Measuring the efficiency of European education systems by combining Data Envelopment Analysis and Multiple-Criteria Evaluation

Tommaso Agasisti
Politecnico di Milano School of Management
Department of Management, Economics and Industrial Engineering
Campus Bovisa - via Lambruschini 4b, Milan, IT 20156, Italy
tommaso.agasisti@polimi.it

Giuseppe Munda
European Commission, Joint Research Centre (JRC), Ispra, Seville, Italy

Ralph Hippe
European Commission, Joint Research Centre (JRC), Ispra, Seville, Italy

Please cite as: Agasisti, T., Munda, G. & Hippe, R. Measuring the efficiency of European education systems by combining Data Envelopment Analysis and Multiple-Criteria Evaluation. J Prod Anal 51, 105–124 (2019). https://doi.org/10.1007/s11123-019-00549-6

Abstract. Education is considered an important factor of economic growth, employment and social inclusion. Yet the economic crisis has also put ever more to the forefront the need to achieve educational goals in the most efficient way. The main objective of this paper is to assess the spending efficiency of European compulsory educational systems, creating a ranking of countries based on the efficiency scores of their systems using a number of standard variables from the literature. To this end, we also present a methodological innovation, that is to combine Data Envelopment Analysis (DEA) with discrete Multiple Criteria Evaluation (MCE), two methods that we consider complementary if used for providing a performance analysis. Moreover, both methods identify a set of common variables (such as some characteristics of teachers, the stock of adults' human capital, and lower expenditures per student) which are associated with higher levels of efficiency of the educational systems. The results show that findings using DEA are largely confirmed by MCE.

Key Words: Compulsory Education; Human Capital; Efficiency Analysis; Data Envelopment Analysis; Multiple-Criteria Evaluation, NAIADE method

JEL Classification: C14, C61, H52, I21

1 Introduction

Over the last couple of decades, educational policies have been characterised by a growing attention to the role that skills and educational results exert on economic and social development of countries and communities. Ever since literature has outlined the potential role of human capital (HC) in the process of economic growth, policy makers have been more and more interested in understanding factors that are correlated with the creation and development of people's HC (Benhabib and Spiegel 1994; Romer 1990; Barro 2001; Hanushek and Woessmann 2008, 2010 and 2012). In this context, the main practical aim of educational policy makers is to create the opportunities for maximising student achievement, by identifying the most efficient means to pursue this objective.

On one hand, the availability of international standardised tests allows for the benchmarking of educational systems across countries, with the aim of understanding the determinants of student achievement that seem more related to test scores, all else equal – e.g. international analyses such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). Several authors have used information from these 'internationally standardised' test scores to derive lessons about national-level outcomes – e.g. the conclusions drawn by Hanushek and Woessmann (2010) or the indications from the OECD's reports (OECD 2014a, 2014b). On the other hand, there is evidence that within the same scholastic system, some schools obtain better educational results than others, and that test scores are dependent on various students' personal characteristics and background (as noted since Coleman et al. 1966) more than on schooling. There are many papers that conduct empirical estimates about the determinants of such differences in educational results within a country between schools and across individuals (Greenwald et al. 1996), and many of them obtain similar findings, e.g. the role of the individual and socioeconomic status (SES) of schools (Perry and McConney 2010; Haveman and Wolfe 1995), the quality of teachers (Darling-Hammond 2000) and peer effects (Sacerdote 2011), etc.

A parallel stream of literature is one that discusses the efficiency of educational systems and organisations, not their absolute performance (i.e. test scores); in other words, the analysis focuses not on the overall results obtained by students, schools (on average) or education system (as a whole), but on the ability of reaching such results by using the least amount of possible resources – or, conversely, maximising educational results with the available resources (Agasisti 2011; Johnes 2004). In this type of analysis, then, inputs enter the picture – i.e. the specific intent of the empirical study is to consider how many resources are employed in order to obtain those results, and not just the level of the educational results. This way, the empirical analysis also has to deal with the collection of data concerning the inputs, and it has to model the process of transforming the inputs (resources) into outputs (educational results). This is the well-known theoretical framework which is known as the 'educational production function' (EPF) in the economics of education literature. As such, two levels of analysis can be considered:

- one that poses its attention on the 'spending efficiency' at country level (i.e. how the financial resources allocated to education are used, and what average educational performance they are able to 'produce'?), and

- the other that considers each single school/university as an organisation that uses financial and human resources to produce (average) educational achievement among its students (this is named 'technical efficiency' of the institution).

The importance of improving efficiency is not confined to single countries or specific grades, but instead, it is a central aspect in recent studies concerning educational challenges as an imperative for the future (Hanushek and Luque 2003; Sutherland et al. 2010) and discussion about ways to improve efficiency is faced by many governments – both domestically as well as international. In parallel, even in recent times, academic literature has debated the strengths and weaknesses of using theories and methods to assess the efficiency of educational provision. For example, Grosskopf et al. (2014) discuss how models of educational efficiency can be leveraged for exploring policy solutions directed at improving educational outcomes, thus solving the methodological problems that are at the heart of any assessment of this kind.

In this paper, we pay attention to the selection of the relevant variables and to the empirical approaches that can be used to measure spending efficiency in educational systems. More specifically, we will innovate current academic literature in the field by combining different approaches, such as frontier methods and multiple-criteria evaluation in order to check the robustness of our main findings. The paper is innovative because current studies that compare the performance of educational systems base their findings on a single method at a time. Instead, in this work, we aim is to test how two complementary approaches reveal the efficiency and performance of educational systems, and whether the findings that are obtained are coherent or not. On the one hand, in order to do this, the paper employs a frontier method – Data Envelopment Analysis (DEA) for assessing the efficiency of the educational system. On the other hand, we will use a discrete multiple-criteria evaluation (MCE) method for highlighting which countries are able to obtain the highest levels of an educational system's performance given the available inputs. This is a complementary way to illustrate the concept of efficiency and has been discussed methodologically by Bouyssou (1999). As such, the paper's innovation stems from observing the 'ranking' of educational systems emerging from the two techniques with the aim of showing whether the two methods generate similar rankings, despite the different assumptions in the technicalities for modelling the educational production process. Another innovation lies in the choice of focusing on country-level data: efficiency measurements are common for single educational organisations (cf. the reviews in Worthington 2001; Johnes 2004; De Witte and López-Torres 2017) but they are only seldom conducted when compared to the overall performance of systems at country-level (cf. Agasisti 2011). In this vein, our paper fills the gap in terms of a lack of empirical literature that explicitly considers the production of educational outputs at an aggregate (country) level.

Frontier methods are the traditional approach in assessing efficiency in education economics and management; the most popular ones being non-parametric methods such as Data Envelopment Analysis (DEA) and parametric methods such as Stochastic Frontier Analysis (SFA). In addition, Multiple-criteria evaluation (MCE) is divided into continuous (i.e. extensions of traditional linear programming methods (cf. Antunes et al. 2016; Steuer 1986)) and discrete (i.e. where the number of options is finite in number) approaches. Multi-criteria evaluation has widely been used in various fields since the sixties both at micro and macro levels of analysis (see e.g. Figueira et al. 2016); common applications in public policy refer to energy, finance, sustainable development, land use and regional planning. In the framework of education policy, the desirability of the peculiar characteristics of multi-criteria evaluation has been advocated by various authors (e.g. Dill and Soo 2005; Guskey 2007; Ho et al. 2006; Malen and Knapp 1997; Nikel and Lowe 2010; Rossell 1993; Stufflebeam 2001; Tzeng et al. 2007). While continuous multi-objective approaches are still related to frontier methods (in particular, they can be considered an attempt to improve DEA techniques). Discrete multi-criteria methods are based on complete different assumptions, such as the non-exclusion of dominated options or non-compensability; from this point of view, they can be considered as a complementary approach. In order to show the potentialities of combining DEA with MCE, this paper analyses the efficiency of compulsory education in the EU. Efficiency is measured by modelling the educational production function based on the literature in the field (De Witte and López-Torres 2017).

In this paper, we will answer the following research questions: which are the most efficient European educational systems? And is the ranking of most performing educational systems affected by the choice of the method that is employed for the empirical assessment? To the best of our knowledge, this is the first paper that combines non-parametric methods with multi-criteria techniques for efficiency analyses in the field.

The paper is organised as follows. In sect. 2, we will introduce the problem of modelling the production of educational outputs and the issue of selecting appropriate variables for this purpose. In sect. 3, we describe the two methodologies that we will employ here for empirical applications: Data Envelopment Analysis (DEA) and Multi-Criteria Evaluations (MCE), together with details of the application of these methods and the description of the dataset. Sect. 4 will present the main findings from the empirical analyses and sect. 5 will discuss these results and finish with concluding remarks.

2 Measuring educational efficiency in practice: the selection of inputs and outputs

From an empirical point of view, the main discussion revolves around the selection of variables that are relevant for the measurement of efficiency: inputs, outputs and contextual variables. With the latter group of indicators, the analyst aims to describing the factors which are statistically associated with higher/lower efficiency scores (i.e. after they are calculated). While efficiency scores (*eff_scores*) represent the ability of transforming inputs (e.g. resources) into outputs (e.g. test scores), this second level of analysis investigates whether there are recurrent factors that are statistically correlated with such scores.

The selection of inputs and outputs is a crucial task in efficiency studies (Coelli et al. 2005; Cooper et al. 2011); indeed, the ability of describing actual efficiency differentials stems from the precision in the description of the production process. Inputs include those factors that are used for producing educational services. They can be classified in three broad groups: (i) financial resources (of various types and with various destinations), (ii) human resources (those devoted to educational activities and support staff), and (iii) facilities – which can be consumables or the use of infrastructure. Outputs are supposed to be a measurement of the results of the educational services that are offered, i.e. the amount of education 'produced' in a given country. Ideally, taken together, such measures should include indications concerning the quality and quantity of the services that are produced, and should refer exclusively to the output – and not the outcome. Indeed, literature in the field of public management associates the concept of effectiveness to the comparison between outcomes and inputs (interesting discussions in Golany and Tamir 1995; Moore 1995). Nevertheless, in the educational literature, outputs are usually measured as achievement, test scores, graduation rates, etc. – something that is more similar to the effects of the educational services, than the quantities which are produced.

