APPLIED ECONOMICS
https://doi.org/10.1080/00036846.2021.2006136

Routledge
Taylor & Francis Group

Plausible values and their use in efficiency analyses with educational data

Juan Aparicio[a], Jose M. Cordero[b] and Lidia Ortiz[a]

[a]Center of Operations Research (CIO). University Miguel Hernandez of Elche (UMH), Alicante, Spain; [b]Department of Economics, University of Extremadura (UEX), Badajoz, Spain

ABSTRACT

There is extensive literature focused on the evaluation of efficiency in the education sector, both at the micro level, analyzing the performance of students or schools, and at the macro level, exploring the behavior of regions or countries. This type of studieshas been driven by exploiting data available in international large-scale assessments, where output measures are usually represented by the so-called plausible values, understood as a representation of the range of the abilities of each student. In this study, we analyze the different options available to incorporate these plausible values in applied studies focused on measuring efficiency and how the results obtained can be affected according to the selected criterion. To do this, we assess the efficiency of Spanish schools participating in PISA using the two most common methodologies in this field: data envelopment analysis and stochastic frontier analysis and considering three different proxies for the educational output: (i) a single plausible value; (ii) an aggregate measure calculated from the ten plausible values available; (iii) an average of the estimates made with the ten plausible values separately. The main conclusion derived from our results is that there are hardly any differences between in the estimates made with different strategies.

KEYWORDS

Efficiency; educational evaluation; plausible values; PISA; data envelopment analysis; stochastic frontiers

I. Introduction

The information provided by international large-scale assessments has become a very useful tool for analyzing the performance of education systems around the world and even, in some countries, serves as a guide for making educational policy decisions (Wiseman 2010). Therefore, it is not surprising that in recent years there has been an exponential growth in the number of studies, projects, and specialized articles in which the data provided by these databases are used, the main exponents of which are PISA (*Programme for International Student Assessment*), TIMSS (*Trends in Mathematics and Science Study*) and PIRLS (*Progress in International Reading Literacy Study*).[1]

In most of these works, regressions are used with a structure based on the so-called educational production function (Hanushek 1979; Todd and Wolpin 2003), in order to identify and quantify the link between different factors related to students´ background and school environment and the outcomes of the educational process, frequently represented by the results obtained in a standardized test of knowledge in various subjects, among which reading comprehension, mathematics or science are the most common.[2] However, there are other alternative approaches in which researchers or educational policy managers and other educational stakeholders may be interested, especially in a context characterized by a generalized increase in costs in the education sector (Eurostat 2021), such as knowing if there is an optimal use of available educational resources (inputs) or if, on the contrary, inefficient practices are being carried out by the agents involved (students or teachers) or as a consequence of certain institutional aspects. In these cases, it will be necessary to estimate efficiency measures of performance relying on a production frontier approach, thus the best practices observed (students or schools) can be identified and serve as a reference or

CONTACT Jose M. Cordero jmcordero@unex.es Department of Economics, University of Extremadura (UEX), 06006, Badajoz, Spain

[1] See Cordero, Cristobal, and Santín (2018) for a review of recent works that use data from the three assessments or Hopfenbeck et al. (2018) for the specific case of PISA.

[2] Reading proficiency is assessed in PISA and PIRLS, while mathematics and science are assessed in TIMSS and in PISA. In addition, it should be noted that, in its latest waves, PISA has added other additional subjects in its assessments, such as financial literacy and problem-solving skills (since 2012) or the so-called global competence (since 2018).

benchmarking to evaluate the rest according to their consumption of inputs and the outputs achieved (Fried et al. 2008).

The present study lies within this second line of research, whose development in recent years has also been remarkable.[3] Specifically, we focus on an issue to which researchers have barely paid attention in their empirical work, even though it may have a relevant impact on the results obtained. We refer to how to deal with the representative information of the educational output, that is, the results obtained by the students in the standardized knowledge tests used in international large-scale assessments. As is well known, these results are not expressed through a single score, but rather each student evaluated receives several scores, commonly known in the specialized literature as plausible values, which are intended to reflect the range of skills of each student (Mislevy et al. 1992). In this way, PISA aims to collect the effect of certain external conditions that are beyond the control of the student at the moment of taking the test.

The technical reports and manuals of international databases provide a detailed description of the statistical procedures that needs to be followed when using regressions in empirical studies using with an education production function structure (see, for example, OCDE 2016; Martin, Mullis, and Hooper 2016), in which plausible values are usually the dependent variable. However, there are no such clear guidelines on how to proceed when the aim of the study is to estimate efficiency measures that reflect the level of performance of a group of students, schools or educational systems through the use of frontier techniques, such as data envelopment analysis (DEA) or stochastic frontier analysis (SFA). Therefore, it is not surprising to observe the existence of an evident disparity of criteria dealing with these plausible values in empirical studies with data from international evaluations applying some of these techniques, where those values are usually considered as the output of the production function. Thus, it is possible to find empirical studies using only one of the available values, others calculating the mean of all the available values and then using this average to estimate efficiency measures and a final group that chooses to estimate efficiency measures for each plausible value and, subsequently, calculates an average from the different estimates.

The objective of the present study is to empirically analyze whether the results of an efficiency analysis can be affected according to the option chosen to incorporate the information provided by the plausible values available in international large-scale assessments. To do this, we estimate efficiency measures for a set of schools using the information contained in the PISA 2015 database, in which there is a total of 10 plausible values for each of the main domains or subjects assessed (reading, mathematics and science). In order to test the robustness of our results, these measurements will be estimated using the two methodological approaches most commonly used in the literature, i.e. the nonparametric DEA model and the parametric SFA model, comparing the results obtained for the three alternatives discussed above. In this way, we intend to verify whether the choice between these options affects the results obtained in the studies that seek to measure the level of efficiency of a set of units operating in the education sector.

The rest of the paper is organized as follows. The following section provides a detailed explanation of the concept of plausible values and describes some basic guidelines for using them correctly in empirical studies. Next, in the third section, a brief description of the two methods that will be used to estimate the efficiency measures is provided. The fourth section describes the main characteristics of the database used in our application and the variables selected to conduct the proposed empirical analysis. In the fifth section, the main results obtained in the estimations are presented and discussed. Finally, the article ends with the conclusions section.

