more precise information for decision making.

The research problem that we intend to solve is whether the consideration of all the information provided by plausible values can lead to different results than those obtained when using traditional methods, which only incorporate limited information to represent educational output. Thus, the efficiency estimates derived from the application of the FDEA approach are compared with the estimates obtained with the traditional DEA method using one single plausible value or the average of plausible values.

The remainder of the paper is organized as follows. In Section 2, we explain the concept of plausible values and some guidelines for handling them in empirical studies. Section 3 introduces the necessary notation and background of the proposed FDEA methodology. Subsequently, in Section 4, we explain the main characteristics of our dataset and the variables selected for our empirical analysis. In Section 5, we present and discuss the main results comparing the estimates obtained with the proposed FDEA approach with those obtained with the traditional DEA approach. Finally, we present the main conclusions in Section 6.

2. Plausible Values and How to Use Them in Empirical Analyses

2.1. What Are Plausible Values?

Measuring cognitive outcomes is one of the main concerns of international large-scale assessments, such as PISA, TIMSS or PIRLS. The assessment instrument often includes multiple questions or items and is carried out in a limited period of time with the aim of providing comparable information about the ability of students and their knowledge in different domains, such as reading, mathematics or science. They rely on a complex psychometric design that involves considering different sub-domains within each subject area. This results in an enormous amount of test material to be covered, since it is not possible to ask every pupil all the available questions. Therefore, all the tested items are divided into multiple blocks or clusters. Since test administration can only take place during a maximum of two hours (this limitation on testing time is based on considerations with respect to reducing student burden, minimizing interruptions in the school schedule and other financial and/or time constraints), students are randomly assigned to complete one particular test booklet, each of which includes a subset of tested items consisting of items from one or more clusters; thus, he/she responds to only a fraction of what constitutes the total assessment pool [18].

As pupils answer only a limited number of questions from the total test item pool, the measurement of individual proficiency is achieved with a substantial amount of measurement error [19]. Thus, traditional methods of estimating individual proficiency would result in biased or inconsistent variance estimates. As an alternative to this problem, plausible value methods are employed as a viable technique to generate proficiency estimates from the limited fraction of administered cognitive items and student background information. Since student ability is not directly observed, it is a latent variable (or latent ability) that could be treated as a missing value; thus, it is necessary to use multiple imputation methods [20,21] to estimate the distribution of proficiency for each student in each subject area (see [22–24] for further details).

Plausible values can be defined as several random values drawn from the distribution of proficiency estimates [11,25]. They are used by applied researchers for different purposes, such as estimating the plausible range and the location of proficiency for groups of students or exploring the relationship between proficiency and various social and educational variables in secondary analysis. However, it is worth mentioning that plausible values are not individual scores in the traditional sense and should, therefore, not be analyzed as multiple indicators of the same score or latent variables [26].

In practice, five plausible values are reported for each student in overall mathematics and science (TIMSS) and another five for each student in overall reading (PIRLS), since this number is sufficient for the accurate estimation of population-level statistics [27]. However, in recent years, some of the large-scale assessment surveys have increased the number of plausible values with the aim of providing better estimates of the variability when a large amount of imputation is required. For example, ten plausible values were used for PISA 2015 and PISA 2018 for each domain (mathematics, reading and science) [28], while only five plausible values were used for PISA 2000 to PISA 2012. There are very few studies that have provided justifications for increasing the number of plausible values used beyond five. One of the few exceptions is the recent work by Bibby [29], which concludes that the sample size has a larger impact on the estimations of population parameters than an increase in the number of plausible values used. The reported plausible values are in scale scores with a mean of 500 and standard deviation of 100 overall, across all participating countries.

2.2. How to Use Plausible Values in Secondary Analyses

Researchers should be aware that secondary analyses should be performed independently on each of the available plausible values so that they can provide appropriate estimates of population statistics, such as means and variances [30]. Specifically, the correct procedure for handling the plausible values provided in the international achievement databases can be divided into four steps, based on the original work of [20]: (i) estimate the statistic/model of interest five times (or ten in the last waves of PISA) using each of the plausible values to obtain five (or ten) separate parameter estimates (β_pv) and the corresponding estimates of the sampling error (σ_pv); (ii) calculate the average of those estimates; (iii) estimate the magnitude of the imputation error; and (iv) calculate the value of the final standard error by combining the average sampling error and the imputation error (note that the secondary analysis model is typically a subset of the latent regression model used to generate the plausible values [31]). Finally, the final parameter estimates and their standard error can be used to conduct hypothesis tests and construct confidence intervals following the usual methods. In order to facilitate the correct implementation of this procedure in secondary analyses, most software specialized in data processing has specific routines or commands to perform estimations with plausible values (for instance, PV [32] or REPEST [33] in Stata).

Although this procedure is clearly described in technical reports and user manuals of different international large-scale assessments, in some empirical studies dealing with plausible values, it is common to find two different shortcuts in the implementation of secondary analyses with econometric techniques, both of which are incorrect. First, analysts often choose to use just one of the five plausible values. With this option, the standard errors of the statistics of interest are generally underestimated, as the uncertainty associated with the measurement of proficiency distributions is ignored. However, PISA analysts indicate that using one or five plausible values in a large sample does not really make a substantial difference [34]. In fact, during the exploratory phase of the data, statistical analyses can be based on a single plausible value, although it is highly recommended to use all the available values in order to improve the accuracy of the estimates, even for large samples.

The second shortcut employed by some authors is to calculate the average of the existing plausible values (five or ten) and use it as if it were the only available estimate of student performance. The main problem with the calculation of this average value as a proxy for performance is that standard errors are severely underestimated (particularly if only a single plausible value is being used), which might lead to misleading results. Therefore, the mean of the available plausible values should never be used in empirical analyses with econometric techniques [19].

2.3. Plausible Values in Efficiency Analyses

In contrast to the existing clear instructions for secondary studies with econometric techniques, the way to proceed when using frontier methods in empirical studies exploiting data from ILSAs generates much more doubt for researchers. In these studies, test scores (or rather plausible values) achieved by students in different domains are usually identified as proxies for the educational output (see Cordero et al. [35] for a review). However, they follow different criteria to deal with plausible values. The most common option consists of using only one plausible value [36–40], usually the first of all available values, i.e., the first shortcut identified in the previous section. Likewise, other studies follow the second shortcut and use the average of all plausible values to estimate efficiency measures [41,42]. Finally, there is a more cumbersome process, which implies estimating one efficiency score for each plausible value and, subsequently, calculating the average efficiency score [43,44].

The main problem of all the aforementioned alternatives is that none of them explicitly considers the fact that the output measures are imprecise since they treat plausible values as crisp values (precise measurements) when they are actually representing a distribution of results. The results obtained from using these crisp values may not adequately reflect the performance of the units evaluated because some of the available information about students' abilities is overlooked.

The present study attempts to shed light on this issue by exploring the extent to which the use of these procedures may affect the results and, more specifically, present an innovative method that allows the incorporation of data on the whole distribution of results into the estimation of efficiency measures of educational performance. This method is described in depth in the following section.