As mentioned above, input variables can reflect both physically available resources (books, building, computers, classrooms, buses, grants, etc.) and expenditures (teaching, research, administrators, support staff) and concerning the extent to which prices are accounted for, they represent two sides of the same coin. The number of teachers is a key input that is used in several studies, and is expressed in various ways – frequently, in the form of a student-teacher ratio. Sometimes, proxies are included in the vector of the inputs themselves as a means to control the differences in the quality of the inputs, teachers' experience or qualifications (among others, cf. Sarrico et al. 2010). Another key input in educational processes refers to the quality of the student intake, especially their socioeconomic background, innate ability and prior achievement level. In many cases, however, this particular input is treated as 'non-discretionary', in the sense that the quality of the student body cannot be modified by the policy makers (a methodological discussion about how dealing with such kind of inputs is provided in Cordero-Ferrera et al. 2008). As such, it is alternatively treated as a contextual factor (this is also the choice adopted in this paper, which focuses on the amount of test-score variability explained by students' socio-economic background as an indicator). A growing body of literature also pays attention to the role that certain managerial practices and/or innovations and/or specific educational processes can play in affecting outputs (cf. Haelermans and De Witte 2012; Mancebón et al. 2012). Therefore, these elements are much more classifiable among the contextual variables than among inputs. In other words, they deal more with the use of inputs, and not with the quantities or qualities of the inputs. The information about the governance models (school autonomy, share of private schools, etc.) is also frequently used for comparing the efficiency of public and private schools – paralleling the literature that compares raw performance between these types of schools (cf. Dronkers and Robert (2008) for an international comparison). The group of contextual variables may also include those variables that reflect the community in which the schools operate: indicators for competition among schools, neighbourhood characteristics, urban/rural areas, educational level of the population in the area. Outputs are typically measured by test scores in a standardised evaluation of achievement. Some studies, however, also consider other output measures, such as the drop-out rates (Alexander et al. 2010) or attendance rates (Grosskopf and Moutray 2001). When considering educational efficiency at the level of the specific system, other outputs may be related to (i) the proportion of students who abandon education without a secondary qualification (early leavers) and (ii) the proportion of students who are neither engaged in education nor in the labour market.

The contextual variables can be divided into sub-groups:

- those variables that are contextual characteristics of the educational institutions (*features* and *processes* set by the institution itself and/or by policy makers). Thus, the institution can indeed modify the system's efficiency by acting on these levers. In this specific sense, exploring the correlations between efficiency scores and these contextual variables can be useful, as evidence can be used (with caution) to understand which recurrent factors can be found in educational systems with higher/lower levels of efficiency;

- those variables that describe the *external context* in which the institution operates (i.e. the wealth characteristics of an area, the proportion of immigrants residing there, etc.). This second sub-group of variables can be broadly considered as related to factors that are 'external' to efficiency measurement (i.e. the school/ HEI cannot modify the features of the place in which it operates, although they have an effect on their operations). Considering this group of factors as a separate group, it is important to calculate the efficiency of educational systems without the risk of bringing external influences into the picture.

In this paper, although we do not incorporate contextual variables into the calculation of efficiency scores, we propose a descriptive analysis of correlations between them and efficiency scores.

3 Methodological approaches for assessing efficiency in education: frontier methods and multiple-criteria evaluation

In this section, we will briefly review the main characteristics of frontier methods and multiple-criteria evaluation and explain why they can be considered as complementary approaches.

3.1 Methodological approach 1–Data Envelopment Analysis

The basic idea at the core of the Data Envelopment Analysis (DEA) is to assess the extent to which the output can be increased, given the available inputs (output-oriented models) or, conversely, the extent to which inputs can be reduced given the produced output (input-oriented models) – Fried et al. (2008); Charnes et al. (2013) and Zhu (2015) all provide a methodological discussion of how the technique for measuring efficiency and performance works. The method is very useful in a multi-input / multi-output context because the technique is able to handle several inputs and outputs at the same time, thus reducing the judgment about the production efficiency to a single-number indicator.

The method is completely non-parametric, as it does not employ any functional form of the production process – this is also a nice property, given that knowledge about the educational production function is still very limited and assumptions about the relationships between inputs and outputs can sometimes be non-verifiable, especially in the context of comparisons between the efficiency of educational systems and not between single institutions.

The DEA methodology has been extensively used in the context of assessing the efficiency of educational institutions, systems and activities – the interested reader can refer to good summaries of existing studies as Worthington (2001); Johnes (2004); De Witte and López-Torres (2017) as well to publications that summarise the applications of DEA in the literature more generally, cf. Liu et al. (2013) and Emrouznejad and Yang (2018).

Mathematically, the DEA method can be illustrated as a problem of maximising the ratio between the sum of outputs and the sum of inputs for each institution (sums of inputs and outputs are obviously standardised to account for different units of measurement). We first define the technical efficiency of each i-th institution (*eff_i*) as follows, considering y_o as outputs [with *$o = (1, \dots s)$*] and x_j as inputs [*$j = (1, \dots, m)$*], with w_o and v_j as the weightings for the o outputs and j inputs, respectively:

$$eff_i = \frac{\sum_{o=1}^{s} w_o y_{oi}}{\sum_{j=1}^{m} v_j x_{ji}} \tag{1}$$

Then, the DEA efficiency score of each i-unit is the one that maximises the unit's efficiency score, by combining the weightings in the optimal way:

$$\max \frac{\sum_{o=1}^{s} w_o y_{oi}}{\sum_{j=1}^{m} v_j x_{ji}} \tag{2}$$

In this sense, the resulting efficiency score is the one that sheds the best possible light on the i-th institution's performance. For obtaining the efficiency scores, the fractional problem is transformed into the dual one, and then solved with linear programming. Specifically, a typical DEA formulation is one where (see Charnes et al. 1978):

$$\min \emptyset_i \tag{3}$$

is subject to:

$$\emptyset_i x_{ji} - \sum_{z=1}^{n} \lambda_z x_{jz} \geq 0 \qquad j = 1, \dots, m \tag{4}$$

$$\sum_{z=1}^{n} \lambda_z y_{oz} \geq y_{oi} \qquad o = 1, \dots s \tag{5}$$

$$\lambda_z \geq 0 \quad \forall z = 1, \dots, n \tag{6}$$

The value $TE = 1/\emptyset_i$ represents the efficiency score of the i-th unit, and mathematically, it is constrained to be in the range [0;1]. The formulation cited above relates to a model that is known as 'input-oriented', i.e. the main assumption is that the unit under observation (i.e. the school, the university) attempts to produce the outputs (attainment, test scores, graduation rates, etc.) while minimising the use of the available resources (personnel, facilities, etc.). It has been argued that such an approach (known as 'input-oriented') is more suited to circumstances where input reductions (i.e. budget cuts) are in action (Cunha and Rocha 2012), such as those used in the recent financial crisis. It is important to note that in the mathematical formula used in (1)–(3) the model assumes that constant returns to scale – and the empirical application performed in this paper is realised coherently.

When compared to the parametric methods for evaluating efficiency, DEA reveals some key advantages: (i) it can employ several inputs and outputs at the same time, (ii) it does not require a specification a priori of the functional form for the production function, (iii) it allows each unit to have its own objectives, through the free/automatic determination of weights for each input/output, and (iv) efficiency is determined by using observed performance levels, that is to say (linear combination of) real units operating in the sector, so that they constitute a real (achievable) reference point. These advantages come at a cost, though. Indeed, the method is completely deterministic, i.e. any deviation of the units from the frontier is considered as a fruit of inefficiency, which can well be due to measurement errors and random noise – and there is no way to check this (as a consequence, efficiency scores cannot be considered in second stages for inferential analyses of their determinants).

For the purpose of this work, we employ a simple DEA approach, although we are obviously aware of the existence of more sophisticated versions of non-parametric methods, such as the bootstrap DEA proposed by Simar and Wilson (2000), the order-m frontiers (Cazals et al. 2002) and the conditional efficiency scores developed by Daraio and Simar (2007). The choice of relying on a very basic DEA model is justified by the main aim of this contribution, which is not to perform exploration into the maximum robust efficiency, but to combine two alternative methods (DEA and MCE, see next section) in order to provide an assessment of the efficiency of educational production at a country level. In this spirit, we value the simplicity of understanding the efficiency scores as obtained through a simplified DEA framework.

3.2 Methodological approach 2–Multi-Criteria Evaluation

Multi-Criteria Evaluation models (MCE) can be used to assess the multidimensional performance of an educational system. MCE has been related to DEA in different ways (for example, Doyle and Green (1993); Joro et al. (1998); Li and Reeves (1999); Madlener et al. (2009)).

A '*discrete multi-criterion problem*' can be formally described as follows (cf. Arrow and Raynaud 1986; Figueira et al. 2016; Munda 2008; Nijkamp et al. 1990; Roy 1996). *A* is a finite set of *N* feasible actions (or alternatives). *M* is the number of different points of view, or evaluation criteria g_m, that are considered relevant to a specific policy problem. This supposes that action *a* is evaluated to be better than action *b* (both belonging to the set *A*), by the *m-th* point of view, then $g_m(a) > g_m(b)$. In this way a decision problem may be represented in an *N* by *M* matrix *P* known as an *evaluation or impact matrix*. In such a matrix, the typical element p_{ij} $(i = 1, 2,\dots, M; j = 1, 2,\dots, N)$ represents the evaluation of the *j-th* alternative by means of the *i-th* criterion (see Table 1). The impact matrix may include quantitative, qualitative or both types of information. In general, in a multi-criterion problem, there is no solution (ideal or utopia solution) optimising all the criteria at the same time and therefore '*compromise solutions*' have to be found.

		Alternatives			
Criteria	**Units**	**a_1**	**a_2**	**a_3**	**a_4**
g_1		$g_1(a_1)$	$g_1(a_2)$	.	$g_1(a_4)$
g_2		.	.	.	.
g_3		.	.	.	.
g_4		.	.	.	.
g_5		.	.	.	.
g_6		$g_6(a_1)$	$g_6(a_2)$	.	$g_6(a_4)$

Table 1. Example of an Impact Matrix

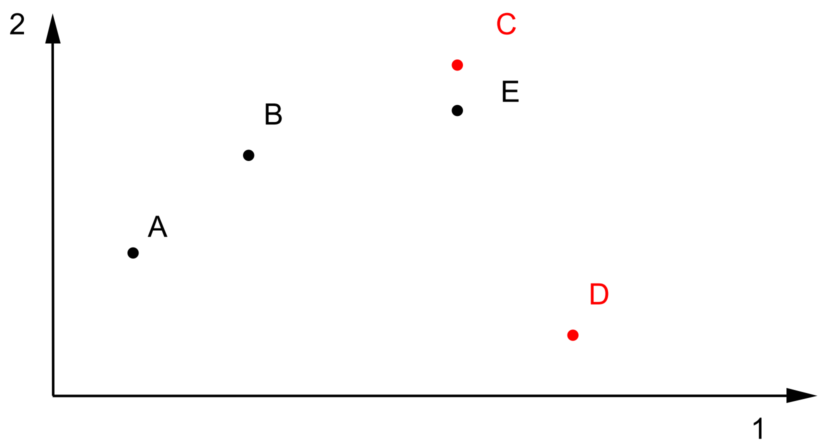

Figure 1. Graphical Representation of Efficiency in a Two-Dimensional Case

In the context of discrete multi-criteria methods, the whole concept that only efficient alternatives have to be taken into account is questioned. From this point of view, as this concept is the key assumption of all frontier based approaches, discrete multi-criteria approaches can be considered as being complementary and particularly useful for testing the robustness of DEA results. The concept of efficient alternatives can easily be illustrated graphically (see Fig. 1 which refers to a 2-criteria state).

