KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and
Applied Economics

Department of Economics

Equity and efficiency in private and public education: a nonparametric comparison

by

Laurens CHERCHYE
Kristof DE WITTE
Erwin OOGHE

Public Economics

Center for Economic Studies
Discussions Paper Series (DPS) 07.25
http://www.econ.kuleuven.be/ces/discussionpapers/default.htm

September 2007

DISCUSSION
PAPER

Efficiency and equity in private and public education: a nonparametric comparison[*]

Laurens Cherchye [†] Kristof De Witte [‡] Erwin Ooghe [§]

Ides Nicaise [¶]

April 22, 2008

Abstract

We present a nonparametric approach for (1) efficiency and (2) equity evaluation in education. Firstly, we use a nonparametric (Data Envelopment Analysis) model that is specially tailored to assess educational efficiency at the pupil level. The model accounts for the fact that typically minimal prior structure is available for the behavior (objectives and feasibility set) under evaluation. It allows for uncertainty in the data, while it corrects for exogenous 'environmental' characteristics that are specific to each pupil. Secondly, we propose two multidimensional stochastic dominance criteria as naturally complementary aggregation criteria for comparing the performance of different school types (private and public schools); these criteria are specifically designed for aggregating

[*] We would like to thank Jan Van Damme for his permission to use the SiBO-data, Frederik Maes and Peter Helsen for their very useful help with the data, and the educational department of the Flemish government for the provision of some supplementary data. In addition, we want to thank Bart Capéau, André Decoster, Erik Schokkaert, Gary Simpson, Wim Moesen, Emmanuel Thanassoulis and other seminar participants in Edinburgh, Aston and Leuven University for insightful discussions.

[†] Centre for Economic Studies, University of Leuven (KU Leuven), Naamsestraat 69, 3000 Leuven, Belgium; post-doctoral Fellow of the Fund for Scientific Research - Flanders; E-mail to Laurens.Cherchye@kuleuven-kortrijk.be.

[‡] Centre for Economic Studies, University of Leuven (KU Leuven), Naamsestraat 69, 3000 Leuven, Belgium; Email to Kristof.Dewitte@econ.kuleuven.be.

[§] European University College, Stormstraat 2, 1000 Brussels, Belgium, and Centre for Economic Studies, University of Leuven (KU Leuven), Naamsestraat 69, 3000 Leuven, Belgium; E-mail to Erwin.Ooghe@econ.kuleuven.be.

[¶] Higher Institute of Labour Studies, University of Leuven (KU Leuven), Parkstraat 47, 3000 Leuven, Belgium; Email to Ides.Nicaise@hiva.kuleuven.be.

pupils' output performance while adjusting for environment-corrected inefficiency. While the first criterion only accounts for efficiency, the second criterion also takes equity into consideration. The model is applied for comparing private (but publicly funded) and public primary schools in Flanders. Our application finds that no school type robustly dominates another type when controlling for the school environment and taking equity into account. More generally, it demonstrates the usefulness of our nonparametric approach, which includes environmental and equity considerations, for obtaining 'fair' performance comparisons in the public sector context.

Keywords: equity; efficiency; private versus public education; nonparametric analysis; Data Envelopment Analysis; stochastic dominance

1 Introduction

An important theme in policy evaluation is whether public funds are used in an efficient and equitable way. In the specific context of education, the comparison between private —but, possibly, publicly funded— schools and public schools is at the heart of a debate, which started with the work of Coleman *et al.* (1982). They find that (1) catholic school students achieve higher standardized test scores than public school students (while controlling for family background); and (2) this is particularly the case for minority students. Therefore, one could conclude that catholic schools were both more efficient and more equitable than public schools in the U.S. at that time. The work of Coleman *et al.* was (and still is) controversial, not only in the public debate (see, e.g., the New York Times articles of April 7, April 12 and April 26, 1981, discussing the consequences of Coleman *et al.*'s results for the introduction of tuition tax credits and/or school vouchers), but also in academics (see, e.g., Cain and Goldberger, 1983, for an overview). In spite of these criticisms, many studies have confirmed the outperformance of public by private schools; see, e.g., the literature overview in Altonji *et al.* (2005a).