II. Plausible values and their use in applied studies

Measuring educational outcomes is one of the main concerns of large-scale assessments. In order to adequately assess the knowledge or skills

[3] See De Witte and López-Torres (2017) for a review of this literature.

of the participants, it would be necessary that they answer a large volume of questions or items with different degrees of difficulty for each evaluated domain (e.g. mathematics, reading or sciences). Nevertheless, since there is a limited time to conduct the test (it does not usually exceed two hours),[4] they can only respond to a limited number of items with different levels of difficulty. Therefore, it is necessary to somehow predict what the results would have been if all questions (instead of only a fraction of the total test item pool) had been answered, assuming that the estimation process will be subject to a substantial amount of measurement error (von Davier, Gonzalez, and Mislevy 2009).

In most large-scale assessments this prediction is made using a psychometric technique called item response theory (Rasch 1960–1980; Carlson and von Davier 2013), which allows for estimating proficiency levels according to the answers provided to the limited fraction of administered cognitive items, their level of difficulty and some contextual variables. This technique makes it possible to obtain a measure of performance for each individual on a common continuous scale, that is, a distribution of values is generated for each domain with its associated probabilities (usually assumed to be a normal distribution). From the distribution of each individual, several random values are usually extracted, the so-called plausible values, which allow us to approximate the range of skills or competencies in each domain assessed (Wu 2005). This distribution uses as a reference the results of the students on an international scale, normally assigning the mean a value of 500 points, with a standard deviation of 100. In turn, these values are usually divided into different levels of difficulty to facilitate their interpretation, so that students who are at a certain level are assumed to be able to answer the items that are at that level and below successfully; being less likely that they are able to solve higher level questions.

Plausible values 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 educational variables in secondary analyses. 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 (Mislevy 1993). Selecting multiple values is necessary to correctly estimate the error variance resulting from the imputation procedure carried out regarding the answers to the questions answered by each individual.

The usual practice of most international assessments has been to provide five plausible values for each student in each of the domains tested since this number is sufficient for accurate estimation of population-level statistics (Luo and Dimitrov 2019). However, in recent years, some international 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), while only five plausible values were used for previous PISA waves from 2000 to 2012.

As mentioned above, the procedure that must be carried out to undertake any estimation with regression techniques is widely known by users of international surveys since it is clearly described in their technical reports and user manuals. It is based upon the original work of Rubin (1987) and can be divided into the following four steps: (i) estimate the statistic/model of interest using each of the plausible values to obtain five (or ten) separate parameter estimates and the corresponding sampling error; (ii) calculate the mean of these parameter estimates; (iii) estimate the magnitude of the imputation error and (iv) calculate the value of the final standard error by combining the mean sampling error and the imputation error. 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, most of the specialized data processing programs have specific routines or commands to make estimates with plausible values, such as PV (Macdonald 2019) or REPEST (Avvisati and Keslair 2020) in Stata.

[4] 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.

In studies focused on obtaining representative efficiency measures of the behavior of students, schools or educational systems, however, the way to proceed is not so clear, thus it is possible to come across different methodological options in the literature. The most common alternative consists of the arbitrary use of a single plausible value of all those available (e.g. Mancebón et al. 2012; De Witte and Kortelainen, 2013; Aparicio, Cordero, and Pastor 2017; Agasisti and Zoido 2018, 2019), which in practice means forgoing much of the information available about the skills demonstrated by the students, in addition to disregarding the uncertainty associated with the measurement of the distribution of their competencies. Another possibility is to calculate the mean of all available plausible values as if there were only a single estimate of students' abilities, and then use these averages to estimate efficiency measures (this is the strategy employed by, for example, Agasisti 2013; Santín and Sicilia 2015; Cordero, Santín, and Simancas 2017; Ben Yahia, Essid, and Rebai 2018). Finally, there is more cumbersome process, which involves estimating an efficiency measure for each of the plausible values and, later, calculating an average value from these efficiency measures (this option was chosen, among others, by De Jorge and Santin 2010; Crespo-Cebada, Pedraja, and Santín 2014). The existence of this disparity of criteria when handling plausible values in applied studies on efficiency measurement leads us to wonder whether the results obtained may be affected (or not) by the decision adopted by the researchers in those empirical studies conducted in the educational setting.

III. Methodology

This section gives an overview of the two frontier methods that have been most widely applied in the previous literature for estimating production frontiers in the education sector (see Johnes 2015; De Witte and López-Torres 2017 and reference therein): DEA and SFA. Consequently, these are the two approaches we have used to estimate efficiency measures in our empirical analysis.

Data envelopment analysis has traditionally been the preferred option for researchers interested in obtaining efficiency measures of units that operate in the education sector, mainly schools (Thanassoulis et al. 2016). The preference for this nonparametric approach is justified mainly by its great flexibility, since it does not require assuming a specific functional form for the production process, being sufficient to assume a set of minimum properties that the set of units evaluated must satisfy (convexity, free availability and minimum extrapolation). This approach also makes it possible to easily incorporate the multidimensional nature of the educational process considering different output indicators, such as the results in the different subjects evaluated in a standardized test (reading comprehension, mathematics and science in the case of PISA). Its main limitations are found in its deterministic nature, which makes it difficult to apply when measurement errors are detected or extreme points exist, and in the fact that it does not have statistical properties (except consistency when the sample size is large), thus making it difficult to test hypotheses.