3. Methodology

First, we provide a brief description of the traditional DEA model, before going on to explain the method we propose to use in our empirical analysis. Therefore, let us consider n DMUs to be evaluated (such as students within our frame of reference). DMU$_j$ consumes $x_j = (x_{1j}, \ldots, x_{mj}) \in R^m_+$, $x_j \neq 0_m$ amounts of inputs to generate $y_j = (y_{1j}, \ldots, y_{sj}) \in R^s_+$, $y_j \neq 0_s$, amounts of outputs. From the data, it is possible to construct an estimation of the underlying technology from which the DMUs were generated. The technology represents the set of all feasible input–output bundles $(x, y) \in R^{m+s}_+$ [45]. Data Envelopment Analysis provides the following estimator of the technology under variable returns to scale (VRS):

$$T_{VRS} = \left\{ (x,y) \in R^m_+ \times R^s_+ : x \geq \sum_{j=1}^{n} \lambda_j x_j, y \leq \sum_{j=1}^{n} \lambda_j y_j, \sum_{j=1}^{n} \lambda_j = 1;\ 0 \leq \lambda_j \leq 1;\ j = 1, \ldots, n \right\} \tag{1}$$

Technical inefficiency can then be calculated as the distance from a DMU to the border of the estimated technology. This distance is usually implemented in practice through two types of approaches: input-oriented models and output-oriented models. For convenience, we only present the output-oriented DEA approach, since this is what we use in our practical example, on the premise that students are always trying to improve their results. Output-oriented models assume that each DMU strives to maximize outputs while using the same level of inputs. Among the existing output-oriented models, the output-oriented radial model is probably the most famous. This approach keeps the inputs of the assessed unit constant but equi-proportionally augments the bundle of outputs. In this way, the output-oriented radial model assuming variable returns to scale for evaluating the unit (x_0, y_0) can be mathematically implemented through the following linear optimization program.

$$\begin{array}{lllll} Max & \phi_0 & & & \\ s.t. & & & & \\ & \sum_{j=1}^{n} \lambda_{j0} x_{ij} \leq x_{i0}, & i = 1, \ldots,\ m & (2.1) & \\ & \sum_{j=1}^{n} \lambda_{j0} y_{rj} \geq \phi_0 y_{r0}, & r = 1, \ldots, s & (2.2) & , \\ & \sum_{j=1}^{n} \lambda_{j0} = 1, & & (2.3) & \\ & \lambda_{j0} \geq 0, & j = 1, \ldots, n & (2.4) & \end{array} \tag{2}$$

The optimal value ϕ_0^* of the above program is the efficiency scores associated with the unit (x_0, y_0). It can be proved that $\phi_0^* > 1$, and with $\phi_0^* = 1$ signaling, that the evaluated DMU is technically efficient. Otherwise, $\phi_0^* > 1$, and there is leeway for improving the outputs while using the same level of inputs.

Additionally, the dual program of (2) is the following linear program:

$$\begin{array}{lllll} Min & \sum_{i=1}^{m} v_{i0} x_{i0} - \pi_0 & & & \\ s.t. & & & & \\ & \sum_{r=1}^{s} u_{r0} y_{r0} = 1, & & (3.1) & \\ & \sum_{i=1}^{m} v_{i0} x_{ij} - \sum_{r=1}^{s} u_{r0} y_{rj} - \pi_0 \geq 0, & j = 1, \ldots, n & (3.2) & . \\ & v_{i0} \geq 0, & i = 1, \ldots, m & (3.3) & \\ & u_{r0} \geq 0, & r = 1, \ldots, s & (3.4) & \end{array} \tag{3}$$

Usual approaches assume that input and output values are precise information (also called "crisp"). However, it may be that the observed values of the variables (inputs and outputs) are imprecise. Within the framework of our research, we elaborate in Section 2 on how student abilities cannot be defined by a unique value (test score), but rather a range or distribution from which we extract different random values. These values (or their average) are frequently used in model (2) to solve the "crisp" traditional DEA radial model. However, they may not necessarily be the most appropriate representation of students' abilities, especially if there is a high spread or variation in the distribution of results. Therefore, it might happen that two students presenting similar values for a single plausible value (or even similar average plausible values) may exhibit a differing pattern with respect to the unobserved distribution of results; so, when it comes to technical efficiency, they could be categorized differently.

Data imprecision may be included in DEA efficiency models through Fuzzy Data Envelopment Analysis (FDEA), where scores in different competences are included as fuzzy numbers, as opposed to crisp numbers, in the DEA model. Emrouznejad et al. [46] provide a taxonomy and review of the FDEA methods in six categories and that can be found in the literature, namely, the fuzzy ranking approach, the fuzzy random/type-2 fuzzy set, the tolerance approach, the fuzzy arithmetic, the possibility approach, and the α-level based approach. Among them, the α-level based approach is probably the most popular FDEA model. In particular, the one proposed by Kao and Liu [17] is the most popular and the most applied in empirical studies. Consequently, this is the option we use in our analysis. We now review the main characteristics of this approach.

The model of Kao and Liu [17] is based on the notion of α-cuts, also known as α-possibility level sets, and the transformation of the FDEA model into a set of standard crisp DEA models. In this framework, $\left(\widetilde{x}_{ij}\right)_\alpha = \left\{z : \mu_{\widetilde{x}_{ij}}(z) \geq \alpha\right\}$ and $\left(\widetilde{y}_{rj}\right)_\alpha = \left\{z : \mu_{\widetilde{y}_{rj}}(z) \geq \alpha\right\}$ represent the α-cut of $\widetilde{x}_{ij}$ and $\widetilde{y}_{rj}$, respectively. Additionally, $\mu(z)$ represents the membership function, which quantifies the degree of truth of the z element. Each α-cut somehow represents a confidence interval for the considered input or output value.

Given that crisp inputs and outputs can be represented as degenerated membership functions with only one value in their domain, we can assume that all inputs and outputs are fuzzy. In this way, the dual of the output-oriented radial model, i.e., model (3), can be formulated as follows:

$$\begin{array}{lll} \widetilde{\phi}_0^* = & Min & \sum\limits_{i=1}^{m} v_{i0}\widetilde{x}_{i0} - \pi_0 \\ & s.t. & \\ & & \begin{array}{lll} \sum\limits_{r=1}^{s} u_{r0}\widetilde{y}_{r0} = 1, & & (4.1) \\ \sum\limits_{i=1}^{m} v_{i0}\widetilde{x}_{ij} - \sum\limits_{r=1}^{s} u_{r0}\widetilde{y}_{rj} - \pi_0 \geq 0, & j = 1,\ldots,n & (4.2) \\ v_{i0} \geq 0, & i = 1,\ldots,m & (4.3) \\ u_{r0} \geq 0, & r = 1,\ldots,s & (4.4) \end{array} \end{array}, \tag{4}$$