Most studies consider differences between public and private schools in the US on the basis of parametric regression techniques. Toma (1996) argues that further investigation using different case-studies and/or different methodologies is needed. First, the US educational system is rather different compared to, e.g., European systems. Private schools in the US historically represent only a small percentage of all pupils. The relative outperformance could vanish if the private educational system operates at a different scale. In addtion, it is not clear to what extent the outperformance of public by private schools in the US has to be attributed to the different ideological vision on education or rather to the different ways these schools are funded. Although Hanushek and Raymond (2005) do not focus on public-private differentials, they show that the different funding systems in the different US states lead to different outcomes. In this study, we compare private (catholic) and public primary schools in Flanders, i.e., a region in Belgium, where the number of pupils in both school types is roughly equal and all schools are fully funded by the government. Second, most studies are parametric and therefore it might be worthwile to complement this analysis with other techniques, e.g., nonparametric techniques.[1] In this study, the methodology consists of two steps, a measurement step (estimating the education production function at the individual level) and an aggregation step (aggregating the actual and potential outcomes for each school type).[2] Both steps are non-parametric, i.e., we impose little a priori structure on the measurement step (to minimize specification error when estimating production functions) and on the aggregation step (to minimize value-laden statements when assessing outcomes).

To set the scene, we briefly present the measurement and aggregation step in more detail and relate them to the existing literature. We use a nonparametric DEA model to measure educational efficiency at the pupil level on the basis of test scores in mathematics and language proficiency (writing and reading in Dutch). We account for the inputs used (which the policy makers do control) as well as for possibly diverging 'environmental' variables —socio-economic status of parents and lagged test score results— that might affect pupil performance (and which often fall beyond the control of policy makers and schools). The environmental variables control for selection issues, assumed to be based on observables only; see, e.g.,

[1] Relatively few studies have compared parametric and non-parametric estimation techniques in education; see, e.g., De Witte *et al.* (2008).

[2] The potential outcome is the actual outcome plus the inefficiency.

Altonji *et al.* (2005b) for selection on unobservables in parametric estimation of education production.

DEA models have been used before to evaluate the educational efficiency at the pupil level; see, e.g., Grosskopf *et al.* (1997, 1999), Portela and Thanassoulis (2001), and the references therein. In the current study, we propose a DEA model that is specially designed for educational efficiency evaluation: while at the input side it uses the minimal 'free disposability' assumption (*in casu*, more input never leads to a lower (potential) performance), at the output side it uses the linear aggregation that is typical for measuring pupil performance in primary education (i.e., aggregate performance results are conventionally defined as weighted sums of the results in separate disciplines). Focusing on linearly aggregated output, it measures educational inefficiency in terms of the difference between the maximally attainable output and the actually achieved output.

Three additional features of our DEA model are worth mentioning. Firstly, it uses linear output aggregation, but it allows for flexible weighting of the different performance dimensions. Essentially, such a flexible weighting allows each pupil to be evaluated in terms of his/her own 'most favorable' weighting scheme, which accounts for 'specialization' in education. At the same time, we avoid undesirable 'extreme' specialization by limiting the range of possible output weights through pre-specified bounds. Secondly, by suitably adapting the methodology of Daraio and Simar (2005, 2007) to our DEA model, it can account for outlier behavior, while it also allows us to explain observed performance differences in terms of diverging environmental characteristics in a nonparametric way. The observed environmental impact as well as the corresponding environment-corrected efficiency results provide an easy-to-implement tool for attention-direction in the political process. Thirdly, economies of scale XXX

To compare the aggregate performance of public and private schools, we suggest two multidimensional stochastic dominance criteria that were introduced by Atkinson and Bourguignon (1982). In this view, the overall performance of a school is defined by the sum of individual performances of their pupils. The individual performance of a pupil is in turn a function of the pupil's actual and potential outcome (i.e., the actual outcome plus the (estimated) inefficiency). Rather than imposing a specific functional form for the individual performance function, Atkinson and Bourguignon (1982) focus on a wide class of functions which all satisfy principles with which everyone can agree; e.g., the individual performance functions increase with the pupil's actual output, ceteris paribus. The resulting dominance criterion tells us that one school (type) outperforms another if the sum of individual performances is higher at the former, for all performance functions satisfying the general principles. It is thus robust (everyone agrees with the resulting statement), but comes at the cost of incompleteness (a school (type) could be better according to some, but worse according to other individual performance functions and is therefore classified as incomparable). We believe these aggregation criteria are particularly useful in the context of DEA efficiency evaluation of the public sector. First, they are nonparametric in nature, which naturally complies with the nonparametric orientation of DEA. Next, it is possible to incorporate a concern for equity (i.e., 'higher efficiency is especially better for pupils with lower test scores'), which is particularly relevant within the context of public policy evaluation. As with DEA, these aggregation criteria are easy-to-implement, which makes them attractive for practical applications.

The remainder of this paper unfolds as follows. The next section presents our research question. Section 3 discusses our methodology for evaluating educational efficiency at the individual pupil level. Section 4 presents the efficiency results, with a main focus on environmental effects. Section 5 discusses the aggregation of the individual efficiencies. A final section 6 summarizes our main conclusions.