The DEA model solves a mathematical programming problem for each observed unit, assigning an efficiency score that reflects the level of use it makes of the factors available to it. Specifically, the formulation of the model to evaluate the performance of a unit taking as a reference its consumption of inputs (x_0) and the level of outputs (y_0) reached, takes the following form[5]

$$\begin{aligned} 1/E_0 = {} & Max\,\phi_0 \\ & s.t. \\ & \sum_{j=1}^{n} \lambda_{j0} x_{ij} \le x_{i0}, \qquad i = 1,\dots,m \\ & \sum_{j=1}^{n} \lambda_{j0} y_{rj} \ge \phi_0 y_{r0}, \qquad r = 1,\dots,s \\ & \sum_{j=1}^{n} \lambda_{j0} = 1, \\ & \lambda_{j0} \ge 0, \qquad j = 1,\dots,n, \end{aligned} \tag{1}$$

[5] This model corresponds to the version that assumes variable returns to scale in production (Banker, Charnes, and Cooper 1984) and output orientation, that is, one is interested in maximizing the results obtained from the available inputs.:

In this case, if $E_0 = 1$, the evaluated unit is considered efficient, since it is not possible to achieve an improvement in its outputs without increasing some of its inputs. On the contrary, if $E_0 < 1$, the unit will be inefficient and has room to increase its output levels while consuming the same amount of resources. In these cases, the value $1/E_0$ precisely reflects the percentage increase in the output that can be achieved.

Although not as popular in this sector, parametric techniques offer another attractive methodological option for measuring efficiency in the field of education as evidenced by the fact that there is a notable number of studies that have relied on this approach to obtain measures of school efficiency (e.g. Barbetta and Turati 2003; Deutsch, Dumas, and Silber 2013). This approach is based on a priori specification of the functional form of the production frontier with constant parameters (β), such as Cobb-Douglas or translog, which must be estimated based on the available observations. The main exponent is the SFA model, introduced in the literature by Aigner, Lovell, and Schmidt (1977) and Meeusen and van Den Broeck (1977), in which it is considered that the evaluated units may be affected by various random factors (v), such as measurement errors derived from the non-consideration of relevant variables in the model regardless of the inefficiency (u) of the evaluated unit. The production function to be estimated will be the following:

$$Y = f(x;\beta) + \varepsilon \;\; \varepsilon = v - u \qquad (2)$$

The main criticisms that these models have received are because the results obtained are strongly conditioned by the assumptions made regarding the specified functional form and the error distribution, especially when only cross-sectional data are available. Another limitation that has traditionally been associated with this approach is the difficulty posed by its use in a multi-output context, as happens in our case with the different subjects evaluated in PISA. However, this problem can be solved by replacing the production frontier by a distance function.[6.]

$$\log y_s = -\log D_O\left(x_0, \frac{y_0}{y_s}\right) + v - u \qquad (3)$$

To estimate this expression, it is necessary to assume a specific functional form. In our case, we opted for the Cobb-Douglas function, so that, after carrying out the necessary transformations to take into account that more than one output is being considered, we have the following expression:

$$\log\left(D\left(x, \frac{y}{y_s}\right)\right) = \beta_0 + \sum_{r=1}^{s-1} \beta_r \log\frac{y_r}{y_s} + \sum_{i=1}^{m} \beta_i \log x_i. \qquad (4)$$

IV. Data and variables

In order to analyze whether the procedure used to incorporate plausible values from a large-scale international database affects the results of an evaluation of the efficiency levels of the units, the OECD PISA 2015 database has been used since it provides a greater number of representative scores of the students' competences to be able to approximate the output of the educational process. Specifically, it provides us with a total of 10 plausible values for each of the participants (15-year-old students) in the three evaluated subjects: reading comprehension, mathematics, and science. In addition, there is also a wide set of variables representative of the school and family environment of the students, obtained from two questionnaires, one completed by the students themselves and the other by school principals.

In our empirical analysis we have selected schools as units of analysis, since these usually constitute the main focus of interest in most of the applied studies on the measurement of efficiency carried out in the educational context (De Witte and López-Torres 2017). The study is limited to the specific context of the Spanish educational system, so we have a total of 201 units (schools) evaluated.[7]

[6] For a detailed explanation of this procedure, see Bogetoft and Otto (2011).:

[7] In fact, 980 Spanish schools participated in the 2015 edition, since all the Autonomous Regions decided to take part with a larger sample so that their results were internationally comparable. However, in our study we have preferred to use only the subsample of 201 schools that represents the country as a whole.

As output variables, the average results of the students belonging to each school have been selected, represented through all the available plausible values, so that we have a total of 30 variables (10 values for each domain evaluated). However, in our estimates of efficiency measures, only three of them will be used (PVMATH, PVREAD and PVSCIE) to be able to carry out the comparison between the different alternative options considered (use of a single plausible value for each domain or the average of all the available plausible values for each domain).

The selection of inputs is a far more complex task, since the database offers an extensive battery of indicators representing educational resources involved in the educational process. As there is no explicit rule to discriminate between them, our decision was based on selecting three variables that are positively (and significantly) correlated with the outputs and that, in addition, count on the theoretical basis of having been previously used in different applied studies focused on measuring school efficiency.[8] The first of these variables is an indicator that represents the economic, social, and cultural status of the students belonging to each school (ESCS),[9] since we consider that students constitute the raw material on which schools are nurtured. The second is an indicator that represents the human capital of schools, approximated through the ratio of teachers per hundred students (TSRATIO). The third indicator tries to approximate the quality of the school's infrastructures or physical resources. Specifically, we used the inverted educational materials shortage index (SCHRES) constructed from principals' opinions about the status of various items (textbooks, library, laboratory supplies, and information technology equipment).[10]

Table 1 shows the descriptive statistics of all the variables used in the study. As can be seen, the mean values of the different plausible values are very similar to each other, although the existence of a certain variability between them is also appreciated, reflected in the magnitude of the deviation and the extreme values, which largely justifies the interest of the proposed study.

V. Results

This section presents the results obtained when estimating efficiency measures for the 201 schools that make up the selected sample considering different alternatives for the management of plausible values as outputs of the production process. Specifically, efficiency measures have been estimated for each school using the three inputs and each of the 10 plausible values (PV1, PV2, … , PV10) representing the three subjects evaluated in

Table 1. Descriptive statistics.