Alternative **C** performs better than **B** in all respects and hence **C** is preferred to **B**. The same can be said for **B** compared to **A**. As such, only **C** and **D** are efficient alternatives. It is necessary to point out that efficiency does not imply that every efficient solution is necessarily preferred to every non-efficient solution; e.g., the non-efficient alternatives **A** and **B** are preferable to the efficient alternative **D** if the second criterion has a higher priority compared to the first criterion. The principle that inefficient solutions can be ignored (often presented as a simple technical step) assumes the acceptance of the following assumptions:

- The assumption that all the relevant criteria have been identified needs to be accepted. If relevant criteria are omitted, there are potential opportunity costs associated with assuming that it is safe to ignore dominated alternatives.

- The assumption that only one alternative (the best) is considered has to be identified needs to be accepted (α problem formulation). This is because 'second best' may have been eliminated during the technical screening if more than one action has to be found. The elimination of the 'inefficient' action may result in an opportunity loss (it is necessary to point out that if the best action is removed from the set of feasible alternatives, then the second best becomes a member of the non-dominated set). If one is interested in the γ problem formulation, then dominated alternatives cannot be eliminated. It is necessary to point out that in public policies, it is often much more useful to have a ranking of policy options than to select just one alternative.

- Arrow's axiom of 'the independence of irrelevant alternatives' states that the choice made in a given set of alternatives *A* only depends on the ordering made with respect to the alternatives in that set. Alternatives outside *A* (irrelevant since the choice must be made within *A*) should not affect the choice inside *A*. Empirical experience does not generally support this axiom; i.e. excluding some actions already inside *A* can have even less justification.

As a conclusion of this discussion we can state then that, when the set of alternatives is a finite one, it makes sense to use mathematical aggregation procedures that do not exclude dominated alternatives a priori. In the framework of efficiency analysis, this conclusion implies that results obtained through traditional frontier methods should always be corroborated by also using non-frontier based mathematical approaches, such as multi-criteria methods.

It is relevant to isolate some properties that can be considered desirable for a discrete multi-criteria method within the framework of efficiency assessment in education systems. In synthesis, the information that is contained in the impact matrix is useful for solving the so-called multi-criterion problem: *Intensity of preference* (when quantitative criterion scores are present), *Number* of criteria in favour of a given alternative, *Weighting* attached to each single criterion, *Relationship* of each single alternative to all the other alternatives.

Combinations of this information generate different aggregation conventions, i.e. manipulation rules relating to the information available in order to arrive at a preferential structure. The aggregation of several criteria implies taking a position concerning the fundamental issue of compensability. Compensability refers to the existence of trade-offs, i.e. the possibility of offsetting a disadvantage relating to some criteria through a sufficiently large advantage concerning another criterion, whereas smaller advantages would not do the same (there is an analogy with the concept of 'weighting' in the DEA). As such, a preference relation is non-compensatory if no trade-off occurs and is otherwise compensatory (Munda 2012). It is important to understand that compensability means that in an education efficiency assessment exercise, an improvement in one of the spending side criteria can easily compensate worsening in an output criterion, e.g. PISA scores in science! An important related point to consider is that the existence of preference independence is a necessary condition in order to use a full compensatory linear aggregation rule (cf. Keeney and Raiffa 1976). From an operational point of view, this means that an additional aggregation function allows for the assessment of the marginal contribution of each criterion separately (as a consequence of the preference independence condition). The marginal contribution of each criterion can then be added together to yield a total value. For example, this implies that among the different aspects of the output variables, there are no synergy phenomena or conflicts, i.e. preference independence considers each single score as being fully unrelated to all the others. Indeed, can courage be evaluated as a positive characteristic of a person, without knowing if he/she is a dedicated criminal or an enthusiastic medical doctor? From an education policy point of view, this implies that, for example, interactions among PISA scores in reading and mathematics are not possible. This is rather unrealistic from a scientific point of view. As such, we can safely state that complete compensability is not desirable for assessing the efficiency of education systems.

Any attempt to measure efficiency should deal with the following two questions: (i) are results robust enough in terms of the selection of specific variables (selection of inputs and outputs)? and (ii) are results robust enough in terms of the selection of a specific method for efficiency analysis? MCE is a consistent approach in order to answer both these questions.

In particular, the first question is dealt with by constructing a multi-criteria impact matrix where the various input and output items are integrated by considering the plurality of available data sources. Here, a difference to the previous DEA analyses is that some contextual variables, such as educational attainment, are considered as output variables.

Finally, the so-called NAIADE method (Munda 1995) has been applied to rank the different countries[1]. This last analysis can be considered as a complementary measure of efficiency of the education systems at country level with respect to the one obtained by DEA; this is aimed at answering question two.

The whole NAIADE mathematical aggregation procedure can be divided into three main steps:

1. paired comparison of alternatives according to each criterion,
2. aggregation of all criteria,
3. ranking of alternatives.

Briefly, given the information on the paired performance of the alternatives according to each single criterion, it is necessary to aggregate these scores in order to simultaneously take all criteria into account. This objective is achieved by using a kind of 'concordance index' aggregating the various degrees of credibility which are obtained according to the criteria that is used. This is done by introducing parameters that allow the *degree of compensability* that is desired in the problem at hand to be established. No criterion weighting is used.

From a mathematical point of view, this concordance index is described by the following equation:

$$\mu_*(a,b) = \frac{\sum_{m=1}^{M} \max(\mu_*(a,b)_m - \alpha, 0)}{\sum_{m=1}^{M} \left| \mu_*(a,b)_m - \alpha \right|} \quad (7)$$

where $\mu_*(a,b)_m$ is the evaluation of an indifference or preference fuzzy relation for each pair of options, according to the *m-th* criterion. Parameter α is a minimum requirement imposed on each fuzzy relation. The higher α is, the lower the compensability that is allowed in the aggregation process. Moreover, in order to have information concerning the diversity (and as such, the possible compensability effects) of the assessments of each single fuzzy relation according to each criterion, a

[1] The NAIADE method can be considered particularly useful for efficiency analyses in the field of education for four main reasons:

1. it has been explicitly designed for public policy applications;
2. it is flexible, as it can deal with different source of information concerning the criterion scores;
3. compensability can be fully controlled;
4. it can also be used for benchmarking exercises.

diversity measure *C* is introduced. By combining both $\mu_*(a,b)$ and its diversity measure, the following global pairwise evaluation can be derived:

$$a\,is\,better\,than\,b = \frac{\mu_{>>}(a,b) \wedge C_{>>} + \mu_{>}(a,b) \wedge C_{>}}{C_{>>} + C_{>}} \tag{8}$$

$$a\,is\,worse\,than\,b = \frac{\mu_{<<}(a,b) \wedge C_{<<} + \mu_{<}(a,b) \wedge C_{<}}{C_{<<} + C_{<}} \tag{9}$$

$$a\,and\,b\,are\,indifferent = \frac{\mu_{==}(a,b) \wedge C_{==} + \mu_{=}(a,b) \wedge C_{=}}{C_{==} + C_{=}} \tag{10}$$

where the operator ^ can be fully compensatory, non-compensatory or partial compensatory.

The final ranking of the alternatives, in a complete or partial pre-order, is obtained by means of the basic idea of *positive* (i.e. credibility that a number of options is worse than the one considered) and *negative* (i.e. credibility that a number of options is better than the one considered) *flows*. Again, the aggregation operator can be fully compensatory, non-compensatory or partially compensatory. All NAIADE mathematical details can be found in Munda (1995).

3.3 Data and modelling

This study is about efficiency in compulsory education at the level of the system. As such, we include data concerning primary and secondary education. The data used in this study is derived from Eurostat and OECD databases[2] (for exact sources, see the Supplementary File). Färe and Karagiannis (2017) discuss how the rules adopted for aggregating individual or institutional data at higher levels of analysis is important in affecting the results; we do not face this problem as we employ data that are already defined, developed and reported at national level by expert international organisations which have the task of maintaining official statistics at country level.

We present a synthetic description of the variables used in Table 2[3]. Among the inputs, the student-teacher ratio is considered as a proxy for the labour intensity of the educational system, and it is obviously influenced by the average size of classes and schools, which may be structurally different across countries. Also, average salaries are different in the various educational systems and this justifies the inclusion of the government expenditure per student (*expstud2avg*). In turn, this neutralises the major heterogeneity in the compensation levels for teachers. As a consequence, the amount of government spending on education (as a proportion of GDP) can indirectly act as a measure of two phenomena: the priority assigned by governments to education and the amount of financial

[2] Whenever possible, preference has been given to Eurostat data because data are usually available for more countries. We have chosen 2012 as our reference year, as most series of data are only available up to 2013. As such, the exercise cannot be performed for recent years. This is because the most relevant output variable, i.e. PISA scores, is only available for 2012 (not 2013). Exceptions include data concerning teachers (age and salaries) which is only available for 2013. Where possible, we calculate an average over the period 2010 to 2012 in order to avoid potential bias in one given year. In the case of expenditure data, we were not able to calculate averages but instead, we chose to only include the year 2012.

[3] In many cases, data are separately available for primary, lower and upper secondary education. For this reason, depending on the individual variable specification, we have summarised the various education levels or averaged them in order to get a measure for both primary and secondary education together. In the latter case, we used weightings in order to assign the appropriate relative share for each education stage (for example, we weighted by the number of pupils at each level) – see Supplementary File 1 for the specific weighting schemes applied in each case.

resources available for investment in education (name of the variable: *expgdp*). When considering outputs, three main indicators have been chosen. Firstly, we used test scores in two domains – reading and mathematics – relying upon countries' averages and rankings as calculated by OECD in its Programme for International Student Assessment (PISA), 2012 edition[4]. Secondly, we consider the proportion of young people in the age range 18–24 who left the school system without a secondary education qualification (esl). This is a proxy for the output dimension that is also targeted by the EU 2020 strategy which aims at reducing this proportion of 'early leavers'. Thirdly, we refer to the number of NEETS, who are people in the age 15–34 who are unemployed, not in education activities and who are not actively searching a job either (neet).

An issue that deserves a specific discussion is the lack of time alignment between the outputs – which represent the current level of educational production at country level – and the inputs. The latter only measures the current level of investment, while the outputs are produced by cumulative investments realised over a long time span. For example, the test scores of students at age 15 (our indicator PISA test scores) are influenced by student-teacher ratios and expenditures realised by countries in terms of compulsory education attended by students in previous years. Two factors limit the possibility of building adequate indicators for cumulated inputs. On one hand, the country-level datasets are relatively recent, and sometimes characterised by missing data or changes in the definition of variables. On the other hand, there are no longitudinal datasets at country level that follow cohorts of students precisely. As a consequence, the present study must assume that current differences in input levels are a proxy of differences in the investments realised by countries over the previous years. The assumption is not too hard when considering that changes of aggregate measures of educational inputs do not vary dramatically in the short run.