2 Data and motivation

In general, the literature finds a positive impact for private schools as they are creating a higher added value given the characteristics of the pupils. However, past research generally evaluates private schools which are privately funded (Toma, 1996). As such, the effect of private schools could be attributed to both its way of funding (and the incentives which result from it), its pupil characteristics (due to the nonrandom selection of pupils) and its potentially more efficient educational system (thanks to ideological background). By considering a specific example, i.e. schooling in the Flemish region, we consider only publicly funded schools which allows us to fully concentrate on the ideological background of the school. Indeed, in Flanders (and in Belgium in general) both public and private schools are publicly funded by the regional government and receive full taxpayer funding since 1914. In return, schools should conform to education programs and regulation which regulates the subjects taught and the language used. The methods for instruction are left to the schools (for an extensive discussion, see Toma, 1996).

Whereas the private schools are mainly catholic in Flanders, public schools both originate from the local and the central level, depending on the level of government initiation. We label the latter, respectively, as local and Flemish public schools. In line with the literature, the general belief in Flanders is that private schools perform better (i.e., the cognitive output of their pupils is thought to be higher on average). However, this statement is somewhat blurred by two counteracting forces related to inputs and environment. While private schools are said to have more pupils with an 'advantageous' family background, they should also receive less funding from the 'Equal Educational Opportunities' programme of the Flemish government (see infra). In the remaining of this section, we will define and describe the inputs, outputs and environment in the Flemish educational system, which, at the same time, will confirm the above belief.

We use data from the SiBO-project. The aim of SiBO is to describe and explain differences in the primary school curriculum of a cohort of Flemish pupils. The dataset is oversampled as it consists of a reference group, which is representative for the Flemish population of primary school pupils, and three additional data sets which allow us to capture specific features of the Flemish region. The oversampling is due to (1) including all public city schools of the city of Ghent, (2) an oversampling to get a sufficient number of schools with a high number of disadvantaged pupils (pupils for whom the schools get additional means in the so-called 'Equal Educational Opportunities' programme of the Flemish government) and (3) an oversampling to obtain a sufficient number of non-traditional schools. Although we use all pupils together, we correct for the sample's non-representative nature in our empirical efficiency evaluation. This leaves us with 3413 pupils (with complete data), of whom 1774 attend private catholic schools, 1039 local public schools and 553 Flemish public schools. The remaining 47 pupils take classes in private non-catholic schools. Because we will not include these pupils in our comparison of school types, the term private schools stands for catholic private schools in the sequel.

We look at the cohort of pupils in their second year of primary education (2004-2005) —at the (normal) age of 7— while we use data from the same pupils in the first year (2003-2004) to retrieve environmental variables. We extract 3 types of variables at the individual level, called inputs, outputs and environmental variables in the sequel.

Financial inputs in primary schools mainly consist of salaries (80%) and operational costs (20%). As we *a priori* assume that the differences in operational costs are unlikely to cause differences in cognitive results, we only focus on inputs related to teaching. Government assigns instruction units to pupils, which can be freely used by their respective schools to finance teachers: 24 instruction units correspond with a full-time teacher. We note that, in practice, the input allocated by the Flemish government to a particular pupil should not perfectly correlate with the input allocated by the school to the same pupil, which implies a possible cause of measurement error. (Our application uses outlier robust inefficiency measures to mitigate this measurement error.) The total number of instruction units assigned to a particular pupil consists of regular (REG) and additional, so-called 'equal educational opportunity' (EEO), instruction units. Regular (per-capita) instruction units are, roughly speaking, the same for all pupils, as they are divided among schools on the basis of a scale which is approximately linear in the number of pupils. The additional EEO instruction units depend on certain 'disadvantageous' pupil characteristics, to wit, the household income consists of replacement incomes only, the pupil is living outside the biological family, the level of education of the mother is low, the pupil's family belongs to a travelling population and —in combination with one of the former characteristics— the home language is different from

Dutch. Table 1 contains some summary statistics for both types of instruction units *REG* and *EEO* over the different school types in Flanders.[3] Overall, local public schools receive most instruction units (per capita), private schools the least, while the Flemish public schools are in between both.

Table 1: (Input) *REG* and *EEO* instruction units per school type.

	school type	all	private	public	
input				local	Flemish
all	average	1.00	0.97	1.07	1.04
	std. dev.	0.28	0.26	0.30	0.28
REG	average	0.88	0.87	0.92	0.86
	std. dev.	0.18	0.18	0.19	0.15
EEO	average	0.12	0.09	0.15	0.19
	std. dev.	0.19	0.17	0.20	0.22

Output is defined on the basis of test scores in three dimensions: mathematics, technical reading and writing, collected at the end of the second year. All scores are set between 0 and 100. We calculate a language proficiency score as the simple average of the reading and writing scores. Table 2 provides summary statistics for the mathematics (*MATH*) and language proficiency score (*DUTCH*) for the different school types in Flanders. Private (catholic) schools do best in both tests. They are followed closely by the local public schools and, at some distance, by the Flemish public schools.