Variable			Mean	Std. Dev.	Min	Max
Inputs	ESCS		4.37	0.68	2.97	5.93
	TSRATIO		9.70	8.93	2.45	100.00
	SCHRES		3.49	1.20	0.04	4.97
Outputs	MATH	PV1MATH	488.45	34.34	382.94	574.85
		PV2MATH	488.13	34.87	376.02	568.58
		PV3MATH	489.26	35.41	398.37	566.29
		PV4MATH	487.90	33.96	368.62	561.28
		PV5MATH	487.91	34.58	387.29	552.90
		PV6MATH	487.19	34.70	385.16	561.85
		PV7MATH	488.97	35.53	382.98	576.83
		PV8MATH	488.37	35.96	383.02	571.59
		PV9MATH	490.32	35.17	379.34	568.11
		PV10MATH	488.79	34.96	388.20	569.31
	READ	PV1READ	497.97	35.70	378.73	584.91
		PV2READ	497.26	35.65	377.04	590.21
		PV3READ	498.92	36.86	385.86	597.86
		PV4READ	497.00	36.64	368.22	596.36
		PV5READ	498.97	37.17	367.87	601.52
		PV6READ	498.89	36.04	369.97	588.16
		PV7READ	498.42	37.72	379.21	596.76
		PV8READ	499.72	36.98	366.59	593.79
		PV9READ	498.30	35.47	392.77	591.55
		PV10READ	498.36	35.22	381.20	586.97
	SCIENCE	PV1SCIE	495.15	35.79	367.45	583.04
		PV2SCIE	495.98	35.23	378.55	580.40
		PV3SCIE	495.74	34.62	370.07	573.50
		PV4SCIE	495.46	35.53	375.95	577.28
		PV5SCIE	495.54	35.66	362.05	581.79
		PV6SCIE	495.61	34.83	376.21	575.34
		PV7SCIE	495.18	35.10	374.22	571.85
		PV8SCIE	495.76	35.00	387.06	575.04
		PV9SCIE	495.59	34.54	390.15	582.26
		PV10SCIE	495.19	34.53	368.85	580.21

Source: Authors' own elaboration.

[8] The same indicators (or very similar ones) are also used as inputs in school efficiency assessments conducted by Thieme, Prior, and Tortosa-Ausina (2013), Agasisti (2014), Aparicio, Cordero, and Pastor (2017) or Agasisti and Zoido (Agasisti and Zoido 2018, 2019).

[9] The ESCS indicator is an index created by PISA analysts based on the information provided by students about the educational level and occupation of their parents and the educational resources and cultural possessions available at home.

[10] The values of the ESCS and SCHRES variables had to be rescaled (adding the minimum value of each variable) to ensure that all units presented positive values.

PISA as outputs. Next, efficiency scores have been estimated again considering the three inputs and the mean of the ten plausible values ($\overline{PV}$) of the three subjects as outputs. Finally, we consider a third option that consist of calculating the average of the efficiency indices obtained with the ten combinations of plausible values.

To facilitate the interpretation of the results obtained with the different methodological options considered are presented in two distinct subheadings.[11] First, we provide the results obtained when applying the DEA model in Subsection 5.1 and, subsequently, those derived from the application of the SFA method adapted to a multi-output context are displayed in Subsection 5.2[12]

Estimation of school efficiencies using DEA

Table 2 shows that the main descriptive statistics of the efficiency scores estimated with DEA, which are quite high in general for all the alternatives considered. The means range from 0.938 (estimate with plausible value 1) to 0.943 (estimate with plausible values 4, 5, and 10). Likewise, efficiency scores derived from averaged values are also very similar. Moreover, we can observe that the average efficiency levels are quite homogeneous and do not suffer from excessively marked deviations (standard deviation ranging from 0.041 to 0.043).

If we look at the number of schools identified as efficient, there are no relevant differences between the estimates derived with the alternative options considered, with a range that oscillates between 23 and 27 (11–13% of units) and an average value of 24. In order to enhance this comparison, we rely on the content of Table 3, which shows the 24 units identified as efficient when using the average of all plausible values ($\overline{PV}$) and the estimated values for those same units with each combination of plausible values. This information allows us to verify that 14 schools (25, 58, 77, 81, 94, 141, 142, 153, 161, 169, 176, 183, 184, 199) are also considered efficient with the ten plausible values available and, consequently, they also obtain a unitary index for the mean of all of them. In the other ten schools, we can also notice that for most of the plausible values a unitary index is also assigned and, for those that are not, the value is very close to 1 (all the inefficiencies detected are less than 3%). Therefore, it can be concluded that the units identified as efficient are practically the same regardless of which option is chosen to incorporate the plausible values into the analysis conducted using DEA.

Another indicator that can be used to check whether there are relevant divergences between the different estimates is the Pearson´s correlation coefficient. In Table 4 we display the values of those coefficients, which are all greater than 0.9 (the values range between 0.90 and 0.97). Therefore, this is additional evidence that there are very few differences between the values of the efficiency scores obtained with the different sets of plausible values and with the average values of all the available observations for each student.

Finally, as an additional robustness check to determine whether the different estimates obtained can be considered similar to each other, we rely on the nonparametric test proposed by Li (1996), later adapted to the specific DEA context by Simar and

Table 2. Descriptive statistics of the efficiency levels estimated by DEA.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{PV}$	Average of 10 PVs
Average Efficiency	0.938	0.940	0.942	0.943	0.943	0.942	0.941	0.941	0.943	0.943	0.942	0.942
Std. Dev.	0.042	0.042	0.042	0.041	0.041	0.043	0.043	0.041	0.041	0.041	0.041	0.042
Min	0.801	0.770	0.798	0.806	0.790	0.796	0.794	0.788	0.794	0.786	0.793	0.792
Q1	0.911	0.911	0.918	0.915	0.915	0.914	0.913	0.916	0.918	0.918	0.915	0.915
Mean	0.941	0.942	0.944	0.943	0.945	0.943	0.942	0.940	0.942	0.947	0.943	0.943
Q3	0.974	0.978	0.973	0.974	0.978	0.978	0.978	0.976	0.976	0.973	0.975	0.976
Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
% Efficient schools	25 (12%)	24 (12%)	27 (13%)	27 (13%)	23 (11%)	24 (12%)	26 (13%)	24 (12%)	26 (13%)	24 (12%)	24 (12%)	

Source: Authors' own elaboration.

[11] The estimates of the efficiency levels have been obtained through the *Benchmarking* package (Bogetoft and Otto 2015) developed for *R* software (R Core Team 2020) and *RStudio* (R Studio Team 2020).