Additionally, the efficiency score $\widetilde{\phi}_0^*$ is also a fuzzy number due to the nature of the data used in the model. That fuzzy number is linked to a membership function. Kao and Liu's model allows this membership function to be calculated from different α-cuts. Next, given a certain level α ($0 < \alpha \leq 1$), we show how the lowest and the highest values of the corresponding α-cut for the membership function of $\widetilde{\phi}_0^*$ can be calculated. In particular, the lowest value is determined through program (5):

$$
\begin{array}{lll}
\left(\widetilde{\phi}_0^*\right)_\alpha^L = & Min & \sum\limits_{i=1}^{m} v_{i0}\left(\widetilde{x}_{i0}\right)_\alpha^U - \pi_0 \\
& s.t. & \\
& & \begin{array}{llr}
\sum\limits_{r=1}^{s} u_{r0}\left(\widetilde{y}_{r0}\right)_\alpha^L = 1, & & (5.1) \\
\sum\limits_{i=1}^{m} v_{i0}\left(\widetilde{x}_{ij}\right)_\alpha^L - \sum\limits_{r=1}^{s} u_{r0}\left(\widetilde{y}_{rj}\right)_\alpha^U - \pi_0 \geq 0, & j \neq 0 & (5.2) \\
\sum\limits_{i=1}^{m} v_{i0}\left(\widetilde{x}_{i0}\right)_\alpha^U - \sum\limits_{r=1}^{s} u_{r0}\left(\widetilde{y}_{r0}\right)_\alpha^L - \pi_0 \geq 0, & & (5.3) \\
v_{i0} \geq 0, & i = 1, \ldots, m & (5.4) \\
u_{r0} \geq 0, & r = 1, \ldots, s & (5.5)
\end{array}
\end{array} . \quad (5)
$$

In model (5), note that the inputs of the evaluated DMU and the outputs of all other units were set to their greatest values, while the outputs of the assessed unit and the inputs of all other DMUs were set to their lowest values. This is the essence of the approach by Kao and Liu [17]. Regarding the highest value of the corresponding α-cut, it can be implemented through model (6):

$$
\begin{array}{lll}
\left(\widetilde{\phi}_0^*\right)_\alpha^U = & Min & \sum\limits_{i=1}^{m} v_{i0}\left(\widetilde{x}_{i0}\right)_\alpha^L - \pi_0 \\
& s.t. & \\
& & \begin{array}{llr}
\sum\limits_{r=1}^{s} u_{r0}\left(\widetilde{y}_{r0}\right)_\alpha^U = 1, & & (6.1) \\
\sum\limits_{i=1}^{m} v_{i0}\left(\widetilde{x}_{ij}\right)_\alpha^U - \sum\limits_{r=1}^{s} u_{r0}\left(\widetilde{y}_{rj}\right)_\alpha^L - \pi_0 \geq 0, & j \neq 0 & (6.2) \\
\sum\limits_{i=1}^{m} v_{i0}\left(\widetilde{x}_{i0}\right)_\alpha^L - \sum\limits_{r=1}^{s} u_{r0}\left(\widetilde{y}_{r0}\right)_\alpha^U - \pi_0 \geq 0, & & (6.3) \\
v_{i0} \geq 0, & i = 1, \ldots, m & (6.4) \\
u_{r0} \geq 0, & r = 1, \ldots, s & (6.5)
\end{array}
\end{array} . \quad (6)
$$

Models (5) and (6) allow the interval $\left[\left(\widetilde{\phi}_0^*\right)_\alpha^L, \left(\widetilde{\phi}_0^*\right)_\alpha^U\right]$ to be determined for different values of α $(0 < \alpha \leq 1)$.

Finally, ranking units to ascertain better performance is a procedure of interest in efficiency analysis. In the case of the Fuzzy Data Envelopment Analysis, there are a few approaches that could be applied. In this paper, again following [1], we implemented the following index:

$$
I_0 = \left[\sum_{k=0}^{h}\left(\left(\widetilde{\phi}_0^*\right)_{\alpha_k}^U - c\right)\right] \Big/ \left[\sum_{k=0}^{h}\left(\left(\widetilde{\phi}_0^*\right)_{\alpha_k}^U - c\right) - \sum_{k=0}^{h}\left(\left(\widetilde{\phi}_0^*\right)_{\alpha_k}^L - d\right)\right], \quad (7)
$$

where $c = \min\limits_{j,k}\left\{\left(\widetilde{\phi}_j^*\right)_{\alpha_k}^L\right\}$ and $d = \max\limits_{j,k}\left\{\left(\widetilde{\phi}_j^*\right)_{\alpha_k}^U\right\}$.

4. Data and Variables

We used data from PISA 2015 in our empirical study. PISA is a triennial study that provides international comparative data on the performance of 15-year-old students in three main domains (in each cycle, one of the domains is in focus, with reading in 2000 and 2009, mathematics in 2003 and 2012, and science in 2006 and 2015) (reading, science and mathematics). In addition, this database contains information about many potential factors that might affect those results, such as variables representing student family background, home resources, school environment or class characteristics (in total, there are more than two hundred variables, including the original variables and the composite indexes constructed from the original information). This information is derived from the responses given by students and school principals to different questionnaires [11].

We selected students as the unit of analysis because they represent the group for which we have imprecise measures of educational output approximated by plausible values. Moreover, there are some additional reasons. First, if we consider that the main purpose pursued in education is the improvement of results, it makes sense that students represent the evaluated units, since they are the ones who must develop their skills based on the resources they have at their disposal [47]. Second, if the observations were aggregated at the school level, the fact that the level of resource utilization by students may differ according to their characteristics cannot be taken into account [48]. Additionally, using school-level averages does not consider the existing dispersion of the data, which may result in inaccurate measures of performance, especially regarding the identification of efficient units [49] For the purpose of our study, we used a representative dataset for Spain, composed of 6700 students from 201 schools (the original dataset includes 6378 students, but we excluded 38 students due to the absence of data in some variables).

With regard to the selection of input variables, we followed a very restrictive efficiency notion that consists of using only one input representing student socioeconomic background as a proxy for the academic quality of the student (this approach is also used in other previous studies [41,50]). Specifically, the input at student level was measured by the students' socioeconomic background (ESCS): an index of the economic, social and cultural status of students created by PISA analysts that captures a range of aspects of a student's family and home background and combines information on parents' education and occupations and cultural possessions at home. The first variable is the higher educational level of any of the students' parents according to the International Standard Classification of Education (ISCED). The second variable is the highest occupational status of any of the students' parents according to the International Socio-economic Index of Occupational Status (ISEI [51]). The third variable is an index of educational possessions related to household economy. This indicator is continuous and is positively correlated with output variables. Using this single input, we evaluated the extent to which a student is making the most of his/her potential abilities, considering his/her socioeconomic background as a proxy for this concept, or to which his/her performance is below the expected level.

Since the ESCS index presented negative values (the values of the PISA index of economic, social and cultural status were standardized to a mean of zero for the total population of students in OECD countries), the values of this variable were rescaled by adding the maximum negative value to all of them; thus, all the new quantities were positive. As a result, the variable fulfils the requirement of isotonicity (i.e., ceteris paribus, more input implies equal or higher level of output), which allows us to preserve the desirable property of translation invariance [52]. This variable was treated as crisp in our empirical analysis.

Output variables are represented by students´ test scores in the three domains assessed by PISA (mathematics, reading and science). For each domain, PISA 2015 provides ten plausible values; thus, we can define our output by using different alternatives that we intend to compare. Thus, we can use a single crisp value for each domain, which in turn could consist of the use of only one plausible value (e.g., PV1MATH, PV1READ or PV1SCIE) or the mean of all the available plausible values ($\overline{\text{PVMATH}}$, $\overline{\text{PVREAD}}$ or $\overline{\text{PVSCIE}}$). The other alternative is to treat them as fuzzy numbers. This possibility implies considering the existing variation among different plausible values, which can be higher or lower for each of the observations, as shown in Table 1, for two randomly selected students.