Table 2: (Output) *MATH* and *DUTCH* per school type.

[3] All reported figures in this paper are weighted in proportion to the inverse of the sampling probability, to correct for the non-representative nature of the dataset.

	school type	all	private	public	
output				local	Flemish
MATH	average	57.08	58.33	57.54	50.74
	std. dev.	19.40	18.95	19.02	20.78
DUTCH	average	55.27	56.49	54.00	51.56
	std. dev.	14.05	13.46	14.19	15.71

Pupil environment is measured by three indices: socio-economic status and entry level in mathematics and language proficiency. First, socio-economic status (*SES*) reflects the cultural, social and economic environment of the pupil's home. We use data that are calculated as the normalized average of the following three variables: average education level (5 categories: 1 to 5), average professional status (7 categories: 1 to 7) and total income of the parents of the pupil (6 categories: 1 to 6); normalization implies that, for each observation, the difference (of the average) with the sample mean is divided by the sample standard deviation. As for the full sample, normalized values vary between -2.41 (minimum) and 2.63 (maximum); see Reynders *et al.* (2005) for further details. Next, the starting level in mathematics (*B-MATH*) and language proficiency in Dutch (*B-DUTCH*) reflect the intellectual antecedents of the pupil, and is equal to the mathematics and language proficiency score of the pupil at the end of the previous year. As with *MATH* and *DUTCH*, these scores are set between 0 and 100.

Table 3 reports summary statistics for *SES*, *B-MATH* and *B-DUTCH*. We find that, on average, private (catholic) schools attract pupils with more 'advantageous' environmental characteristics compared to local public schools and —to an even greater extent— Flemish public schools. Notice that the differences in *EEO* instruction units between the different school types (reported in Table 1) reflect the differences in *SES*.

Table 3: (Environment) *SES*, *B-MATH* and *B-DUTCH* per school type.

	school type	all	private	public	
environment				local	Flemish
SES	average	0.03	0.11	-0.01	-0.35
	std. dev.	0.85	0.83	0.85	0.83
B-MATH	average	53.26	53.91	54.05	48.44
	std. dev.	19.06	18.66	18.75	20.69
B-DUTCH	average	47.25	47.92	47.29	43.92
	std. dev.	9.74	9.69	9.42	9.87

To summarize, our data roughly confirm the widely held belief that private (catholic) schools in Flanders perform better, while they receive less teaching inputs as a consequence of their more 'advantageous' pupil population. Our main research question is how we must assess these output differences in a fair way, i.e., by taking the differences in inputs and environment into account.

3 Efficiency measurement: method

Consider a general educational system that is characterized, at the level of each pupil, by p inputs and q outputs. We denote the corresponding input vector by $x \in \mathbb{R}^p_+$, and the output vector by $y \in \mathbb{R}^q_+$; in our application, $p = 1$ and the input is the sum of the *REG* and *EEO* instruction units, while $q = 2$ and the outputs are the *MATH* and *DUTCH* scores. The set of all feasible combinations of educational inputs and outputs is the *feasibility set*

$$F = \left\{(x, y) \in \mathbb{R}^{p+q}_+ \mid x \text{ can produce } y\right\}.$$

Educational efficiency analysis relates educational input to educational output. As such, empirical efficiency evaluation essentially requires two steps: (1) we need to empirically estimate the feasibility set F; (2) we have to evaluate observed efficiency by using an (in)efficiency measure that has a meaningful interpretation in terms of the underlying educational objectives. These two issues are discussed next. Subsequently, we discuss two additional issues that will be important for our empirical application: (3) we need to account for outlier observations in the empirical efficiency evaluation; and (4) we want to correct the observed (in)efficiency scores for environmental characteristics, which will also allow us to visualize the impact of the latter on the former. We construct the model step by step.

3.1 Empirical feasibility set

Usually, the 'true' feasibility set F is not observed. To deal with such incomplete information, the nonparametric approach suggests to start from the set of n observed input-output vectors $S \subseteq F$ ($|S| = n$); it assumes that observed input-output combinations are certainly feasible (e.g., Varian, 1984). In addition, we assume that inputs and outputs are freely disposable, which means:

$$\text{if } (x, y) \in F \text{ then } (x', y') \in F \text{ for } x' \geq x \text{ and } y' \leq y.$$

Taken together, these assumptions define the *empirical feasibility set*

$$\widehat{F} = \left\{(x', y') \in \mathrm{I\!R}_+^{p+q} \mid \; x' \geq x \text{ and } y' \leq y \text{ for } (x, y) \in S\right\};$$

i.e., the free disposal hull (FDH) of the set S (e.g., Deprins *et al.*, 1984; Tulkens, 1993).