[12] This procedure requires normalizing all the outputs by means of one of them. In our case, we have usedscience as a reference, so that the adjusted parametric model would be the following:x_0

Table 3. Comparison between efficient schools with different alternatives (DEA).

SCHOOL	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{PV}$	Average of 10 PVs
15	0.997	0.996	0.990	1.000	0.971	1.000	1.000	1.000	1.000	1.000	1.000	0.995
25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
52	0.985	1.000	1.000	1.000	0.996	0.995	1.000	1.000	1.000	1.000	1.000	0.998
55	1.000	1.000	1.000	1.000	1.000	0.984	1.000	1.000	1.000	1.000	1.000	0.998
58	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
69	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
76	1.000	0.988	0.994	1.000	0.999	1.000	0.982	0.997	1.000	0.986	1.000	0.995
77	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
81	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
94	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
118	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000
135	1.000	1.000	1.000	1.000	1.000	0.994	0.983	1.000	1.000	0.997	1.000	0.997
141	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
142	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
145	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
153	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
161	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
169	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
176	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
183	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
184	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
193	0.983	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	1.000	1.000	0.998
194	0.989	0.979	0.992	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996
199	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Source: Authors' own elaboration.

Table 4. Correlation coefficients of the efficiency levels estimated with DEA.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{PV}$
PV1	1.0000										
PV2	0.9067	1.0000									
PV3	0.9016	0.9189	1.0000								
PV4	0.9178	0.9193	0.9084	1.0000							
PV5	0.9051	0.9137	0.9166	0.9010	1.0000						
PV6	0.9068	0.9211	0.9265	0.9132	0.9223	1.0000					
PV7	0.9123	0.9135	0.9408	0.9180	0.9150	0.9206	1.0000				
PV8	0.9228	0.9199	0.9207	0.9275	0.9044	0.9298	0.9349	1.0000			
PV9	0.9184	0.9215	0.9182	0.9172	0.9063	0.9363	0.9270	0.9237	1.0000		
PV10	0.9243	0.9101	0.9264	0.9163	0.9045	0.9131	0.9307	0.9220	0.9062	1.0000	
$\overline{PV}$	0.9537	0.9613	0.9589	0.9593	0.9483	0.9613	0.9630	0.9658	0.9582	0.9603	1.0000

Source: Authors' own elaboration.

Zelenyuk (2006), which allows us to check whether the distributions of the efficiency scores obtained from different data can be considered to be analogous. The values shown in Table 5 confirm this phenomenon, since the associated probabilities of this test are not significant in any case, so we can conclude that there is a high level of homogeneity in the efficiency levels estimated with the eleven alternative DEA models considered.

Estimation of school efficiencies using SFA

Table 6 shows the descriptive statistics of efficiency scores derived from the estimates made with 11 different SFA models, one for each combination of plausible values and another using the mean of all the value ($\overline{PV}$). Once again, no large divergences were observed between different estimates, with an even smaller fluctuation than those registered with DEA, with values ranging between 0.956 and 0.963. Neither are relevant discrepancies detected in the variability of the estimates since the standard deviations of the scores obtained are very low and similar to each other (variation from 0.018 to 0.026).

These results are equivalent to those shown above for DEA in the sense that there do not seem to be significant divergences between using one of the available plausible values or the average value of all of them.

As we checked with the estimates made with DEA, to be able to further investigate the possible divergences between these estimates, we calculated Pearson´s correlation coefficients between different estimates obtained with SFA. Additionally, we performed the test of equality

Table 5. Test of equality of distributions between indices estimated with DEA.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{PV}$
PV1	-										
PV2	0.115 (0.884)	-									
PV3	0.421 (0.572)	0.598 (0.433)	-								
PV4	0.259 (0.762)	−0.2 (0.799)	−0.147 (0.852)	-							
PV5	−0.080 (0.932)	−0.262 (0.752)	0.205 (0.801)	−0.269 (0.743)	-						
PV6	0.121 (0.872)	0.071 (0.929)	−0.014 (0.989)	−0.256 (0.75)	−0.173 (0.827)	-					
PV7	0.274 (0.733)	0.085 (0.922)	0.109 (0.883)	−0.08 (0.921)	−0.066 (0.933)	−0.291 (0.719)	-				
PV8	0.159 (0.837)	0.78 (0.254)	0.281 (0.731)	0.439 (0.555)	0.345 (0.658)	0.193 (0.814)	0.656 (0.327)	-			
PV9	0.437 (0.574)	0.437 (0.585)	−0.039 (0.953)	−0.015 (0.99)	0.146 (0.858)	−0.247 (0.752)	−0.088 (0.919)	−0.096 (0.893)	-		
PV10	0.065 (0.944)	0.216 (0.794)	−0.11 (0.902)	−0.223 (0.777)	−0.339 (0.664)	−0.247 (0.746)	0.014 (0.983)	0.035 (0.968)	−0.141 (0.871)	-	
$\overline{PV}$	−0.035 (0.966)	0.142 (0.86)	0.049 (0.958)	0.026 (0.975)	−0.17 (0.83)	−0.133 (0.859)	−0.013 (0.982)	0.111 (0.888)	−0.046 (0.951)	−0.199 (0.801)	-

Statistic:(p-value). Source: Authors' own elaboration.

Table 6. Descriptive statistics of the efficiency levels estimated by SFA.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{PV}$	Average of 10 PVs
Average Efficiency	0.957	0.958	0.963	0.961	0.957	0.956	0.956	0.957	0.960	0.956	0.961	0.958
Std. Dev.	0.025	0.023	0.018	0.021	0.024	0.026	0.024	0.025	0.021	0.026	0.021	0.023
Min	0.857	0.851	0.881	0.871	0.854	0.846	0.856	0.857	0.873	0.838	0.867	0.861
Q1	0.945	0.949	0.955	0.953	0.946	0.946	0.944	0.947	0.952	0.944	0.952	0.949
Mean	0.961	0.962	0.966	0.966	0.962	0.959	0.962	0.962	0.964	0.962	0.964	0.963
Q3	0.975	0.976	0.976	0.975	0.976	0.976	0.975	0.975	0.976	0.976	0.976	0.976
Max	0.990	0.988	0.988	0.990	0.989	0.990	0.989	0.989	0.990	0.989	0.990	0.989