Table 1. Statistics of fuzzy variables for two randomly selected students.

Student	Domains	Mean	SD	Variation Coefficient
1635	MATHS	497.14	41.98	0.08
	READING	511.04	45.79	0.09
	SCIENCE	496.84	23.64	0.05
5102	MATHS	479.41	13.48	0.03
	READING	504.24	10.68	0.02
	SCIENCE	526.04	17.65	0.03

As a preliminary step to the application of the approach suggested by Kao and Liu [17], we need to model the values of the three variables representing the output (PVMATH, PVREAD and PVSCIE) as a particular fuzzy number. To this end, we estimated a kernel function for each student from the data corresponding to each variable. We also calculated the skewness coefficient for each kernel, obtaining values close to zero, which is indicative that the kernel functions are quite symmetrical. This implies that the mean values of the distributions are close to the median and the mode. Figure 1 presents several examples of the shape of the estimated kernel functions for the PVMATH variable for different students.

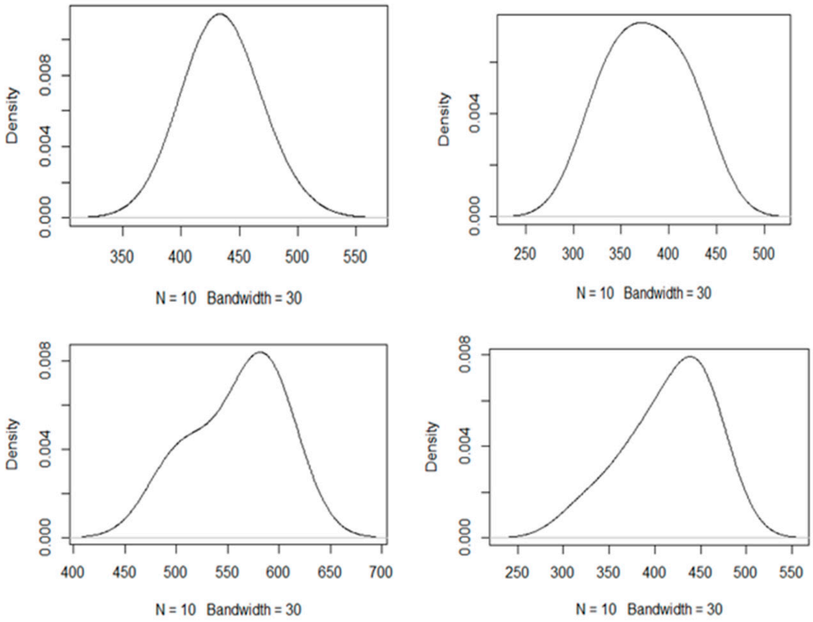

Figure 1. Examples of kernel functions for the PVMATH variable.

Kao and Liu's approach determines the α-cut for each value α, with $0 < \alpha \leq 1$. The α-cut corresponds to an interval. To this end, the y-axis in the kernel must be rescaled so that the maximum is equal to one for the rescaled kernel function. For example, if we consider student #1635 from the data sample, we obtain different intervals (α-cuts) for the values $\alpha = 0.7$, $\alpha = 0.8$, $\alpha = 0.9$ and $\alpha = 1$. The corresponding α-cuts are shown in Table 2. Given the set of α-cuts, we may determine the membership function of the fuzzy efficiency score of each DMU (student), following the steps described in Section 3. Table 3 reports the descriptive statistics for all the variables employed in our study.

Table 2. α-cuts for student 1635 (randomly selected).

	MATH		READ		SCIE	
α	MATH^L	MATH^U	READ^L	READ^U	SCIE^L	SCIE^U
0.7	445.1619	533.7739	459.1364	547.8178	474.0604	514.9951
0.8	453.0232	520.3045	468.1629	537.2065	478.1630	510.2898
0.9	462.4251	507.2320	478.7103	525.5529	482.8929	504.7852
1	483.6888	483.6888	501.9167	501.9167	493.4934	493.4934

Table 3. Descriptive statistics.

	Variable		Mean	SD	Min	Max
Input	ESCS (Crisp)		4.06	1.18	0.15	7.59
Outputs	MATH (Fuzzy)	PV1MATH	490.56	82.97	182.21	763.90
		PV2MATH	490.30	82.91	188.22	743.36
		PV3MATH	491.45	83.53	202.83	822.88
		PV4MATH	490.00	83.04	120.56	793.86
		PV5MATH	490.05	82.97	192.22	800.69
		PV6MATH	489.15	82.57	145.55	766.85
		PV7MATH	491.03	85.34	181.32	796.75
		PV8MATH	490.78	83.51	188.53	755.75
		PV9MATH	492.13	83.50	189.36	770.91
		PV10MATH	491.21	83.86	163.28	797.26
	READ (Fuzzy)	PV1READ	499.63	85.55	161.77	779.97
		PV2READ	498.72	85.83	190.47	757.95
		PV3READ	500.65	86.00	174.56	789.86
		PV4READ	498.52	85.83	162.16	746.98
		PV5READ	500.42	86.98	158.96	758.82
		PV6READ	500.86	86.70	164.17	734.15
		PV7READ	499.96	86.75	118.88	752.60
		PV8READ	501.32	85.36	192.69	767.53
		PV9READ	499.77	84.13	175.96	755.31
		PV10READ	499.89	86.66	163.25	767.98
	SCIE (Fuzzy)	PV1SCIE	497.14	86.47	210.70	754.33
		PV2SCIE	497.53	86.81	190.18	763.32
		PV3SCIE	497.60	85.94	186.66	805.02
		PV4SCIE	497.23	87.48	147.04	789.23
		PV5SCIE	497.27	87.13	191.37	760.99
		PV6SCIE	497.50	86.80	187.20	745.63
		PV7SCIE	497.07	86.78	194.79	752.90
		PV8SCIE	497.70	87.01	222.69	763.39
		PV9SCIE	497.37	86.53	214.96	755.82
		PV10SCIE	496.99	86.73	195.06	758.77

5. Results

In this section, we illustrate the potential divergences that might arise in efficiency measures depending on how we deal with plausible values by applying different approaches to the data from PISA. We chose three alternative methods (the resolution of these approaches was carried out by programming algorithms with R software [53]). First, we estimated a standard DEA using ESCS as our input and the mean values of all the plausible values for each domain as our outputs, i.e., treating PVMATH, PVREAD and PVSCIE as crisp (Model A). Second, we also treated PVMATH, PVREAD and PVSCIE as crisp, but we calculated a standard DEA for each set of plausible values (ten values), once more including ESCS as the only input (Model B) (these estimations were conducted using the R package “lpSolveAPI” [54]). Both estimations are relatively simple, since we only need to solve linear programming models. Finally, we obtained fuzzy efficiency estimates using the approach by Kao and Liu [17] for different α-cuts (α = 0.7, 0.8, 0.9 and 1) incorporating

PVMATH, PVREAD and PVSCIE as fuzzy numbers through kernel functions and ESCS as a crisp input variable (Model C) (we also estimated efficiency scores for other values (e.g., α = 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6), but we do not report these results because the intervals are too wide). This is much more complex, since it is necessary to previously estimate the kernel functions of each of the fuzzy variables (PVMATH, PVREAD and PVSCIE) (we obtained these estimates using the "density" function, implemented within the R stats package [53], opting for the use of the "Gaussian" smoothing method). Once the kernels were estimated, we performed the α-cuts on the kernels, generating, for each fuzzy variable and for each α, an interval of values. Next, we created a database for each α-cut with the interval data of all the fuzzy variables and incorporated the information of the rest of the crisp variables. Finally, after obtaining these databases, we applied the fuzzy linear programming model (5) and (6) of Kao and Liu [17] (again, this was also solved using the R package "lpSolveAPI" [54]), which allowed us to obtain, for each student, the values of the efficiency associated with each α.