We briefly discuss the interpretation of the assumptions that underlie the construction of $\widehat{F}$. Firstly, 'free disposability of inputs' means that more input never implies a decrease of the (maximally achievable) output. We believe this is a reasonable assumption in the current context, where inputs stand for instruction units and outputs stand for pupil performance (in alternative disciplines). Secondly, 'free disposability of outputs' means that more output never implies a decrease of the (minimally required) input. Once more, we believe this assumption is tenable in our specific context.

Finally, the assumption $S \subseteq F$ excludes measurement errors and atypical observations, such that all observed input-output vectors are comparable (or, alternatively, that all relevant input and output dimensions are included in the analysis). Admittedly, this assumption may seem problematic in our application, which compares pupils that may be characterized by different background characteristics (that are not explicitly included in our set of conditioning/environmental variables; see infra: conditional inefficiency measure), and which uses inputs that may be characterized by measurement errors (see supra). Therefore, as we will explain further on, we will use an efficiency evaluation method that mitigates the impact of potential outliers within the observed set S.

3.2 Inefficiency measure

In line with the usual practice in primary education, we focus on output performance (see, e.g., Worthington, 2001). Specifically, we use an inefficiency measure which is, for a given input, equal to the maximally possible output performance minus the actual output performance. The output performance is measured as a weighted sum of the output performances in alternative disciplines (captured by the q constituent components of each output vector y), which again reflects the usual practice in primary education. Suppose, that we are to evaluate a pupil observation $(x_E, y_E) \in S$ (also referred to as 'observation E' in what follows) and that the relevant output weights are given by $w_E \in \mathbb{R}^q_+$. For the empirical feasibility set $\widehat{F}$, educational inefficiency for this pupil is defined as

$$\theta_E = \max_{(x,y)\in\widehat{F}} \left\{ \frac{w_E \cdot (y - y_E)}{w_E \cdot g} \mid x \leqslant x_E \right\},$$

with $g \in \mathbb{R}^q_+$ an aggregation vector that defines the denominator as a weighted sum of the output weights; we use $w_E \cdot g > 0$. For the given input level, the measure takes the difference of (linearly aggregated) maximal output performance over actual output performance; this difference is normalized by dividing through the weighted sum $w_E \cdot g$. Clearly, $\infty > \theta_E \geq 0$. Efficiency implies $\theta_E = 0$; and higher inefficiency values generally reveals more inefficiency. In our application, we set the aggregation vector g equal to a q-dimensional vector of ones, which implies that the denominator is simply the (equally weighted) sum of weights. We believe this specification of g is appropriate in our application context because the outputs

(*MATH* and *DUTCH*) are measured in a comparable measurement unit: it naturally corrects for the scale of the output weights w_E (i.e., κw_E obtains the same results as w_E for all $\kappa > 0$), while treating the (directly comparable) output dimensions identically. But it should be clear that, in general, our method also allows for other specifications of g, which accounts for the possibility that different outputs are expressed in different measurement units.[4]

The measure θ_E assumes that the weighting vector w_E is fixed *a priori*. In our application, we will focus on an alternative inefficiency measure that allows for flexible weighting. This is particularly relevant in the present context, because the teaching curricula are typically different among schools. Specifically, for each pupil observation we choose 'most favorable' weights $\widehat{w}_E$ that minimize the inefficiency of the input-output vector under evaluation; this conveniently allows for 'specialization' in learning (at the school level and/or the pupil level): e.g., if pupils perform relatively well in mathematics, then this discipline gets a relatively high weight in their inefficiency measure. To avoid undesirable 'extreme' specialization, we impose that the endogenously selected relative output weights $\widehat{w}_E$ should respect upper and lower bounds, which are captured by the set $W_E \subseteq \mathbb{R}^q_+$ characterized in terms of linear constraints ($\widehat{w}_E \in W_E$ satisfying $\widehat{w}_E \cdot g > 0$). (The construction of W_E for our empirical application is discussed in the beginning of section 4.) This yields the empirical *inefficiency measure*

$$\widehat{\theta}_E = \min_{\widehat{w}_E \in W_E} \max_{(x,y)\in \widehat{F}} \left\{ \frac{\widehat{w}_E \cdot (y - y_E)}{\widehat{w}_E \cdot g} \mid x \leqslant x_E \right\}.$$

Clearly, for $w_E \in W_E$ we have $\theta_E \geq \widehat{\theta}_E \geq 0$. The measure $\widehat{\theta}_E$, with endogenously defined most favorable weights, has a directly similar interpretation as the measure θ_E, with *a priori* fixed weights w_E.