Source: Authors' own elaboration.

of distributions proposed by Li, Maasoumi, and Racine (2009). As was the case for DEA, the high values reported in Table 7 (all values are higher than 0.89) suggest that there is a high correlation between all pairs of values, while the values of the Li´s test displayed in Table 8 corroborate that there are not significant divergences among those estimates.[13]

Thus, the results obtained by applying the SFA approach to the available data support the findings previously presented for the most widely used technique in the educational field, i.e. DEA, thus when researchers use only one plausible value or aggregate values for estimating efficiency measures in this context, obtain practically the same results. In other words, the decision between choosing a

Table 7. Correlation coefficients of the efficiency levels estimated with SFA.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{PV}$
PV1	1.000										
PV2	0.938	1.000									
PV3	0.926	0.943	1.000								
PV4	0.908	0.926	0.900	1.000							
PV5	0.925	0.932	0.938	0.899	1.000						
PV6	0.924	0.942	0.950	0.914	0.945	1.000					
PV7	0.922	0.943	0.944	0.912	0.942	0.944	1.000				
PV8	0.928	0.941	0.937	0.916	0.928	0.934	0.949	1.000			
PV9	0.938	0.941	0.930	0.904	0.936	0.946	0.947	0.936	1.000		
PV10	0.936	0.937	0.940	0.908	0.926	0.946	0.936	0.936	0.933	1.000	
$\overline{PV}$	0.966	0.969	0.959	0.952	0.962	0.966	0.963	0.967	0.962	0.966	1.000

Source: Authors' own elaboration.

[13] To calculate the statistic, the *npdeneqtest* function of the *np* library (Hayfield and Racine 2008) developed for the R software (R Core Team 2020) and *RStudio* (R Studio Team 2020) has been used.

Table 8. Test of equality of distributions between indices estimated with SFA.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	$\overline{\text{PV}}$
PV1	-										
PV2	−0.166 (0.846)	-									
PV3	1.626 (0.038)	−0.052 (0.21)	-								
PV4	1.099 (0.109)	−0.325 (0.364)	−0.955 (0.995)	-							
PV5	−1.282 (0.981)	−1.239 (0.633)	1.485 (0.033)	0.839 (0.075)	-						
PV6	−0.842 (0.779)	−1.308 (0.694)	2.960 (0.002)	1.986 (0.012)	−0.045 (0.418)	-					
PV7	−1.421 (0.927)	−2.048 (0.911)	0.568 (0.083)	−0.197 (0.22)	−1.153 (0.906)	−0.380 (0.461)	-				
PV8	−0.670 (0.878)	−1.376 (0.921)	0.678 (0.195)	−0.099 (0.458)	−0.451 (0.876)	−0.300 (0.765)	−0.520 (0.972)	-			
PV9	0.809 (0.331)	0.880 (0.9)	−0.299 (0.959)	−0.636 (0.987)	1.380 (0.192)	1.612 (0.168)	0.887 (0.56)	−0.183 (0.796)	-		
PV10	−2.222 (0.837)	−2.024 (0.513)	0.950 (0.013)	−0.191 (0.064)	−1.513 (0.553)	−1.778 (0.668)	−1.700 (0.828)	−1.977 (0.758)	−1.153 (0.162)	-	
$\overline{\text{PV}}$	−0.739 (0.658)	−1.890 (0.929)	−0.506 (0.49)	−0.774 (0.579)	−0.408 (0.594)	0.382 (0.209)	−0.239 (0.701)	−1.009 (0.787)	−1.691 (0.867)	1.732 (0.249)	-

Statistic:(p-value). Source: Authors' own elaboration.

plausible value or an average value of all available values seems to be irrelevant for the interpretation of the results obtained in an efficiency analysis, regardless of the technique used to estimate efficiencies.

This means that the conclusions reached in the different applied papers reviewed in Section II (e.g. De Witte, Maasen van den Broek, and Groot ; Crespo-Cebada, Pedraja, and Santín 2014; Santín and Sicilia 2015; Cordero, Santín, and Simancas 2017; Aparicio, Cordero, and Pastor 2017; Agasisti and Zoido 2018, 2019; Ben Yahia, Essid, and Rebai 2018) would not be affected by the choice adopted regarding the treatment of plausible values by their authors, since our findings suggest that the estimates of efficiency indices obtained with different approaches (using a single value, mean values or average of different values) are very similar to each other.

This is also a very useful finding for practitioners interested in conducting secondary analyses using frontier techniques, as DEA or SFA, using data from international large-scale assessments for different purposes such as simply measuring the efficiency of schools, regions or countries using cross-sectional data or panel data, exploring the factors that might explain inefficiency levels of those units using some of the many methodological options that exist in the literature (Bădin, Daraio, and Simar 2014) or decomposing the different sources of inefficiency adopting a metafrontier approach (Battese, Rao, and O'Donnell 2004; O'Donnell, Rao, and Battese 2008). For any of these purposes, the treatment of plausible values for the estimation of efficiency measures should not be a concern given that our results reveal that the estimated efficiency scores will not be affected by the decision they make on this issue.

VI. Concluding remarks

The present study has analyzed an issue that had not been studied so far in the growing literature on the measurement of efficiency in the educational context, namely, how to incorporate information on educational output in studies that use information from large-scale international assessments like PISA, TIMSS or PIRLS. Those datasets provide several measures representing student outcomes, the so-called plausible values, which may raise doubts for researchers about how to deal with them in their applied studies. The fact that these output measures represent a distribution of results and not an exact score is problematic for researchers working in this field, since the techniques commonly used to measure efficiency do not have statistical properties. Given this limitation, practitioners must decide between choosing one of the available plausible values, calculating the mean of all of them, or estimating a measure for each plausible value and then calculating an average of all the estimated measures.

With the purpose of offering a solution to this open question, we have conducted an empirical study using data about a sample of Spanish schools participating in PISA 2015. This survey provides ten plausible values as a representation of the abilities demonstrated by each participating student in three core domains (mathematics, reading comprehension and science), which we use as a proxy of educational outcomes in an educational production function that also includes several representative indicators of school inputs, such as students´ characteristics or educational (human and physical) resources.