Table 4 reports the main descriptive statistics of efficiency estimates for each alternative. For Model 3, we present for each α both the lowest and the highest values of the confidence interval for the fuzzy efficiency score. Therefore, we obtained intervals as $\left[\left(\widetilde{\varphi}_0^*\right)_\alpha^L, \left(\widetilde{\varphi}_0^*\right)_\alpha^U\right]$. Moreover, we also calculated the index I_0 (Equation (7)), which reflects a summary of the different results determined for the set of αs. We assume variable returns to scale and an output orientation in all estimations.

Table 4. Descriptive statistics of efficiency scores for different approaches.

			MIN	**Q1**	**Median**	**Mean**	**Q3**	**MAX**
Model A	**Score**		**1**	**1.22**	**1.34**	**1.38**	**1.49**	**2.78**
Model B	**PV1**		1	1.27	1.39	1.43	1.55	3.37
	PV2		1	1.26	1.38	1.42	1.54	3.16
	PV3		1	1.29	1.41	1.45	1.57	2.92
	PV4		1	1.27	1.39	1.44	1.55	4.33
	PV5		1	1.27	1.38	1.42	1.55	2.78
	PV6		1	1.25	1.37	1.42	1.54	3.26
	PV7		1	1.24	1.36	1.40	1.52	3.39
	PV8		1	1.26	1.38	1.42	1.54	2.84
	PV9		1	1.26	1.37	1.41	1.53	2.68
	PV10		1	1.27	1.39	1.43	1.55	2.98
Model C FDEA (different α-cuts)	**0.7**	**E^L**	1	1.10	1.20	1.23	1.33	2.18
		E^U	1	1.36	1.49	1.55	1.68	3.44
	0.8	**E^L**	1	1.13	1.23	1.26	1.36	2.27
		E^U	1	1.33	1.46	1.51	1.64	3.27
	0.9	**E^L**	1	1.16	1.27	1.30	1.40	2.40
		E^U	1	1.31	1.43	1.48	1.60	3.08
	1	**E^L**	1	1.24	1.35	1.40	1.51	2.72
		E^U	1	1.24	1.35	1.40	1.51	2.72
		Ij *	1	1.07	1.10	1.11	1.13	1.65

(*) All p-values are approximately 0.000 with a level of 0.1%.

From our results, we observe that, in general terms, the average scores calculated with the traditional DEA are similar to the estimated values obtained with fuzzy DEA for $\alpha = 1$. Indeed, the correlation among them is higher than 0.9 (and statistically significant) in all cases, as we can see in Table 5. In our opinion, this is due to two main reasons. First, considering $\alpha = 1$, it yields the mode of the data for each variable and student as the corresponding α-cut. Second, the modes are generally close to the means of the data in this real example due to the symmetry of the data distributions. Consequently, the results of programs (5) and (6) for $\alpha = 1$ are identical and generate an optimal value (score) very similar to the optimal value of model (2), where PVMATH, PVREAD and PVSCIE are incorporated as crisp variables. Of course, in the case of asymmetric data distributions, the efficiency scores for the traditional model and for $\alpha = 1$ could be different.

Table 5. Correlation coefficients among efficiency scores estimated with different approaches.

		Model A	**Model B**										**Model C**	
			PV1	**PV2**	**PV3**	**PV4**	**PV5**	**PV6**	**PV7**	**PV8**	**PV9**	**PV10**	**FDEA (α = 1)**	**Ij ***
Model A		1.000												
Model B	**PV1**	0.931	1.000											
	PV2	0.924	0.854	1.000										
	PV3	0.925	0.856	0.839	1.000									
	PV4	0.931	0.858	0.859	0.848	1.000								
	PV5	0.932	0.861	0.853	0.858	0.859	1.000							
	PV6	0.933	0.860	0.858	0.850	0.862	0.857	1.000						
	PV7	0.938	0.865	0.857	0.856	0.869	0.869	0.869	1.000					
	PV8	0.925	0.852	0.859	0.835	0.854	0.852	0.857	0.863	1.000				
	PV9	0.934	0.865	0.856	0.856	0.865	0.862	0.865	0.871	0.860	1.000			
	PV10	0.933	0.867	0.851	0.862	0.856	0.855	0.860	0.865	0.856	0.864	1.000		
Model C	**FDEA (α = 1)**	0.991	0.926	0.918	0.920	0.925	0.926	0.923	0.931	0.918	0.929	0.926	1.000	
	Ij *	0.983	0.921	0.917	0.912	0.922	0.916	0.923	0.927	0.911	0.924	0.919	0.979	1.000

(*) All p-values are approximately 0.000 with a level of 0.1%.

Although we note that the values of the index I_0 are highly correlated with the rest of scores calculated using traditional DEA, it is also noteworthy that their mean values are significantly lower. This can be observed more clearly if we examine the divergences for students presenting a high dispersion in output data. In Table 6, we report the estimated scores for all the considered alternatives for students with the highest values of standard deviation in the fuzzy variables. Here, we can see that the standard DEA tends to overestimate the level of inefficiency, since their efficiency scores are clearly higher than the value of the index I_0. The information displayed in this table also highlights that there is a high level of variation across efficiency scores calculated with each set of plausible values, that cannot be detected when plausible values are aggregated (Model A).

If we focus on students identified as efficient according to standard DEA models, which are presented in Table 7, we detect that most of them present an index I_0 higher than one and, for almost a half, are not considered as efficient in the fuzzy DEA model for α = 1. Moreover, some students are fuzzy efficient exclusively for α = 1, but the upper bound of the α-cuts for all other possible levels is greater than one. Thus, we cannot be totally sure that these students are performing efficiently, although this would be the conclusion to be drawn from an analysis performed with a standard DEA. Therefore, our results suggest that the consideration of all the variability reported by the different plausible values using an FDEA model may modify the conclusions obtained to a certain extent, at least as far as the identification of efficient units is concerned.

Table 6. Efficiency scores for students with the highest dispersion in their plausible values.