To conclude, we note that the empirical inefficiency measure can be computed by simple linear programming. Specifically, given the construction of $\widehat{F}$, the computation proceeds in two steps. The first step identifies the set of observations that dominate the evaluated

[4] In this respect, it is also worth indicating that, for general g, the 'empirical' inefficiency measure $\widehat{\theta}_E$ (cfr. infra) is formally similar to the so-called 'directional distance function'; see, for example, the duality results in Chambers *et al.* (1998, p. 358). These authors also provide a discussion on possible specifications of g; while they focus on profit efficiency, the analogy with our setting is straightforward.

observation in input terms:

$$D_E = \{(x, y) \in S \mid \; x \leqslant x_E\}.$$

The second step involves the linear programming problem. As a preliminary note, we recall that $\widehat{w}_E \cdot g > 0$ in the above definition of $\widehat{\theta}_E$, so that we can use the normalization $\widehat{w}_E \cdot g = 1$ (because the set W_E only restricts the relative output weights). As such, we can compute

$$\widehat{\theta}_E = \min_{u, \widehat{w}_E \in W_E} \left\{ u - \widehat{w}_E \cdot y_E \mid \begin{array}{c} \widehat{w}_E \cdot g = 1 \\ u \geq \widehat{w}_E \cdot y \quad \forall y : (x, y) \in D_E \\ \widehat{w}_E \in W_E \end{array} \right\}.$$

This is a linear programming problem given that the set W_E is characterized by linear constraints. The fact that merely linear programming is required for the computation of the empirical inefficiency measure $\widehat{\theta}_E$ (after a trivial check of input dominance) makes it attractive for practical applications.

3.3 Outlier-robust inefficiency measure

To mitigate the impact of (potential) outlier behavior and to allow for uncertainty in the observed sample S, we use the order-m method as suggested by Cazals *et al.* (2002); we adapt the method for the specific inefficiency measure $\widehat{\theta}_E$ defined above. Essentially, in terms of the terminology introduced above, this boils down to repeatedly drawing (with replacement) R subsets $D_E^{r,m}$ ($r = 1$, ..., R) from the dominating set D_E; each subset $D_E^{r,m}$ contains (*at most*) m (> 1) (*different*) input-output vectors that are selected from D_E, i.e., $D_E^{r,m} \subseteq D_E$ and

$|D_E^{r,m}| \leq m$.[5] For each $D_E^{r,m}$ we compute the corresponding empirical inefficiency measure

$$\widetilde{\theta}_E^{r,m} = \min_{u,\widehat{w}_E \in W_E} \left\{ u - \widehat{w}_E \cdot y_E \;\middle|\; \begin{array}{c} \widehat{w}_E \cdot g = 1 \\ u \geq \widehat{w}_E \cdot y \quad \forall y : (x,y) \in D_E^{r,m} \\ \widehat{w}_E \in W_E \end{array} \right\},$$

which again uses linear programming. Subsequently, the outlier-robust *order-m inefficiency measure* is defined as the arithmetic average

$$\widetilde{\theta}_E^{m} = \frac{\sum_{r=1}^{R} \widetilde{\theta}_E^{r,m}}{R}.$$

Referring to Cazals *et al.* (2002), this measure has attractive statistical properties and conveniently mitigates outlier behavior. See also Simar (2003) for a related discussion.[6] As a final note, because it can well be that $(x_E, y_E) \notin D_E^{r,m}$, we may have $\widetilde{\theta}_E^{r,m} < 0$ and thus also $\widetilde{\theta}_E^{m} < 0$. We label such observations as 'super-efficient'.

3.4 Environment-corrected inefficiency measure

To capture environmental effects, we use the procedure outlined by Daraio and Simar (2005, 2007). Like before, we adapt this method to the specific inefficiency measure under consideration by implementing the Daraio and Simar procedure in the outlier correction, which in turn is an adaptation of the simple efficiency evaluation model.

Suppose we want to take up k environmental characteristics, which corresponds to a k-dimensional vector z of environmental indicators associated with each input-output vector (x, y); in our application, $k \leq 3$ and the vector z captures *SES*, *B-MATH* and/or *B-DUTCH.* For the evaluated observation E, the Daraio-Simar procedure computes an environment-

[5] Remark that, to correct for the non-representative nature of our dataset, we take the probability of drawing a pupil proportional to the inverse of the probability that this pupil appears in the sample due to the specific sampling design. A similar qualification applies to the environment-corrected inefficiency measure where we weight the Kernel functions by the inverse of the sampling probability.