In our application we have estimated efficiency scores using the two techniques most commonly used in the literature, data envelopment analysis (DEA) and stochastic frontiers (SFA) and considering the aforementioned alternative strategies (using a single plausible value, the mean value and the mean of estimations with different plausible values). Our findings reveal that the results hardly vary in the different considered scenarios, since there are quite high and significant correlations between the different estimated efficiency scores as well as no significant differences are observed between the distributions of the efficiency scores. Therefore, we can conclude that the decision made by the researcher on how to incorporate the information offered by the plausible values does not condition the results of the analysis and that any of the three options considered are perfectly valid. This conclusion may be considered as a positive result for practitioners and users of international databases like PISA, TIMSS, PIRLS or PIAAC who intend to estimate efficiency measures of performance using traditional frontier methods, since they can opt for any of the three considered alternatives without their choice having any effect on the results.

Acknowledgements

J. Aparicio and L. Ortiz thank the grant PID2019-105952GB-I00 funded by Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033. 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. He would also like to express his gratitude to the "European Regional Development Fund (FEDER). A way to make Europe" and to Junta de Extremadura through grant GR18106.

Disclosure statement

No potential conflict of interest was reportedby the authors.

Funding

The work was supported by the Government of Extremadura: GAEDAF Research Group [GR18106]; Spanish Ministry of Science and Innovation and the State Research Agency [ECO2017-83759-P/AEI/10.13039/501100011033,PID2019-105952GB-I00/ AEI / 10.13039/501100011033].

ORCID

Juan Aparicio http://orcid.org/0000-0002-0867-0004
Jose M. Cordero http://orcid.org/0000-0001-8783-6748

References

Agasisti, T. 2013. "The Efficiency of Italian Secondary Schools and the Potential Role of Competition: A Data Envelopment Analysis Using OECD-PISA2006 Data." *Education Economics* 21 (5): 520–544. doi:10.1080/09645292.2010.511840.

Agasisti, T. 2014. "The Efficiency of Public Spending on Education: An Empirical Comparison of EU Countries." *European Journal of Education* 49 (4): 543–557. doi:10.1111/ejed.12069.

Agasisti, T. and P. Zoido. 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. doi:10.3102/0013189X18777495.

Agasisti, T. and P. Zoido. 2019. "The Efficiency of Schools in Developing Countries, Analysed Through PISA 2012 Data." *Socio-Economic Planning Sciences* 68: 100711. doi:10.1016/j.seps.2019.05.002.

Aigner, D., C. A. K. Lovell, and P. Schmidt. 1977. "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics* 6 (1): 21–37. doi:10.1016/0304-4076(77)90052-5.

Aparicio, J., J. M. Cordero, and J. T. Pastor. 2017. "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. doi:10.1016/j.omega.2016.09.008.

Avvisati, F., and F. Keslair. 2020. REPEST: Stata Module to Run Estimations with Weighted Replicate Samples and Plausible Values. Boston, MA, USA: College Department of Economics.

Bădin, L., C. Daraio, and L. Simar. 2014. "Explaining Inefficiency in Nonparametric Production Models: The State of the Art." *Annals of Operations Research* 214 (1): 5–30. doi:10.1007/s10479-012-1173-7.

Banker, R. D., A. Charnes, and W. W. Cooper. 1984. "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." *Management Science* 30 (9): 1078–1092. doi:10.1287/mnsc.30.9.1078.

Barbetta, G. P. and G. Turati. 2003. "Efficiency of Junior High Schools and the Role of Proprietary Structure." AAnnals of Public and Cooperative Economics 74 (4): 529–552. doi:10.1111/j.1467-8292.2003.00234.x.

Battese, G. E., D. P. Rao, and C. J. O'-Donnell. 2004. "A Metafrontier Production Function for Estimation of Technical Efficiencies and Technology Gaps for Firms Operating Under Different Technologies." *Journal of Productivity Analysis* 21 (1): 91–103. doi:10.1023/B:PROD.0000012454.06094.29.

Ben Yahia, F., H. Essid, and S. Rebai. 2018. "Do Dropout and Environmental Factors Matter? A Directional Distance Function Assessment of Tunisian Education Efficiency." *International Journal of Educational Development* 60: 120–127. doi:10.1016/j.ijedudev.2017.11.004.

Bogetoft, P. and L. Otto. 2011. *Benchmarking with DEA, SFA, and R*. New York: Springer.

Bogetoft, P. and L. Otto. 2015. "Package 'Benchmarking'. Data Envelopment Analyses (DEA) and Stochastic Frontier Analyses (SFA) – Model Estimations and Efficiency Measuring." https://cran.r-project.org/web/packages/Benchmarking/Benchmarking.pdf .

Carlson, J. E. and M. von Davier. 2013. "Item Response Theory." *ETS Research Report Series* 2013 (2): 1–69. doi:10.1002/j.2333-8504.2013.tb02335.x.

Cordero, J. M., D. Santín, and R. Simancas. 2017. "Assessing European Primary School Performance Through a Conditional Nonparametric Model." *The Journal of the Operational Research Society* 68 (4): 364–376. doi:10.1057/jors.2015.42.

Cordero, J. M., V. Cristobal, and D. Santín. 2018. "Causal Inference on Education Policies: A Survey of Empirical Studies Using PISA, TIMSS and PIRLS." *Journal of Economic Surveys* 32 (3): 878–915. doi:10.1111/joes.12217.

Crespo-Cebada, E., F. Pedraja, and D. Santín. 2014. "Does School Ownership Matter? an Unbiased Efficiency Comparison for Regions of Spain." *Journal of Productivity Analysis* 41 (1): 153–172. doi:10.1007/s11123-013-0338-y.

De Jorge, J. and D. Santin. 2010. "Determinantes de la Eficiencia Educativa En la Unión Europea." *Hacienda Pública Española* 193: 131–155.

De Witte, K., and Kortelainen, M. (2013). What explains student performance in a heterogeneous environment? Estimating conditional efficiency with continuous and discrete environmental variables. *Applied Economics*, 45(17): 2401–2412.