Seven key country-level dimensions have been selected for a description of contextual specific features of the educational system. These are used for analysing correlations with efficiency scores[5]. Two variables relate to the characteristics of teachers: their average age and salaries (*teacherage* and *teachersalavg*, respectively). The proportion of immigrants (*migr*) can be an indicator about difficulties faced by educational institutions to cope with individuals who come from different backgrounds and experiences. We also consider the measure concerning the index of economic, social and cultural status (ESCS) of students (stud_background); this indicator could also be interpreted as a proxy for the quality of student intake and as such, should be included among the inputs. As anticipated, however, we tend to consider it as non-discretionary (i.e. cannot be easily altered by specific policies decided by policy makers). As such, we prefer to list this variable among the contextual variables. The subsequent interpretation is to check whether the 'technical efficiency' of educational systems is somehow affected by the (exogenous) socio-economic characteristics of the students they have to educate. In particular, we focus on the amount of test scores which can be explained by the students' socioeconomic background (within country), which can be interpreted as a proxy for the equity of the educational system. The numeracy scores obtained by adults (country average) in the OECD

[4] We have not dealt with the problem of defining and interpreting plausible values for these scores (following the technical instructions provided by OECD (2014b) and commented on by Wu (2005)) as we directly used the country averages as officially calculated and reported by OECD in their institutional reports.

[5] We rely on a simple correlation analysis between efficiency scores and the contextual factors that describe the characteristics of educational systems. We are well aware of the drawbacks of two-stage procedures as pointed out by Simar and Wilson (2011). Indeed, the problem of the two-stage approach is that it relies on a separability condition between the input-output space and the space of the external factors, assuming that these factors have no influence on the attainable set, which may not hold in some situations. For this reason, we limit ourselves to simple descriptive correlations and do not perform any second-stage regression of the kind criticised under the methodological indications recalled here.

Programme for International Assessment of Adult Competences (PIAAC) is a measure of the stock of human capital in a given country, which also considers a qualitative dimension. A similar indicator, which only focuses on quantity, is the average highest level of education of the adult population (*eduatt2*). Lastly, the share of public expenditure devoted to education (*expsharetot*) relates to direct information about how much education is a political priority for a country, and it would be interesting to see if these data have any correlation with the spending efficiency on education of various countries.

Variable	Description	Year
Input variables		
Teacher ratio	student/teacher ratio [ISCED 1-3]	average 2010-2012
expstud2avg	Government expenditure per student in primary and secondary education (constant PPP$)	average 2010-2012
expgdp	Total public expenditure on primary, lower and upper secondary education as % of GDP	2012
Output variables		
pisa_read	PISA reading scores	2012
pisa_math	PISA mathematics scores	2012
esl	100 - (share of early schooling leavers) [18-24 year olds]	average 2010-2012
neet	100 - (share of NEET) [15-34 year olds, non-employed persons not in education or training]	average 2010-2012
Contextual variables		
teacherage	Average teachers' age, lower secondary (TALIS)	2013
teachersalavg	Teachers' Salaries (in Euro, converted using PPS for household final consumption expenditure), averaging primary, lower secondary and upper secondary education.	2013
migr	100 - (share of migrants among 15 year old pupils participating in PISA)	2012
piaac_num	PIAAC numeracy scores	2012
eduatt2	Higher educational attainment level from 25 to 64 years, ISCED 5-8	average 2010-2012
expsharetot	Total public expenditure on primary, lower and upper secondary education as % of public expenditure	2012
stud_background	Amount of test scores' variability explained by the PISA Index of economic, social and cultural status	2012

Table 2. Overview of variables

Note: Given the short span of the period analysed, the use of constant prices (as expressed with PPP, Purchasing Power Parity) instead of current prices should not represent a sensible change in the results of the empirical exercise.

Model	**Inputs**	**Outputs**
Model_1 (DEA)	• (inverse of) students:teachers ratio • expenditure per student	• PISA test score, mathematics • (100-) % early school leavers
Model_2 (DEA)	• (inverse of) students:teachers ratio • expenditure per student • expenditure as % GDP	• PISA test score, reading • (100-) % early school leavers
Model_3 (DEA)	• (inverse of) students:teachers ratio • expenditure per student • expenditure as % GDP	• PISA test score, reading • (100-) % NEET
Model_4 (DEA)	• (inverse of) students:teachers ratio • expenditure per student • expenditure as % GDP	• PISA test score, mathematics • (100-) % early school leavers • (100-) % NEET

Table 3. The models estimated for calculating efficiency scores

	teachstud	expstud2avg	expgdp	pisa_read	pisa_math	esl	neet
Teachstud	1						
expstud2avg	0.1649	1					
Expgdp	-0.1016	0.5022	1				
pisa_read	-0.1173	0.3263	0.2191	1			
pisa_math	-0.1139	0.2841	0.0489	0.9164	1		
Esl	-0.0316	0.1528	-0.016	0.1391	0.2775	1	
Neet	0.1046	0.6672	0.3422	0.3084	0.4483	0.4165	1

Table 3.a Correlations between inputs and outputs

As a first step, we apply DEA to input and output variables, with the aim of calculating the efficiency scores. Then we corroborate these results with a technique based on MCE. Following this, we try to understand the relationship between the results obtained (efficiency scores) and some contextual variables, in a descriptive way. Table 3 describes the various DEA models which are used to carry out the efficiency analysis, each of which are characterised by different combinations of inputs and outputs. Here, it is worth providing some explanations about the choices of the various models and for this purpose we refer to the Table 3a, which reports correlations between inputs and outputs. Firstly, the correlations between PISA maths and reading scores are very high (>0.9), so they can be used alternatively without altering the meaning of the main results. As a consequence, we employed reading for models 2 and 3, while using mathematics for models 1 and 2. We decided not to use them simultaneously in order to avoid collinearity, given the limited number of observations. The first model is the baseline, with just two inputs and two outputs, while in all the other models we added the third input for a more complete picture of the resources available. The models 2, 3 and 4 use alternative sets of outputs in order to check the sensitivity of efficiency results (due to the different specifications of the outputs). The high correlations (as presented in Table 4) across efficiency scores reveal that the rankings of countries produced by alternative models are substantially similar. In this vein, our preferred model is number 4, where three outputs are jointly considered. It is important to note that early school leavers and NEET students are only moderately correlated (around 0.4), so the inclusion of both indicators aims at capturing two really different dimensions of educational systems' results. This is our preferred model because it includes all the three outputs together, using all the inputs available. As such, it is the most complete representation of the efficiency of educational spending that is possible with the data employed in this work.

All the models have been estimated assuming input-orientation and constant returns-to-scale (CRS).

4 Efficiency analysis of compulsory education in the EU Member States–results

4.1 Baseline results from Data Envelopment Analysis

Tables 4 and 5 contain the main findings obtained through the DEA efficiency analysis.[6] Looking at the average efficiency scores from the different models, we understand how adding more inputs and outputs increases the average score; however, the correlation matrix indicates that no major effects influence the relative ranking of efficient educational systems.

[6] Note that we cannot compute efficiency scores for all countries, as we have some missing values in the data on both the input and output side. However, the missing values affect the same countries in each model (for example, both PISA maths and reading scores are not available for Malta), so that each time the same countries are dropped in the calculation of the efficiency scores. In other words, we cannot compute scores for Croatia, Denmark, Greece and Malta in all models.

Panel A. Correlations between the efficiency scores obtained through the different models

	Model 1	Model 2	Model 3	Model 4
Model 1	1			
Model 2	0.9484	1		
Model 3	0.9227	0.9875	1	
Model 4	0.9540	0.9971	0.9812	1

Panel B. Average efficiency scores obtained through the different models

	Model 1	Model 2	Model 3	Model 4
Mean	0.8214	0.8614	0.8535	0.8544

Table 4. The efficiency scores obtained through the different models

The results reveal that, on average, the educational output level produced in the European countries can be maintained while reducing the amount of resources invested by 15–18%, depending on the specific model used for the analysis. When considering our preferred specification (Model 4), a group of efficient countries can be observed: Germany, Estonia, the Netherlands, Romania, Slovak Republic and the UK. These educational systems are quite heterogeneous when compared against each other, confirming the idea that there are several different ways to be efficient. For example, exploring the dataset, we find that Estonia is characterised by low expenditures per student and very high test scores; while Romania, which reports PISA test scores well below the international average, has an even lower level of expenditure per student proportionally. As such, it appears that the educational system is making the most of the (few) available resources. The Italian educational system, for example, is judged to be relatively inefficient because, despite reporting average test scores similar to those of the Slovak Republic, it spends 1/3 more per student and almost 1% more of its (higher) GDP on education. Spain, like some other countries, is estimated to have an inefficient educational system because of the high proportion of early school leavers (26%) and neets (22%). Overall, countries from continental Europe and Nordic countries appear to have inefficient educational systems because, notwithstanding fairly high outputs (good PISA test scores, low proportions of early school leavers and neets), they spend an amount of money per student well above the international average (typically, >12,000$ PPP).

4.2 Complementary approach: multi-criteria analysis

In this subsection, we derive an assessment of the efficiency of educational systems by using a specific multi-criteria evaluation approach (NAIADE), with the aim of checking the coherence of results with DEA analyses. Specifically, we carry out a multidimensional evaluation, where the various input and output items are all integrated together.

In multi-criteria terms, efficiency can be defined as a compromise solution between inputs and outputs. On the input side, we include all the expenditure variables; this is a comparative advantage of using the multi-criteria evaluation. As such, we consider expenditure on education as a share of

GDP (in two specific definitions, expgdp and expgdp-2), the expenditure on education as a share of total expenditure (expshare) and the educational expenditure per student (exp-stud2). The same applies to the output point of view, where we have considered pisa-reading, pisa-maths, pisa-scien. early school-leavers, NEET and educational attainment.

Country	DEA	Country	DEA
AT	0.70	IT	0.78
BE	0.76	LT	0.65
BG	0.95	LU	0.63
CY	0.70	LV	0.86
CZ	0.94	NL	1.00
DE	1.00	PL	0.79
EE	1.00	PT	0.63
ES	0.76	RO	1.00
FI	0.93	SE	0.77
FR	0.89	SI	0.87
HU	0.78	SK	1.00
IE	0.93	UK	1.00

Table 5. The DEA efficiency scores by country, Model 4

The ranking presented in Fig. 2 can therefore be considered as a multidimensional measure of country efficiency. Each option is characterised by its strength (positive flow) and weakness (negative flow). The intersection between these two evaluations provides the final ranking. When two options are not connected by an arrow, the situation is described as a so-called incomparability relation,

i.e. according to the information available, no clear relationship of preference or indifference can be derived between these two options. This ranking has been derived by limiting the compensability among criteria as much as possible. As we have already discussed, a low degree of compensability can here be considered a desirable property. However, in the search of the assessment of result robustness, we also test how the final ranking varies if one allows higher degrees of compensability. With the exception of the extreme case where the maximum degree of compensability is used, the country ranking appears very stable.