[6] Cazals *et al.* (2002) actually consider an efficiency measure that does not consider linear but monotonic aggregation of the outputs. But their main results carry over to the linear variant that we consider. A similar qualification applies for our use of the procedure of Daraio and Simar (2005) to account for environmental effects in the efficiency evaluation exercise. In fact, these authors also focus on input efficiency, while we translate their procedure towards output efficiency.

corrected inefficiency measure by conditioning on the corresponding value z_E of the environmental vector: it selects input-output vectors $(x, y) \in D_E$ with z in the neighborhood of z_E. This gives us the *conditional inefficiency measure*

$$\widehat{\theta}_E\,(z_E) = \min_{u,\widehat{w}_E \in W_E} \left\{ u - \widehat{w}_E \cdot y_E \;\middle|\; \begin{array}{ll} \widehat{w}_E \cdot g = 1 & \\ u \geq \widehat{w}_E \cdot y & \forall y : (x, y) \in D_E(z_E) \\ \widehat{w}_E \in W_E & \end{array} \right\},$$

with $D_E(z_E) = \{(x, y) \in D_E \mid |z_E - z| \leq h\}$ and h a Kernel bandwidth vector. In our application, when the number of conditioning variables k is larger than 1, we first apply a so-called Mahalanobis transformation to decorrelate the environmental variables (see, e.g., Mardia *et al.*, 1979). Afterwards, we perform a sequential Kernel estimation —as if all environmental variables were independently distributed— to compute the optimal bandwidth vector (via the likelihood cross-validation criterion). Similar to before, outlier-robust conditional inefficiency measures $\widetilde{\theta}_E^m\,(z_E)$ can be obtained by the order-m method; in that case, we use the Kernel estimates to repeatedly draw subsets of size m.

4 Efficiency measurement: application

In this section, we focus on visualizing the impact of the environmental variables *SES*, *B-MATH* and *B-DUTCH* on educational efficiency at the pupil level, by using the outlier-robust order-m inefficiency measures described in the previous section. For these measures, an additional consideration concerns the specification of the parameters R (the number of drawings with replacement) and m (the number of input-output vectors selected from D_E in each drawing). In the following, we discuss empirical results for $R = 50$ and $m = 100$ as, from these values on, the number of super-efficient observations (see supra) in the sample is robust at around 1%; Daraio and Simar (2007) use similar criteria for defining m. Still, at this point it is worth stressing that we have also experimented with other values for R ($R = 10, 25, 100$) and m ($m = 10, 25, 50, 125, 150$); these alternative configurations generally obtained the same qualitative conclusions. For compactness, we do not include all these results in the current paper, but they are available from the authors upon simple request.

As discussed before, our application avoids ‘extreme’ specialization in either $DUTCH$ or $MATH$ by focusing on a restricted set $W_E \subseteq \mathbb{R}^q_+$ (with $q = 2$), which captures upper and lower bounds of the relative output weights. To construct these bounds, we divide the number of hours spent on $DUTCH$ in the classroom by the sum of the number of instruction hours spent on $DUTCH$ and $MATH$. This reflects the weight attached to $DUTCH$ (relative to $MATH$) in the second year of primary education. The average equals 0.54 —and is very similar for the different school types— while the 1 and 99-percentile values equal 0.44 and 0.71, respectively. These 1 and 99-percentile values will serve as (relative) weight restrictions for $DUTCH$ (and hence 0.56 and 0.29 for $MATH$). To check the sensitivity of our main results with respect to this particular specification of W_E, we have also considered extreme scenarios with no weight flexibility (i.e., using 0.50 as a fixed weight for the two outputs $DUTCH$ and $MATH$) and full weight flexibility (i.e., $W_E = \mathbb{R}^q_+$, with $\widehat{w}_E \cdot g = 1$ for $\widehat{w}_E \in W_E$). Our main qualitative results appeared to be robust for these alternative weight bounds; the corresponding results will not be reported in the current paper, but they are available from the authors upon simple request.

4.1 Outlier-robust inefficiency measures

Before visualizing the impact of the different environmental variables under study, Table 4 provides summary statistics for alternative outlier-robust order-m inefficiency measures (individual efficiency scores are available upon request). We report results for the full sample (see the column ‘all’) and for the subsamples that correspond to the different school types (private schools, local public schools and Flemish public schools).