	SD			Model A	Model B										Model C (FDEA)	
Student	**MATH**	**READ**	**SCIE**	**Score**	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**	**PV6**	**PV7**	**PV8**	**PV9**	**PV10**	$\alpha = 1$	**Ij ***
185	38.37	77.36	46.35	2.20	1.92	2.40	2.24	2.18	1.91	2.93	1.69	2.42	2.10	2.60	2.24	1.36
557	32.81	58.84	31.50	1.97	1.72	2.16	1.99	1.82	1.81	2.34	2.10	1.88	1.93	1.81	1.96	1.26
583	46.70	43.38	55.29	2.00	1.75	2.23	1.94	2.38	2.05	2.64	2.08	2.02	2.37	1.95	2.05	1.32
906	51.18	32.09	41.62	2.11	2.24	1.91	1.95	2.44	2.14	2.48	2.47	2.14	2.11	2.44	2.22	1.36
958	46.23	57.57	37.43	2.28	2.76	1.91	2.19	2.13	1.95	2.41	2.41	2.64	2.44	2.13	2.38	1.45
1219	40.72	45.78	32.16	1.92	2.04	2.06	1.80	2.26	1.63	2.27	2.07	2.06	2.27	1.93	2.08	1.27
1226	60.37	48.15	53.14	2.78	3.37	2.47	2.92	4.33	2.78	3.02	3.39	2.23	2.37	2.54	2.72	1.65
2252	35.56	56.73	46.84	1.53	1.58	1.47	1.70	1.62	1.41	1.63	1.60	1.73	1.80	1.70	1.60	1.16
2900	36.67	60.53	33.98	1.63	1.60	1.90	1.53	1.72	1.79	1.64	1.79	1.76	1.88	1.54	1.81	1.20
2925	42.93	65.98	33.59	2.05	2.05	2.46	1.63	2.48	1.75	1.93	1.96	2.00	1.92	1.96	1.99	1.30
3258	55.95	37.06	40.64	2.12	2.71	2.02	2.05	1.83	1.69	1.95	2.58	1.89	2.15	2.36	2.04	1.38
3316	47.05	63.21	31.78	1.86	1.76	2.08	1.67	1.74	1.68	2.17	1.77	2.23	1.92	1.90	1.78	1.27
3381	48.36	55.56	47.17	1.94	1.68	2.61	1.83	1.78	1.92	2.00	2.32	1.99	2.06	2.13	1.93	1.30
3542	56.59	34.18	30.17	2.25	2.53	2.24	2.32	1.97	2.73	2.41	2.36	2.20	2.11	2.19	2.34	1.37
4010	53.58	25.39	31.14	2.02	1.62	2.38	2.07	1.90	2.05	2.37	1.86	1.94	2.20	1.82	2.04	1.29
4170	24.73	54.92	43.73	2.26	2.40	2.46	2.25	2.42	2.45	2.07	2.44	2.57	2.44	1.76	2.41	1.41
4351	33.06	44.46	41.12	2.04	2.15	2.18	2.27	2.34	2.28	2.13	1.94	1.75	2.49	2.10	2.10	1.30
4694	64.17	44.90	23.01	1.99	2.04	2.24	1.87	2.30	1.67	2.12	1.61	2.30	1.89	2.11	2.06	1.28
4888	22.12	61.87	41.77	1.96	1.92	1.92	2.18	2.01	1.86	1.88	2.08	1.91	2.24	2.17	1.98	1.25
5212	49.70	29.40	38.51	1.77	1.65	2.12	1.72	1.97	1.77	2.00	1.95	1.76	1.90	1.65	1.80	1.21
5963	34.25	63.22	34.64	2.26	2.52	1.98	2.48	2.37	2.19	2.50	2.09	1.99	1.96	2.06	2.12	1.35
6486	62.03	57.37	53.83	1.72	1.76	3.16	1.79	1.62	1.64	1.85	1.87	1.85	1.47	2.03	1.73	1.27
6489	29.86	50.01	33.10	1.77	2.04	1.70	1.86	1.99	2.11	1.90	1.64	1.97	1.74	1.83	1.89	1.22

(*) All *p*-values are approximately 0.000 with a level of 0.1%.

Table 7. Efficient units according to the traditional DEA compared to efficiency scores with other alternatives (FDEA and each PV).

	Mod. A	Mod. B										Mod. C—FDEA (Different α-Cuts)								
												0.7		0.8		0.9		1		
Student	**Score**	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**	**PV6**	**PV7**	**PV8**	**PV9**	**PV10**	E^L	E^U	E^L	E^U	E^L	E^U	E^L	E^U	**Ij ***
502	1.00	1.00	1.00	1.03	1.00	1.01	1.02	1.00	1.00	1.00	1.00	1	1.05	1	1.03	1	1.01	1.00	1.00	1.01
613	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	1	1.00	1	1.00	1.00	1.00	1.00
952	1.00	1.08	1.04	1.09	1.05	1.00	1.05	1.00	1.03	1.09	1.07	1	1.12	1	1.10	1	1.08	1.04	1.04	1.03
1098	1.00	1.00	1.03	1.00	1.00	1.07	1.01	1.04	1.06	1.04	1.01	1	1.11	1	1.07	1	1.05	1.02	1.02	1.02
1795	1.00	1.02	1.08	1.00	1.00	1.00	1.00	1.09	1.11	1.00	1.01	1	1.11	1	1.08	1	1.05	1.00	1.00	1.03
2062	1.00	1.00	1.11	1.01	1.10	1.00	1.02	1.04	1.02	1.14	1.09	1	1.12	1	1.10	1	1.08	1.01	1.01	1.03
2853	1.00	1.00	1.03	1.00	1.00	1.00	1.18	1.00	1.19	1.10	1.07	1	1.10	1	1.07	1	1.03	1.00	1.00	1.04
2863	1.00	1.00	1.05	1.08	1.06	1.02	1.00	1.00	1.00	1.04	1.01	1	1.04	1	1.02	1	1.00	1.00	1.00	1.01
2907	1.00	1.02	1.12	1.10	1.15	1.06	1.02	1.00	1.13	1.00	1.00	1	1.16	1	1.13	1	1.10	1.02	1.02	1.03
3312	1.00	1.00	1.00	1.00	1.02	1.00	1.00	1.00	1.04	1.00	1.00	1	1.06	1	1.05	1	1.03	1.00	1.00	1.02
4274	1.00	1.02	1.02	1.06	1.02	1.13	1.03	1.04	1.00	1.00	1.13	1	1.08	1	1.06	1	1.03	1.00	1.00	1.03
5874	1.00	1.01	1.03	1.11	1.00	1.00	1.09	1.00	1.05	1.04	1.10	1	1.12	1	1.09	1	1.05	1.00	1.00	1.03
6126	1.00	1.08	1.00	1.00	1.01	1.01	1.04	1.03	1.05	1.00	1.10	1	1.11	1	1.09	1	1.08	1.02	1.02	1.02
6467	1.00	1.00	1.03	1.00	1.13	1.03	1.01	1.00	1.07	1.00	1.00	1	1.06	1	1.05	1	1.01	1.00	1.00	1.02
6654	1.00	1.00	1.00	1.00	1.00	1.00	1.05	1.00	1.03	1.00	1.00	1	1.00	1	1.00	1	1.00	1.00	1.00	1.01

(*) All *p*-values are approximately 0.000 with a level of 0.1%.