Overall the following three groups of countries appear:

High Performance = (FI, SI, EE, PL, CZ, DE, NL)

This group contains all countries which present a high performance and do not present any incomparability relation with other countries, their performance is a clear cut one. It should be noted that while Finland looks definitely better than all other countries in this group (strength 0.87 and weakness 0.24), the other ones are all very close (strength between 0.76 and 0.68, weakness between 0.40 and 0.45).

Medium Performance = (BE, LV, SK, AT, LT, HU, DK, LU, RO, IE)

Low Performance = (ES, FR, BG, UK, SE, IT, PT, CY) Both medium and low performance countries are characterised by the presence of incomparability relations; their performance is a less clear cut one.

Medium performance countries have been considered the ones which are in the set fulfilling both conditions maximum weakness equal to 65 and minimum strength bigger than 50. All remain countries are considered low performance ones.

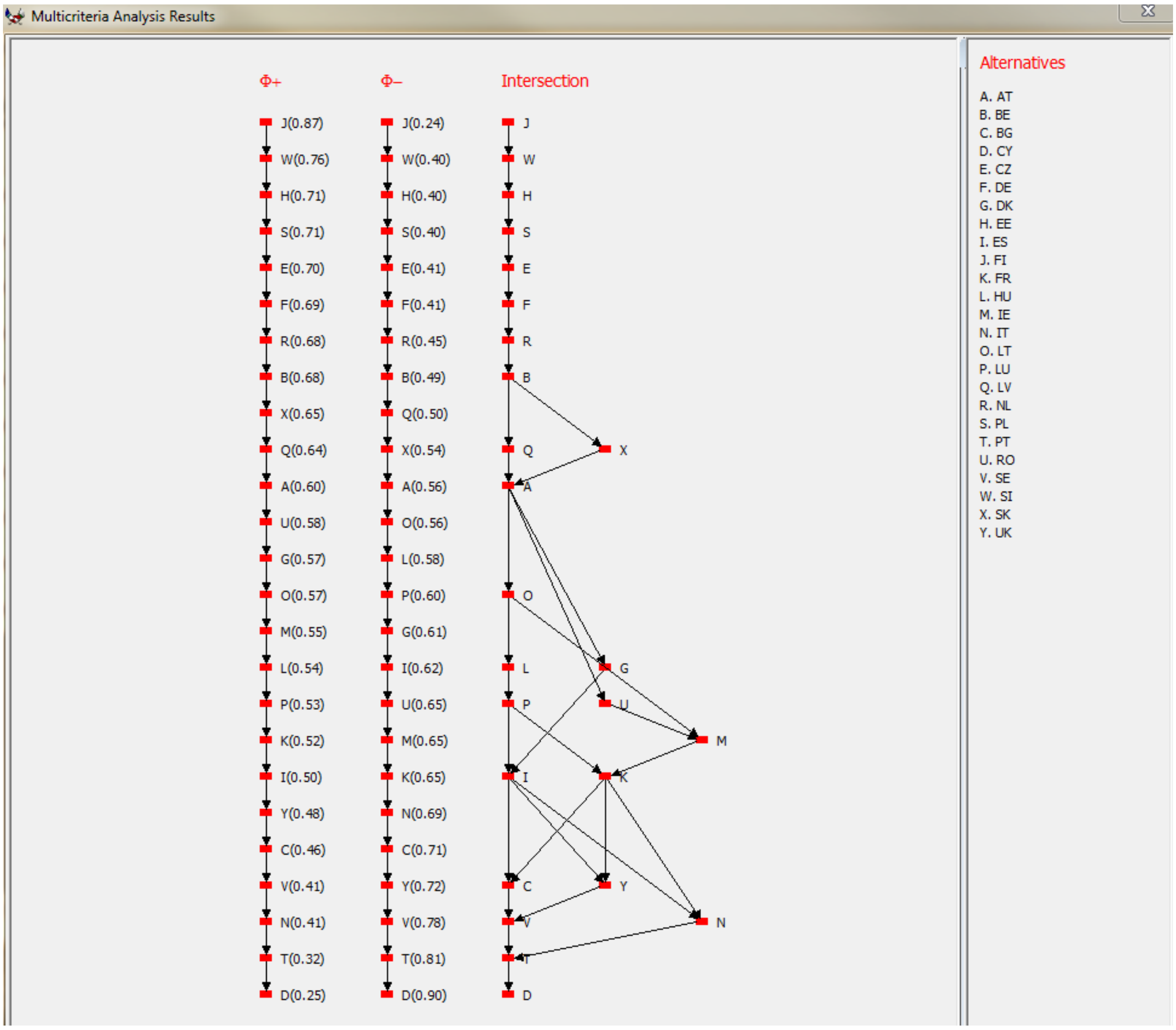

Figure 2. Multi-criteria ranking according to input and output items

Compensability = minimum (Countries in the top positions are the ones presenting better overall efficiency scores)

In the case of maximum compensability, which we reiterate is not a desirable property, countries can be clustered as follows:

High Performance = (FI, SI, SK, CZ, LV, HU, RO)

Medium Performance = (BG, DE, BE, LT, ES, IT, EE, PL)

Low Performance = (NL, AT, LU, DK, IE, SE, PT, UK, CY, FR)

The effects of complete compensability are very evident. While countries such as FI, SI and CZ are still top performers and some others continue to be at the bottom, there are countries that present a completely different performance in comparison, with lower compensatory degrees. For example, NL is now evaluated as a bottom performer while ES and IT become medium efficiency countries. The reason why this happens is that, in a complete compensatory framework, spending inputs (if medium-high) can completely overcompensate good outcomes (this is the case for e.g. NL, AT, LU) or (if medium-low) bad outcomes (e.g. IT or ES). Although multi-criteria analysis is based on completely different methodological assumptions than DEA, overall the results are robust. The only exceptions being BG, which DEA evaluates as a top performer while multi-criteria analysis considers it as a bottom performer, and FR and the UK which DEA evaluates much better than MCA (in one approach, UK is a top country while in the other, it is consistently considered a bottom one). In this case, differences in results between DEA and MCE are due to the following main factors: the degree of compensability that is allowed in the mathematical aggregation (high in the case of DEA and variable from low to high in the case of MCE); DEA being a frontier based evaluation method while NAIADE is not; the information that is used (plurality of data sets in the MCE case); the use of indifference and preference thresholds in MCE while being not used at all in DEA (see the annex for more information on this point).

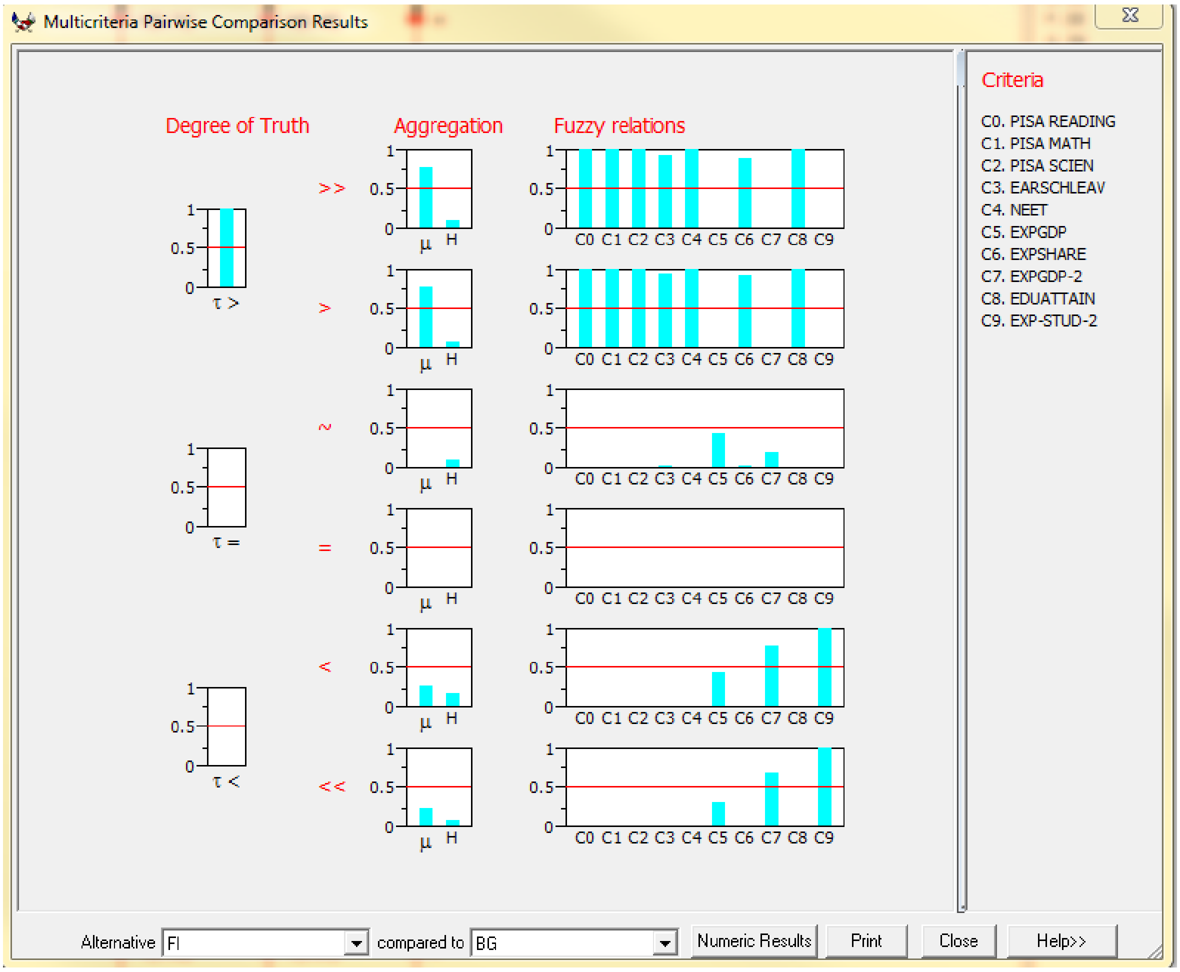

Figure 3. Pair-wise comparison between FI and BG

In order to better understand the multi-criteria performance of these three countries, we compare each of them with the top performer, i.e. Finland (see Figs 3–5). The first two columns of results back up the credibility of the statement that the overall evaluation considers an option better, equal or worse than another one. The third column provides relevant information for the benchmark exercise.

In fact, the credibility of the evaluation is referred to each single evaluation criterion.

In the NAIADE framework, an assessment is considered 'credible' if its 'degree of truth' is higher than 0.5. In all the three cases, the first two columns are corroborating the statement that overall, Finland is a more efficient country than the other ones according to the great majority of criteria. One should note that this benchmarking exercise (enabled by the NAIADE method) can also help policy-makers to derive policy priorities. For example, in the case of the UK, by looking at the performance for each of the single criteria used, it is possible to see that for criteria 8 and 9, the UK performs better than Finland. The UK however definitely performs worse than Finland when considering all the other criteria. This is something which should be considered as a possible policy priority.