De Witte, K. and L. López-Torres. 2017. "Efficiency in Education: A Review of Literature and a Way Forward." *The Journal of the Operational Research Society* 68 (4): 339–363. doi:10.1057/jors.2015.92.

Deutsch, J., A. Dumas, and J. Silber. 2013. "Estimating an Educational Production Function for Five Countries of Latin America on the Basis of the PISA Data." *Economics of Education Review* 36: 245–262. doi:10.1016/j.econedurev.2013.07.005.

Eurostat. 2021. *GDP per Capita, Consumption per Capita and Price Level Indices*. Eurostat Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=GDP_per_capita,_consumption_per_capita_and_price_level_indices .

Fried, H. O., C. A. K. Lovell, S. S. Schmidt. 2008. "Efficiency and Productivity." In *The Measurement of Productive Efficiency and Productivity Growth*, edited by H. O. Fried, C. A. K. Lovell, and S. S. Schmidt H. O. Fried; C. A. K. Lovell, and S. S. Schmidt, 3–91. New York: Oxford University Press.

Hanushek, E. A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *The Journal of Human Resources* 14 (3): 351–388. doi:10.2307/145575.

Hayfield, T. and J. S. Racine. 2008. "Nonparametric Econometrics: The Np Package." *Journal of Statistical Software* 27 (5): 1–32. doi:10.18637/jss.v027.i05.

Hopfenbeck, T. N., J. Lenkeit, Y. El Masri, K. Cantrell, J. Ryan, and J.-A. Baird. 2018. "Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment." *Scandinavian Journal of Educational Research* 62 (3): 333–353. doi:10.1080/00313831.2016.1258726.

Johnes, J. 2015. "Operational Research in Education." *European Journal of Operational Research* 243 (3): 683–696. doi:10.1016/j.ejor.2014.10.043.

Li, Q. 1996. "Nonparametric Testing of Closeness Between Two Unknown Distribution Functions." *Econometric Reviews* 15 (3): 261–274. doi:10.1080/07474939608800355.

Li, Q., E. Maasoumi, and J. S. Racine. 2009. "A Nonparametric Test for Equality of Distributions with Mixed Categorical and Continuous Data." *Journal of Econometrics* 148 (2): 186–200. doi:10.1016/j.jeconom.2008.10.007.

Luo, Y. and D. M. Dimitrov. 2019. "A Short Note on Obtaining Point Estimates of the IRT Ability Parameter with MCMC Estimation in Mplus: How Many Plausible Values are Needed?." *Educational and Psychological Measurement* 79 (2): 272–287. doi:10.1177/0013164418777569.

M. O. Martin, I. V. S. Mullis, and M. Hooper, eds. 2016. *Methods and Procedures in TIMSS 2015*Boston CollegeTIMSS & PIRLS International Study CenterDisponible en

Macdonald, K. 2019. "PV: Stata Module to Perform Estimation with Plausible Values." In *Statistical Software Components S456951*. Boston, MA, USA,: College Department of Economics.

Mancebón, M.-J., J. Calero, Á. Choi, and D. P. Ximénez-de-Embún. 2012. "The Efficiency of Public and Publicly Subsidized High Schools in Spain: Evidence from PISA-2006." *The Journal of the Operational Research Society* 63 (11): 1516–1533. doi:10.1057/jors.2011.156.

Meeusen, W. and J. van Den Broeck. 1977. "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error." *International Economic Review* 18 (2): 435–444. doi:10.2307/2525757.

Mislevy, R. J. 1993. "Should "Multiple Imputations" Be Treated as "Multiple Indicators"?." *Psychometrika* 58 (1): 79–85. doi:10.1007/BF02294472.

Mislevy, R. J., A. E. Beaton, B. Kaplan, and K. M. Sheehan. 1992. "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses." *Journal of Educational Measurement* 29 (2): 133–161. doi:10.1111/j.1745-3984.1992.tb00371.x.

O'-Donnell, C. J., D. P. Rao, and G. E. Battese. 2008. "Metafrontier Frameworks for the Study of Firm-Level Efficiencies and Technology Ratios." *Empirical Economics* 34 (2): 231–255. doi:10.1007/s00181-007-0119-4.

OCDE. 2016. *PISA 2015 Technical Report PISA*. Paris: OECD Publishing.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Studio Team. 2020. *RStudio: Integrated Development for R*. Boston: RStudio, PBC.

Rasch, G. 1960-1980. Probabilistic Models for Some Intelligence and Attainment tests, *Danish Institute for Educational Research: Copenhagen*. Expanded edition (1980) ed. The University of Chicago Press.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons.

Santín, D. and G. Sicilia. 2015. "Measuring the Efficiency of Public Schools in Uruguay: Main Drivers and Policy Implications." *Latin American Economic Review* 24 (1): 1–28. doi:10.1007/s40503-015-0019-5.

Simar, L. and V. Zelenyuk. 2006. "On Testing Equality of Distributions of Technical Efficiency Scores." *Econometric Reviews* 25 (4): 497–522. doi:10.1080/07474930600972582.

Thanassoulis, E., K. De Witte, J. Johnes, G. Johnes, G. Karagiannis, and M. C. Portela. 2016. "Applications of Data Envelopment Analysis in Education." In *Data Envelopment Analysis: A Handbook of Empirical Studies and Applications*, edited by J. En Zhu, pp. 367–438. New York: Springer.

Thieme, C., D. Prior, and E. Tortosa-Ausina. 2013. "A Multilevel Decomposition of School Performance Using Robust Nonparametric Frontier Techniques." *Economics of Education Review* 32: 104–121. doi:10.1016/j.econedurev.2012.08.002.

Todd, P. E. and K. I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113 (485): F3–F33. doi:10.1111/1468-0297.00097.

von Davier, M., E. Gonzalez, and R. Mislevy. 2009. "Plausible Values: What are They and Why Do We Need Them?." *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* 2: 9–36.

Wiseman, A. N., ed. 2010. *The Impact of the International Achievement Studies on National Education Policymaking*. Bingley: Emerald.

Wu, M. 2005. "The Role of Plausible Values in Large-Scale Surveys." *Studies in Educational Evaluation* 31 (2–3): 114–128. doi:10.1016/j.stueduc.2005.05.005.