6. Conclusions

In this paper, we used a novel method to incorporate all the information provided by the plausible values available in international large-scale educational databases as an approximation of educational output to estimate efficiency measures. Specifically, we applied the fuzzy DEA approach proposed by Kao and Liu [17]. This methodology provides more precise information than conventional methods, such as DEA, which only allows for incorporating a part of the available information about students as proxies of educational output. Therefore, the main objective of this research was to determine the extent to which having this additional information on the units evaluated can change the results in terms of identifying best practices and ranking units. In order to be able to make this comparison, we applied both methods to assess the performance of a sample of Spanish students participating in PISA 2015.

Our results reveal that the estimated measures of performance obtained with the fuzzy DEA approach present high levels of correlation with the efficiency scores calculated using traditional DEA models. Therefore, in principle, when researchers use only one plausible value or aggregate values into a single aggregate measure, they obtain similar results to those obtained if they were to consider the whole distribution of results, represented by all the available plausible values that are expressed by a kernel function within our framework. This can be considered as a positive result for practitioners using traditional frontier methods in this area, since it appears that not considering all available information on students' abilities does not seem to generate relevant biases in the estimated efficiency measures.

However, we also found some noteworthy divergences among the estimated scores with both alternatives. First, we noticed that standard DEA may overestimate the level of inefficiency for some students, especially those with a higher dispersion in their results, i.e., plausible values more different from each other. Second, we also observed that a high proportion of units identified as efficient in standard DEA models are not identified as efficient in the fuzzy DEA model; thus, we cannot be completely sure that units identified as being efficient by traditional DEA are actually efficient.

In view of the above, we claim that empirical studies using microdata from international comparative surveys for estimating measures of performance using DEA should try to account for the existing variation among plausible values as a representation of the output (test scores). Otherwise, there might be an overestimation of the level of inefficiency of some units as well as a misidentification of efficient units, which could also affect the measures of the remaining evaluated units, since they are commonly used as references in DEA.

Finally, we would like to mention that we are aware that our study presents a series of limitations that should lead us to interpret the results with some caution. Probably the most important of these limitations is that we only incorporated one input variable in our empirical analysis, which is not very common in efficiency studies. However, this decision was made in an attempt to minimize the potential problems of loss of discriminatory power of nonparametric techniques, such as DEA or FDEA. Another potential limitation arises from the fact that in our application, we only used data referring to one country (Spain), so it would be advisable to extend the scope of the study to a broader context considering information on other countries. In the same way, it could also be interesting to use more recent information, such as the data available in PISA 2018, although in principle, this change would not be relevant since the newest wave of this survey offers the same number of plausible values for each student (10). Finally, it is also worth mentioning that we only used one of the multiple existing models to implement FDEA, Kao and Liu´s procedure; thus, a potential extension of the present study could be to applied other alternative FDEA approaches to test the robustness of our results.

Author Contributions: Conceptualization, J.A. and J.M.C.; Methodology, J.A.; Data curation, L.O.; writing—original draft preparation, J.A., J.M.C. and L.O.; writing—review and editing, J.A., J.M.C. and L.O. All authors have read and agreed to the published version of the manuscript.

Funding: J. Aparicio and L. Ortiz thank the financial support of the Spanish Ministry of Science and Innovation and the State Research Agency under grant PID2019-105952GB-I00/AEI/10.13039/501100011033, and Jose M. Cordero also acknowledges the support from the Spanish Ministry of Science and Innovation and the State Research Agency through grant ECO2017-83759-P/AEI/10.13039/501100011033 and the Regional government of Extremadura (Junta de Extremadura) through grant GR18106.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this paper are publicly available in the OECD website: https://www.oecd.org/pisa/data/2015database/, accessed on 4 July 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reynolds, D.; Teddlie, C. *The International Handbook of School Effectiveness Research*; Routledge: London, UK, 2002.
2. Creemers, B.; Kyriakides, L. *The Dynamics of Educational Effectiveness*; Routledge: London, UK, 2008.
3. Ergüzen, A.; Erdal, E.; Ünver, M.; Özcan, A. Improving Technological Infrastructure of Distance Education through Trustworthy Platform-Independent Virtual Software Application Pools. *Appl. Sci.* **2021**, *11*, 1214. [CrossRef]
4. Dospinescu, O.; Dospinescu, N. Perception Over E-Learning Tools in Higher Education: Comparative Study Romania and Moldova. In Proceedings of the IE 2020 International Conference, Madrid, Spain, 20–23 July 2020; Volume 10.
5. Gustafsson, J.-E. Effects of International Comparative Studies on Educational Quality on the Quality of Educational Research. *Eur. Educ. Res. J.* **2008**, *7*, 1–17. [CrossRef]
6. Fischman, G.E.; Topper, A.M.; Silova, I.; Goebel, J.; Holloway, J.L. Examining the influence of in-ternational large-scale assessments on national education policies. *J. Educ. Policy* **2019**, *34*, 470–499. [CrossRef]
7. Rutkowski, L.; Gonzalez, E.; Joncas, M.; von Davier, M. International large-scale assessment data: Issues in secondary analysis and reporting. *Educ. Res.* **2010**, *39*, 142–151. [CrossRef]
8. Sjøberg, S. PISA and 'real life challenges': Mission impossible? In *PISA According to PISA: Does PISA Keep What It Promises*; LIT: Wien, Vienna, 2007; pp. 241–263.
9. Sáez-López, J.-M.; Domínguez-Garrido, M.-C.; Medina-Domínguez, M.-D.-C.; Monroy, F.; González-Fernández, R. The Competences from the Perception and Practice of University Students. *Soc. Sci.* **2021**, *10*, 34. [CrossRef]
10. Mislevy, R.J.; Beaton, A.E.; Kaplan, B.; Sheehan, K.M. Estimating Population Characteristics From Sparse Matrix Samples of Item Responses. *J. Educ. Meas.* **1992**, *29*, 133–161. [CrossRef]
11. OECD. *PISA 2015 Technical Report*; PISA, OECD Publishing: Paris, France, 2016.
12. Boston College: TIMSS & PIRLS International Study Center. Methods and Procedures in TIMSS 2015. 2016. Available online: http://timssandpirls.bc.edu/publications/timss/2015-methods.html (accessed on 21 May 2021).
13. Foy. *TIMSS 2015 User Guide for the International Database. TIMSS & PIRLS*; International Study Center, International Association for the Evaluation of Educational Achievement: Boston, MA, USA, 2017.
14. De Witte, K.; López-Torres, L. Efficiency in education: A review of literature and a way forward. *J. Oper. Res. Soc.* **2017**, *68*, 339–363. [CrossRef]
15. Zadeh, L. The concept of a linguistic variable and its application to approximate reasoning—I. *Inf. Sci.* **1975**, *8*, 199–249. [CrossRef]
16. Hatami-Marbini, A.; Emrouznejad, A.; Tavana, M. A taxonomy and review of the fuzzy data en-velopment analysis literature: Two decades in the making. *Eur. J. Oper. Res.* **2011**, *214*, 457–472. [CrossRef]
17. Kao, C.; Liu, S.-T. Fuzzy efficiency measures in data envelopment analysis. *Fuzzy Sets Syst.* **2000**, *113*, 427–437. [CrossRef]
18. Berezner, A.; Adams, R.J. Why large-scale assessments use scaling and item response theory. In *Implementation of Large-Scale Education Assessments*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2017.
19. von Davier, M.; Gonzalez, E.; Mislevy, R. Plausible values: What are they and why do we need them? *IERI Monogr. Ser. Issues Methodol. Large-Scale Assess.* **2009**, *2*, 9–36.
20. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: Hoboken, NJ, USA, 1987.
21. Schafer, J.L. Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Stat. Neerl.* **2003**, *57*, 19–35. [CrossRef]
22. Von Davier, M.; Sinharay, S. Analytics in international large-scale assessments: Item response the-ory and population models. In *Handbook of International Large-Scale Assessment: Background, Technical Issues, And Methods of Data Analysis*; Chapman and Hall/CRC: London, UK, 2013; pp. 155–174.
23. Marsman, M.; Maris, G.K.J.; Bechger, T.M.; Glas, C.A.W. What can we learn from plausible val-ues? *Psychometrika* **2016**, *81*, 274–289. [CrossRef] [PubMed]
24. Laukaityte, I.; Wiberg, M. Using plausible values in secondary analysis in large-scale assessments. *Commun. Stat. Theory Methods* **2016**, *46*, 11341–11357. [CrossRef]
25. Wu, M. The role of plausible values in large-scale surveys. *Stud. Educ. Eval.* **2005**, *31*, 114–128. [CrossRef]
26. Mislevy, R.J. Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika* **1993**, *58*, 79–85. [CrossRef]