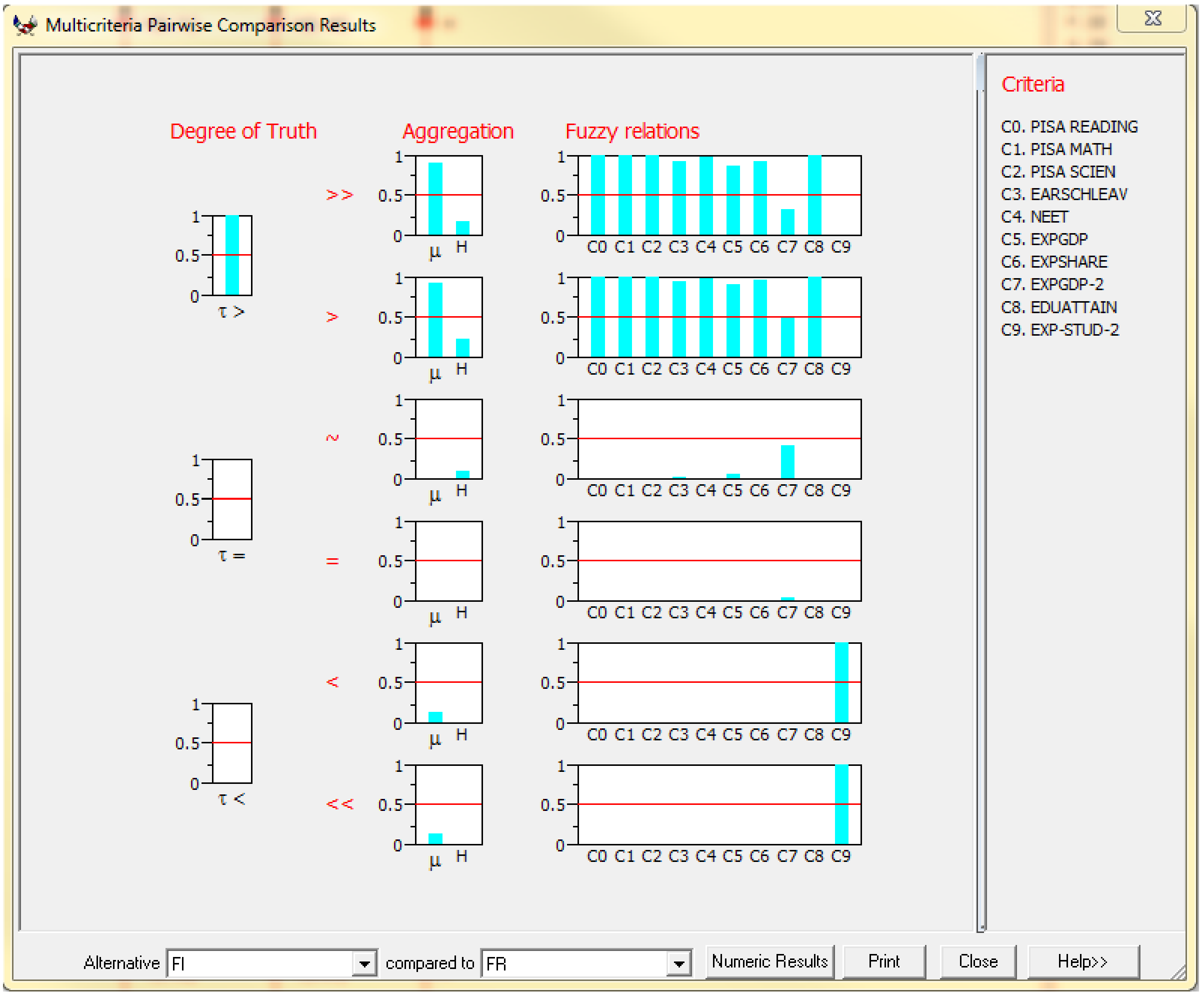

Figure 4. Pair-wise comparison between FI and FR

5 Discussion of key results and concluding remarks

5.1 Interpreting the results: some preliminary insights

Efficiency analyses, as in any other evaluation study, may present a number of risks, e.g. oversimplification, wrong policy conclusions due to model misspecification and biased results caused by hidden subjective judgments in the design process. Uncertainty and sensitivity analyses can gauge the robustness of the results obtained and help the framing of the debate around the conceptual framework used, i.e. which representation of reality has been considered. Efficiency scores should be derived through a plurality of methodological approaches:

- Non-parametric methods allow the statistical impact of contextual variables in production processes and efficiency to be shown. Together with more traditional second-stage regressions and descriptive analyses, these methods should be employed to reveal how efficiency estimates do indeed mask the influence of factors that are beyond the control of educational policy makers (see a list in Table 6).

- A golden standard of any system for measuring efficiency should consist of defining whether results have the following two properties: (i) they are robust in the selection of specific variables (inputs and outputs selection), and (ii) they are robust in the selection of a method for efficiency analysis. Multi-criteria evaluation can be useful for tackling both questions as they allow many variables to be dealt with simultaneously and it is complementary in terms of frontier based methods, such as DEA (e.g. it also considers dominated alternatives, it is non-compensatory and it can use both qualitative and quantitative scores).

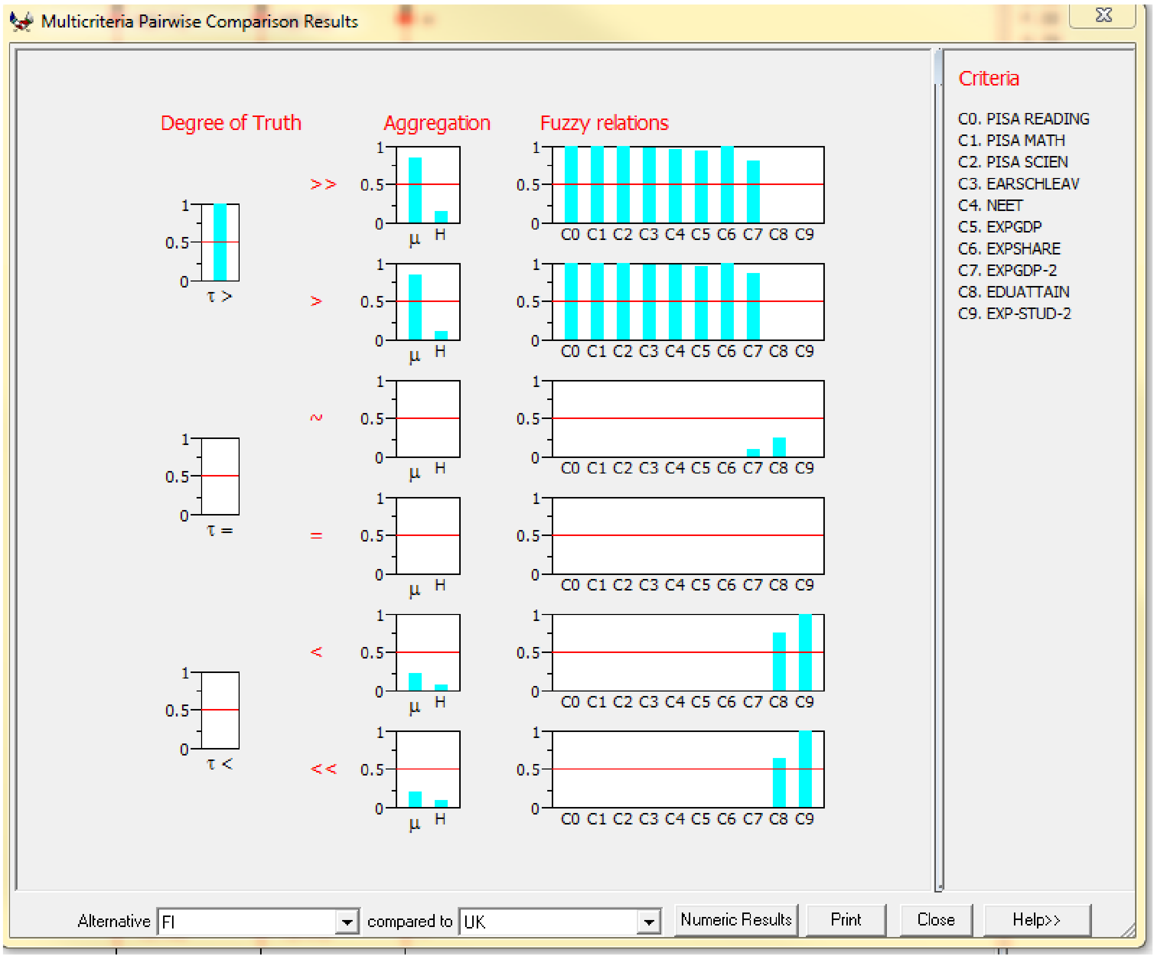

Figure 5. Pair-wise comparison between FI and UK

In the section that follows, we will comment on evidence that stems from a correlational analysis that compares efficiency scores with key contextual variables. Detailed figures and tables are available in Agasisti et al. (2017).

We reiterate that measuring efficiency is different from measuring effectiveness (academic performance measured by test scores and graduation rates). Some countries have students who obtain, on average, excellent results in terms of test scores in mathematics (e.g. PL and BE) whose educational system has been deemed as 'relatively inefficient' in this analysis, because of a high amount of resources that are used in order to get these results – more precisely, other countries in the sample employ lower levels of resources to obtain a comparable level of performance. Conversely, there are countries where the educational system is 'relatively efficient' – i.e. the resources are used at their most, for getting the highest possible level of output – but still the academic results are unsatisfactory, e.g. students' results in PISA are still very low e.g. RO and SK. As a consequence, the interpretation of results must take the definition of 'efficiency' into account. As such, we focus our attention on correlations between contextual variables and DEA efficiency scores and we do not consider the results from NAIADE. As explained above, the latter model considers more a 'global performance' measure than 'traditional efficiency'. We are aware that, from a policy perspective, an educational system is not desirable if it produces low educational outputs, even if it does it efficiently, that is to say with low levels of resources employed in the process. The general aim of any educational policy should consist of reaching an adequate (high) level of outputs using the necessary amount of resources (and no more than this level). As such, instead of reducing the level of inputs, policymakers should find viable strategies for stimulating and getting higher results. Future research should be dedicated to employing output-oriented DEA approaches and understanding how outputs can be expanded while holding inputs constant. Also, future research should look at clusters of countries as emerging from DEA and multicriteria approaches simultaneously and look for features in educational systems that are jointly associated with their efficiency and performance levels. We shall leave these topics for further steps in our research.

Panel A. Descriptive statistics

Variables' label	Variable	Obs	Mean	Std. Dev.	Min	Max
teachers_age	Average age of teachers	15	44.5	2.7	39.2	48.9
teachers_salary	Average salary for teachers	15	29,223.8	18,386.5	9,329.0	80,789.5
%_native	% native students	16	88.8	10.7	53.6	99.3
PIAAC_scores	Adults' test scores in PIAAC	16	267.6	12.2	245.8	282.2
%_adults_TE	% adults (25-64) with a tertiary educ- degree	24	28.1	8.2	14.5	39.0
%_educ_spend	% public spending devoted to education	24	6.6	1.8	4.2	11.2
Stud_background	% of test scores explained by SES	24	16.2	4.4	8.6	24.6

Panel B. Statistical correlations across variables

	Eff	teachers_ age	teachers_ salary	%_non-native	PIAAC_ scores	%_adults _TE	%_ educ_ spend	Stud_b ackgro und
Eff	1.00							
teachers_age	-0.31	1.00						
teachers_salary	-0.41	0.01	1.00					
%_native	0.49	0.03	-0.90	1.00				
PIAAC_scores	0.24	-0.10	-0.05	0.07	1.00			
%_adults_TE	-0.02	-0.07	0.41	-0.43	0.12	1.00		
%_educ_spend	-0.20	-0.17	0.25	-0.36	-0.26	0.51	1.00	
Stud_background	0.33	-0.08	-0.19	0.23	0.31	0.04	0.16	1.00

Table 6. Some key characteristics of the educational systems (contextual factors)

5.2 Correlations between efficiency scores and the characteristics of educational systems

In this section, we present a discussion of the main correlations emerging between efficiency scores (as calculated by means of Data Envelopment Analysis) and some characteristics of the educational systems. In principle, such exploration could be conducted also considering the results emerging from the multi-criteria analysis; this can represent a potential extension of this line of research – thus, we leave the MCA as a robustness check for DEA scores only.

All else being equal, more efficient educational systems are those where the expenditure per student is lower (although some efficient educational systems are characterised by higher-than-average spending). As a policy consequence, higher spending in education should be accompanied by over proportional ability in terms of achieving better academic results. Investing more resources in educational systems is not enough to guarantee better results; an efficient use of those resources is a prerequisite for raising the overall qualitative level of the systems themselves. In this perspective, the better approach for policy-making is not allocating more public and private money to education *tout court*, but instead, searching for those interventions and activities which are more related to better results.

Resources then, should go hand-in-hand with methods for assessing the results achieved; and in the current age of tight budgets, the evaluation of the dynamics of expenditure growth should be based on the benchmarking of results (i.e. by comparing practices and characteristics of educational systems with similar spending levels and different levels of outputs produced). In this sense, our findings are consistent with evidence provided by existing literature that demonstrates that the specific use of resources matters in affecting educational outputs, more than the absolute level of spending (see a theoretical model about the different cost-effectiveness of various educational interventions in Pritchett and Filmer 1999).

In the same vein, there is a positive correlation between the student-teacher ratio and the estimated efficiency of an educational system (around 0.74, results available from the authors). This finding is suggestive of the negative relationship between the amount of (human) resources available and efficiency (holding the output constant). The policy interpretation of this finding is far from being straightforward, however. For sure, educational policy-makers cannot intend this finding as a justification for reducing the teaching labour force, nor can they consider it as a suggestion for increasing average class sizes. Instead, the result only corroborates the intuition that a higher investment for hiring more teachers is not necessarily conducive to higher levels of educational results per se. To pursue this objective, instead, human resources should work in a productive manner. The available data at country level do not allow researchers to take into account quality differentials in the average workforce of teachers, which could be responsible for at least part of the efficiency differentials that emerge across countries. Apart from these average differences, it is also the case that differences within countries do exist and these lead to an overall difference in the efficiency use of human resources. As such, no further strong conclusions can be easily derived unless one explores the distribution of the student-teacher ratio by school, within countries – something which is well beyond the scope of this analysis.

Finally, in Table 6 we report on the descriptive statistics of the variables that we use for providing evidence of correlations between the efficiency scores obtained by means of DEA and some important characteristics of the educational system. The selection of these characteristics depends upon two factors: (i) the findings from existent literature; (ii) data availability. Of course, the relationships between efficiency scores and contextual variables must be interpreted as correlational and not causal. Moreover, given the small number of observations, we do not check statistical significance but only the direction of the correlation (i.e. qualitative correlations). In this perspective, we do not claim any prescriptions supported by robust statistical and econometric approaches, e.g. as pointed out by Walker (2009) – something that is well beyond the scope of this work, given its focus on country level variables – instead, we aim to show important descriptive evidence that can inform future research concerning the determinants of efficiency scores for educational systems. We believe that our findings constitute a first step in a wider research effort that uses and combines additional and diversified econometric and statistical techniques, and also employs larger datasets and more precise variables that will be available over time. Despite the methodological cautions expressed above, from a policy point of view, there are several lessons that can be drawn from our results.

To begin with, educational systems where teachers are younger are, on average, also more efficient. Although no causal claims can be derived from the present study, this statistical correlation is helpful for stimulating institutional reflections at a country level; in fact two different explanations may exist. The first one is that a younger body of teachers can help the efficient use of available resources. For instance, younger teachers may be more motivated or equipped with the most 'updated' skills, and they may employ resources in a more efficient way so that the number of teachers that is required to obtain the expected outputs may be lower than in the past. If this is the case, policy makers should reflect on three important steps that deal with policies about training human resources appropriately. Firstly, plans and incentives to hire younger teachers, and which favour a turnover between older teachers and younger ones, could be welcome and justified under an efficiency argument. Secondly, studies and analyses concerning the differences in the use of time and resources by teachers of different ages could be helpful in informing about more efficient practices and habits. Thirdly, salaries and economic incentives could be directed towards the remuneration of more productive young teachers; indeed, reducing the amount of money attributed on the basis of teachers' age and experience, and improving funds allocated to more productive efficiency, can result in improving the efficiency of an educational system, all else being equal. However, in the literature, some studies suggest that older and more experienced teachers may lead to the higher performance of students; although this is not a strictly linear relationship, as the very oldest seem to perform again on a lower level. A possible explanation is that in some countries, older teachers have more power to select which class they take, so that they may opt for the classes with the better (and easier) students. The second explanation is that younger teachers are simply paid less; all else equal (i.e. assuming equal productivity along the curve of teachers' age), this would result in lower expenditures per student, in front of the same academic results. Data support a negative relationship between a country's average salary paid to teachers, and its educational system's efficiency score. In terms of policy implications, this would result in completely different suggestions than those formulated above. At system level, it would not be that desirable to keep the salaries artificially low (i.e. below the productivity ratio) just for pursuing a more efficient process of educational production. Albeit this approach can have a positive payoff in the short-run, when the average levels of educational outputs do not experience negative shocks, it can instead be the case that negative effects appear in the medium-long run. For example, the teachers could be demotivated by persistently too low salaries and their productivity could drop year after year – eventually, the best part of them could even leave the profession. Furthermore, low salaries can create a barrier preventing the attraction and retention of promising young talents and productive teachers, undermining the positive effects for system-level educational efficiency associated with having a body of teachers more skewed towards the younger ones. On the other side of the equation, therefore, policy makers of those countries where teachers (even the younger ones) are paid more, should assess which initiatives must be taken to stimulate a more productive use of teaching human resources, including the opportunity of lifelong training and the consideration of more performance-oriented incentives and salaries (although the academic literature in these areas is quite sceptical about the real effectiveness of these approaches).

Countries where a higher proportion of students are natives (i.e., non-immigrants) have, on average, a more efficient educational system. This negative correlation between the proportion of immigrants and the educational system's efficiency can be explained by a higher cost associated with educating non-natives, because of obstacles related to culture, background, history and practices – overcoming these obstacles is, by the way, an explicit aim of various educational policies in Europe. As often happens, this is a clear case where efficiency and equity objectives are potentially in conflict. This finding at a country level comes by stimulating important policy implications. The political aim of promoting social inclusion for students who are not natives is strong enough to justify some inefficiency stemming from having a high proportion of immigrant students, as this will also help future benefits in terms of overall productivity. However, policy makers and educational leaders should challenge themselves in identifying institutional approaches and educational practices which can maximise the productivity of the resources invested for the social (and educational) inclusion of immigrant students. For example, it may be the case that the inefficiency of the educational systems is derived from allocating money to some practices and initiatives which do not stimulate the educational experience of immigrant students, and their subsequent educational performance. In the same vein, different initiatives could be developed to guarantee the current educational performance of immigrant students with less money invested. Overall, given that policy makers opted for investing resources directed at including immigrant students, educational experts and teaching leaders should devote their efforts towards using the available additional money for raising the educational performances of these students. The analyses based on the efficiency approach, in this context, could provide a good guide in order to check whether the available resources (that are invested together, for reasons that go beyond any efficiency consideration and instead follow priorities based around equity) are used in the most productive directions.

The efficiency of an educational system is positively correlated with the competences of the adults, at a country level, as measured through the OECD program known as PIAAC - Programme for the International Assessment of Adult Competencies. Stimulating the efficiency of an educational system is not only worthwhile because of the responsibility in the best possible use of public resources; it also contributes to long-term benefits in terms of adults' skills and competences (on the other hand adults with better skills and competences contribute to the education system's overall efficiency by providing a better family background to students). This finding also constitutes a justification for once again bringing attention to policies for improving the lifelong learning (LL) in various countries. The traditional argument for LL is that adults need continuous updating of their competences to fully live in a society and economy that continuously requires new skills – and this is certainly true. However, our study points at the existence of a 'spillover' effect. Indeed, adults who are better educated are able to work in a more productive way towards the creation of positive conditions that favour a more productive use of educational resources which, in turn, leads to better educational current outputs and impacts positively on the subsequent efficiency of the educational system.

The amount of public spending on education, as a proportion of total public spending, is not statistically correlated with the efficiency of an education system in a country. However, many efficient educational systems are in countries where the proportion of public spending on education is relatively low. As mentioned above, our findings suggest that probably the way of employing the available resources matters more in terms of efficiency than the prioritisation of education in public spending. At the same time, policy considerations should also consider that there are countries where educational spending is high, and the efficiency of the system is also high. These are countries in which high levels of spending did result in higher-than-average performance and which are able to secure future prosperity in the country through their high level of educational outputs. These countries should represent the desirable benchmark, more than those countries which are only deemed as efficient because they spend relatively little on education (without getting educational performance that is high enough). In the light of policy-making, educational leaders should aim at understanding how the high-spending, high-results employ their resources so that new funding requests can be channelled through initiatives for improving the efficiency of educational systems overall.