27. Luo, Y.; Dimitrov, D.M. A Short Note on Obtaining Point Estimates of the IRT Ability Parameter With MCMC Estimation in Mplus: How Many Plausible Values Are Needed? *Educ. Psychol. Meas.* **2019**, *79*, 272–287. [CrossRef]
28. OECD. *PISA 2018 Technical Report*; PISA, OECD Publishing: Paris, France, 2019.
29. Bibby, Y. Plausible Values: How Many for Plausible Results? Ph.D. Thesis, University of Melbourne, Melbourne, Australia, 2020.
30. Goldstein, H. International comparisons of student attainment: Some issues arising from the PISA study. *Assess. Educ. Princ. Policy Pract.* **2004**, *11*, 319–330. [CrossRef]
31. Braun, H.; von Davier, M. The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assess. Educ.* **2017**, *5*, 1–16. [CrossRef]
32. Macdonald, K. *PV: Stata Module to Perform Estimation with Plausible Values*; Statistical Software Compo-nents S456951; College Department of Economics: Boston, MA, USA, 2019.
33. Avvisati, F.; Keslair, F. *REPEST: Stata Module to Run Estimations with Weighted Replicate Samples and Plausible Values*; Statistical Software Components; College Department of Economics: Boston, MA, USA, 2020.
34. OECD. *PISA Data Analysis Manual, SPSS Second Edition*; OECD Publishing: Paris, France, 2009.
35. Cordero, J.M.; Polo, C.; Santín, D.; Simancas, R. Efficiency measurement and cross-country differ-ences among schools: A robust conditional nonparametric analysis. *Econ. Model.* **2018**, *74*, 45–60. [CrossRef]
36. De Witte, K.; Kortelainen, M. What explains the performance of students in a heterogeneous envi-ronment? Conditional efficiency estimation with continuous and discrete environmental variables. *Appl. Econ.* **2013**, *45*, 2401–2412. [CrossRef]
37. Agasisti, T.; Zoido, P. Comparing the Efficiency of Schools Through International Benchmarking: Results From an Empirical Analysis of OECD PISA 2012 Data. *Educ. Res.* **2018**, *47*, 352–362. [CrossRef]
38. Agasisti, T.; Zoido, P. The efficiency of schools in developing countries, analysed through PISA 2012 data. *Socio-Econ. Plan. Sci.* **2019**, *68*, 100711. [CrossRef]
39. Aparicio, J.; Cordero, J.M.; Pastor, J.T. The determination of the least distance to the strongly effi-cient frontier in Data Envelopment Analysis oriented models: Modelling and computational aspects. *Omega* **2017**, *71*, 1–10. [CrossRef]
40. Aparicio, J.; Cordero, J.M.; Gonzalez, M.; Lopez-Espin, J.J. Using non-radial DEA to assess school efficiency in a cross-country perspective: An empirical analysis of OECD countries. *Omega* **2018**, *79*, 9–20. [CrossRef]
41. Cordero, J.M.; Prior, D.; Simancas, R. A comparison of public and private schools in Spain using robust nonparametric frontier methods. *Central Eur. J. Oper. Res.* **2015**, *24*, 659–680. [CrossRef]
42. Cordero, J.M.; Santín, D.; Simancas, R. Assessing European primary school performance through a conditional nonparametric model. *J. Oper. Res. Soc.* **2017**, *68*, 364–376. [CrossRef]
43. De Jorge, J.; Santin, D. Determinantes de la eficiencia educativa en la Unión Europea. *Hacienda Pública Española* **2010**, *193*, 131–155.
44. Crespo-Cebada, E.; Pedraja-Chaparro, F.; Santín, D. Does school ownership matter? An unbiased efficiency comparison for regions of Spain. *J. Prod. Anal.* **2014**, *41*, 153–172. [CrossRef]
45. Banker, R.D.; Charnes, A.; Cooper, W.W. Some Models for Estimating Technical and Scale Ineffi-ciencies in Data Envelopment Analysis. *Manag. Sci.* **1984**, *30*, 1078–1092. [CrossRef]
46. Emrouznejad, A.; Tavana, M.; Hatami-Marbini, A. The State of the Art in Fuzzy Data Envelopment Analysis. In *Performance Measurement with Fuzzy Data Envelopment Analysis*; Springer: Berlin/Heidelberg, Germany; pp. 1–48.
47. Waldo, S. On the use of student data in efficiency analysis: Technical efficiency in Swedish upper sec-ondary school. *Econ. Educ. Rev.* **2007**, *26*, 173–185. [CrossRef]
48. Santin, D. La medición de la eficiencia de las escuelas: Una revisión crítica. *Hacienda Pública Española* **2006**, *177*, 57–82.
49. Aparicio, J.; Cordero, J.M.; Ortiz, L. Measuring efficiency in education: The influence of imprecision and variability in data on DEA estimates. *Socio-Econ. Plan. Sci.* **2019**, *68*, 100698. [CrossRef]
50. Thieme, C.; Prior, D.; Tortosa-Ausina, E. A multilevel decomposition of school performance using robust nonparametric frontier techniques. *Econ. Educ. Rev.* **2013**, *32*, 104–121. [CrossRef]
51. Ganzeboom, H.; De Graaf, P.M.; Treiman, D.J. A standard international socio-economic index of occupational status. *Soc. Sci. Res.* **1992**, *21*, 1–56. [CrossRef]
52. Cooper, W.W.; Seiford, L.M.; Tone, K. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*; Springer: Berlin/Heidelberg, Germany, 2007.
53. R Core Team. Package "Stats.". RA Lang. Environment Stat. Comput. Vienna, Austria: R Foundation for Statistical Computing. 2021. Available online: https://www.R-project.org (accessed on 21 October 2020).
54. Konis, K.; Konis, M.K. Package 'lpSolveAPI'. 2020. Available online: https://cran.r-project.org/web/packages/lpSolveAPI/lpSolveAPI.pdf (accessed on 21 October 2020).