5.3 Concluding remarks

Equity and efficiency are not necessarily set in a trade-off setting. Although we reveal that, in some countries, high efficiency comes at the cost of lower equality, there are also cases where the two dimensions (efficiency and equity) go hand-in-hand. In this sense, it is likely that the internal characteristics of the educational system's design can affect the relative importance of equity and efficiency and the potential for their co-existence. From a policy perspective, optimal conditions arise when higher levels of educational output can be obtained without sacrificing the ability of disadvantaged students to obtain good results. Such conditions can be achieved when the average results are not only satisfactory, but the distribution around the average results is also quite narrow; indeed, in these circumstances, the majority of students are able to achieve high educational outputs. Policy makers should seek to create conditions which use resources in the most productive way. Reducing the spending while keeping outputs constant is a risky approach, as it is likely to reproduce social inequalities between students with different socioeconomic backgrounds. In this sense, resources should be allocated to helping disadvantaged students close potential educational gaps. For example, such an approach could work simultaneously in improving the efficiency and equity of the educational system. In other words, the results of our efficiency analyses suggest that policy makers could keep an eye on raising average educational results while, at the same time, not leaving disadvantaged students behind.

Finally, we would like to stress the importance of data for proper efficiency analyses. Although we made all efforts to use reliable and complete data, there is no doubt that data availability could be improved a lot. Here, efforts conducted by international authorities and organisations (such as Eurostat and OECD) will be necessary and are more than welcome.

Acknowledgements

This article builds on previous work carried out for the JRC's Centre for Research on Education and Lifelong Learning (CRELL). The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission. We are extremely grateful to the Editor and to two anonymous reviewers for their detailed, challenging and precise comments/suggestions which substantially helped us improve a previous draft of the paper.

References

Agasisti T (2011) Performances and spending efficiency in higher education: a European comparison through non-parametric approaches. Educ Econ 19(2):199–224

Agasisti T, Hippe R, Munda G (2017) Efficiency of investment in compulsory education: empirical analyses in Europe; EUR 28607 EN. : Publications Office of the European Union, Luxembourg (Luxembourg), p JRC106678. https://doi.org/10.2760/975369

Alexander WRJ, Haug AA, Jaforullah M (2010) A two-stage double-bootstrap data envelopment analysis of efficiency differences of New Zealand secondary schools. J Prod Anal 34(2):99–110

Antunes CH, Alves MJ, Clímaco J (2016) Multi-objective Linear and Integer Programming. Springer, New York, NY

Arrow KJ, Raynaud H (1986) Social choice and multicriterion decision making. M.I.T. Press, Cambridge

Barro R. J (2001) Human capital and growth. American Economic Review 91(2):12–17

Benhabib J, Spiegel MM (1994) The role of human capital in economic development evidence from aggregate cross-country data. J Monetary Econ 34(2):143–173

Bouyssou D (1999) Using DEA as a tool for MCDM: some remarks. J Oper Res Soc 50(9):974–978

Cazals C, Florens JP, Simar L (2002) Nonparametric frontier estimation: a robust approach. J Econ 106(1):1–25

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Charnes A, Cooper WW, Lewin AY, Seiford LM (eds.) (2013) Data envelopment analysis: theory, methodology, and applications. Springer Science & Business Media, New York

Coelli TJ, Rao DSP, O'Donnell, CJ, Battese GE (2005) An introduction to efficiency and productivity analysis. Springer Science & Business Media, New York

Coleman JS, Campbell EQ, Hobson CJ, McPartland J, Mood AM, Weinfeld FD, York R (1966) Equality of educational opportunity. U.S. Government Printing Office, Washington, DC, 1066–5684

Cooper WW, Seiford LM, Zhu J (Eds.) (2011) Handbook on data envelopment analysis (Vol. 164). Springer Science & Business Media New York

Cordero-Ferrera JM, Pedraja-Chaparro F, Salinas-Jiménez J (2008) Measuring efficiency in education: an analysis of different approaches for incorporating non-discretionary inputs. Appl Econ 40(10):1323–1339

Cunha M, Rocha V (2012) On the efficiency of Public higher education institutions in Portugal: an exploratory study. FEP Economics and Management, Working Paper

Daraio C, Simar L (2007) Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. J Prod Anal 28(1–2):13–32

Darling-Hammond L (2000) Teacher quality and student achievement. Education Policy Analysis Archives 8:1

De Witte K, López-Torres L (2017) Efficiency in education: a review of literature and a way forward. J Oper Res Soc 68(4):339–363

Dill DD, Soo M (2005) Academic quality, league tables, and public policy: a cross-national analysis of university ranking systems. Higher Education 49(4):495–533

Doyle J, Green R (1993) Data envelopment analysis and multiple criteria decision making. Omega 21(6):713–715

Dronkers J, Robert P (2008) Differences in scholastic achievement of public, private government-dependent, and private independent schools: a cross-national analysis. Educ Policy 22(4):541–577

Emrouznejad A, Yang GL (2018) A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. Socio-Economic Planning Sciences 61:4–8

Färe R, Karagiannis G (2017) The denominator rule for share-weighting aggregation. Eur J Oper Res 260(3):1175–1180

Figueira J, Greco S, Ehrgott M (eds.) (2016) Multiple-criteria decision analysis. State of the art surveys. Springer International Series in Operations Research and Management Science, New York, NY Fried HO, Lovell CK, Schmidt SS (Eds.) (2008) The measurement of productive efficiency and productivity growth. Oxford University Press, Oxford, UK

Golany B, Tamir E (1995) Evaluating efficiency-effectiveness-equality trade-offs: a data envelopment analysis approach. Manag Sci 41 (7):1172–1184

Greenwald R, Hedges LV, Laine RD (1996) The effect of school resources on student achievement. Rev Educ Res 66(3):361–396 Grosskopf S, Moutray C (2001) Evaluating performance in Chicago public high schools in the wake of decentralization. Econ Educ Rev 20(1):1–14

Grosskopf S, Hayes KJ, Taylor LL (2014) Efficiency in education: research and implications. Appl Econ Persp Policy 36 (2):175–210

Guskey TR (2007) Multiple sources of evidence: an analysis of stakeholders' perceptions of various indicators of student learning. Educ Measurement: Issues and Practice 26(1):19–27

Haelermans C, De Witte K (2012) The role of innovations in secondary school performance–Evidence from a conditional efficiency model. Eur J Oper Res 223(2):541–549

Hanushek EA, Luque JA (2003) Efficiency and equity in schools around the world. Econ Educ Rev 22(5):481–502

Hanushek EA, Woessmann L (2008) The role of cognitive skills in economic development. J Econ Lit 46(3):607–668

Hanushek EA, Woessmann, L (2010) The economics of international differences in educational achievement (No. w15949). National Bureau of Economic Research

Hanushek EA, Woessmann L (2012) Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. J Econ Growth 17(4):267–321

Haveman R, Wolfe B (1995) A review of methods and findings. J Econ Lit 33(4):1829–1878

Ho W, Dey PK, Higson HE (2006) Multiple criteria decision-making techniques in higher education. Int J Educ Manag 20(5):319–337

Johnes J (2004) Efficiency measurement. International Handbook on the Economics of Education, chapter 16

Joro T, Pekka K, Jyrki W (1998) Structural comparison of data envelopment analysis and multiple objective linear programming. Manag Sci 44(7):962–970

Keeney R, Raiffa H (1976) Decisions with multiple objectives: preferences and value trade-offs. Wiley, New York, NY

Li X-B, Reeves GR (1999) A multiple criteria approach to data envelopment analysis. Eur J Oper Res 115(3):507–517

Liu JS, Lu LY, Lu WM, Lin BJ (2013) A survey of DEA applications. Omega 41(5):893–902

Madlener R, Antunes CH, Dias LC (2009) Assessing the performance of biogas plants with multi-criteria and data envelopment analysis. Eur J Oper Res 197(3):1084–1094

Malen B, Knapp M (1997) Rethinking the multiple perspectives approach to education policy analysis: implications for policy-practice connections. J Educ Policy 12(5):419–445

Mancebón MJ, Calero J, Choi Á, Ximénez-de-Embún DP(2012) The efficiency of public and publicly subsidized high schools in Spain: Evidence from PISA-2006 J Opera Res Soc 63 (11):1516–1533

Moore, M. H. (1995). Creating public value: Strategic management in government. Harvard Business University Press, Cambridge, MA Munda G (1995) Multicriteria evaluation in a fuzzy environment. Physica-Verlag, Contributions to Economics Series, Heidelberg

Munda G (2008) Social multi-criteria evaluation for a sustainable economy. Springer, Heidelberg, New York, NY

Munda G (2012) Intensity of preference and related uncertainty in non-compensatory aggregation rules. Theory and Decision 73 (4):649–669

Nikel J, Lowe J (2010) Talking of fabric: A multi-dimensional model of quality in education. Compare 40(5):589–605

Nijkamp P, Rietveld P, Voogd H (1990) Multicriteria Evaluation in Physical Planning. North-Holland, Amsterdam

OECD (2014a) PISA 2012 Results in focus. What 15-year-olds know and what they can do with what they know. OECD, Paris

OECD (2014b) PISA 2012 Technical report. OECD, Paris

Perry L, McConney A (2010) Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. The Teachers College Record 112(4):7–8

Pritchett L, Filmer D (1999) What education production functions really show: a positive theory of education expenditures. Econ Educ Rev 18(2):223–239

Romer PM (1990). Human capital and growth: theory and evidence. In Carnegie-Rochester Conference Series on Public Policy (Vol. 32, pp. 251–286). North-Holland

Rossell CH (1993) Using multiple criteria to evaluate public policies: The case of school desegregation. American Politics Quarterly 21 (2):155–184

Roy B (1996) Multicriteria methodology for decision analysis. Kluwer, Dordrecht

Sacerdote B (2011) Peer effects in education: How might they work, how big are they and how much do we know thus far? Handbook of the Economics of Education 3:249–277

Sarrico CS, Rosa MJ, Coelho IP (2010) The performance of Portuguese secondary schools: an exploratory study. Quality Assurance Educ 18(4):286–303

Simar L, Wilson PW (2000) A general methodology for bootstrapping in non-parametric frontier models. J Appl Stat 27(6):779–802

Simar L, Wilson PW (2011) Two-stage DEA: caveat emptor. J Prod Anal 36(2):205

Steuer R (1986) Multiple criteria optimization; theory, computation, and application, Wiley Series in Probability and Mathematical Statistics – Applied, Wiley

Stufflebeam D (2001) Evaluation models. New Directions for Evaluation Issue 89:7–98

Sutherland D, Price R, Gonand F (2010) Improving public spending efficiency in primary and secondary education. OECD J: Econ Stud 2009(1):1–30

Tzeng GH, Cheng-Hsin C, Chung-Wei L (2007) Evaluating intertwined effects in e-learning programs: a novel hybrid MCDM model based on factor analysis and DEMATEL. Expert Systems with Applications 32:1028–1044

Walker WE (2009) Does the best practice of rational-style model-based policy analysis already include ethical considerations? Omega 37(6):1051–1062

Worthington AC (2001) An empirical survey of frontier efficiency measurement techniques in education. Educ Econ 9(3):245–268

Wu M (2005) The role of plausible values in large-scale surveys. Studies in Educational. Evaluation 31(2-3):114–128

Zhu, J (ed.) (2015) Data envelopment analysis: A handbook of models and methods (Vol. 221). Springer, New York