Table 4: Some summary statistics for the robust inefficiency measures.

	school type	all	private	public	
environment				local	Flemish
$\emptyset$	average	26.99	25.74	27.68	31.61
	std. dev.	13.85	13.28	13.78	15.49
	minimum	-5.70	-3.50	-3.67	-5.70
	maximum	74.10	74.03	74.10	71.39
SES	average	26.97	25.39	27.43	31.17
	std. dev.	13.80	13.25	13.78	15.27
	minimum	-3.02	-2.85	-3.02	-2.34
	maximum	75.96	73.45	75.96	70.03
B-MATH	average	24.17	23.00	25.12	27.88
	std. dev.	12.34	11.87	12.28	13.62
	minimum	-5.74	-5.42	-4.92	-5.74
	maximum	72.46	61.63	71.52	72.46
B-DUTCH	average	23.61	22.61	24.34	27.04
	std. dev.	12.35	11.80	12.39	14.03
	minimum	-6.54	-1.37	-6.54	-0.86
	maximum	65.41	62.88	65.41	60.09
B-MATH, B-DUTCH & SES	average	17.18	16.34	18.52	18.70
	std. dev.	10.16	9.78	10.35	11.04
	minimum	-17.52	-1.99	-17.52	-3.39
	maximum	55.14	49.72	55.14	53.43

Let us first regard the unconditional inefficiency values (with environment = $\emptyset$). Table 4 reports an average inefficiency score of 26.99 for all pupils in our sample. In words, the average pupil achieves an output level that is 26.99 points below the best possible performance for (at most) the same amount of instruction units (= REG + EEO = input). To interpret this result, we recall that aggregate output performance is measured as a weighted sum of the output performance in the disciplines *MATH* and *DUTCH* (using 'most favorable' weights for each individual pupil), and that the *MATH* and *DUTCH* scores are both set between 0 and 100. As such, this average shortage of 26.99 points should be compared to a ('theoretical') maximum possible shortage of 100 points. Next, we also observe much variation in the inefficiency scores over pupils. For example, the standard deviation in the inefficiency values is 13.85; and the maximum inefficiency value amounts to 74.10 points, while the minimum value equals -5.70.[7] Note, finally, that we find differences in the distributions for different school types; for example, the average inefficiency value for private schools (25.74) is below that for local public schools (27.68), which in turn is below that for Flemish public schools (31.61).

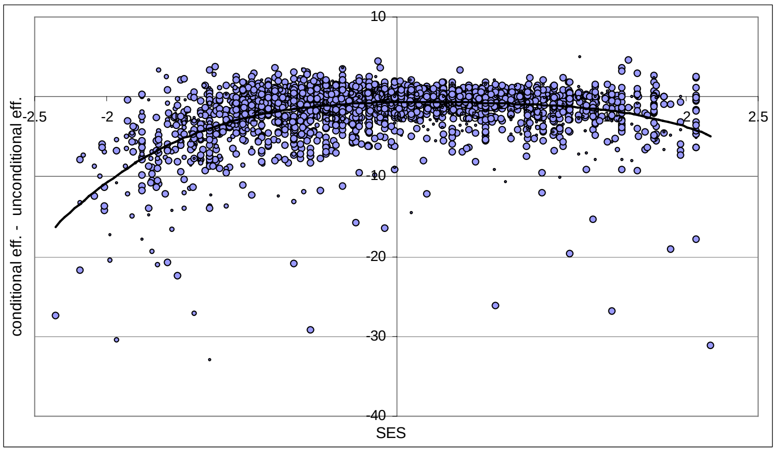
In the following, we investigate to what extent these patterns in the distribution of the inefficiency scores can be attributed to environmental differences, as captured by the variables *SES*, *B-MATH* and *B-DUTCH*. The summary statistics in Table 4 provide some preliminary insights. We first consider the separate impact of the social and cultural environment of a pupil's home (captured by *SES*) and the cognitive antecedents of the pupil (captured by *B-MATH* and *B-DUTCH*). As expected, we find that all three variables influence the pupils' inefficiency values. For example, when focusing on the full sample (see the column 'all'), the average inefficiency reduces (marginally) to 26.97 when controlling for *SES*, and it reduces (more substantially) to 24.17 and 23.61 when controlling for, respectively, *B-MATH*, and *B-DUTCH*. In addition, we observe a decrease in the variation of the inefficiency values; for example, the standard deviation reduces to 13.80, 12.34 and 12.35 when conditioning on, respectively, *SES*, *B-MATH* and *B-DUTCH*. This indicates that each individual variable can explain the observed variation in the inefficiency values to some extent. Finally, if we simultaneously control for *SES*, *B-MATH* and *B-DUTCH*, we observe a further and rather substantial decrease of the average inefficiency value (to 17.18 for 'all') as well as the standard deviation of inefficiency values (to 10.16 for 'all'). This suggests that simultaneous consideration of all three environmental variables can effectively yield additional 'explanatory' value

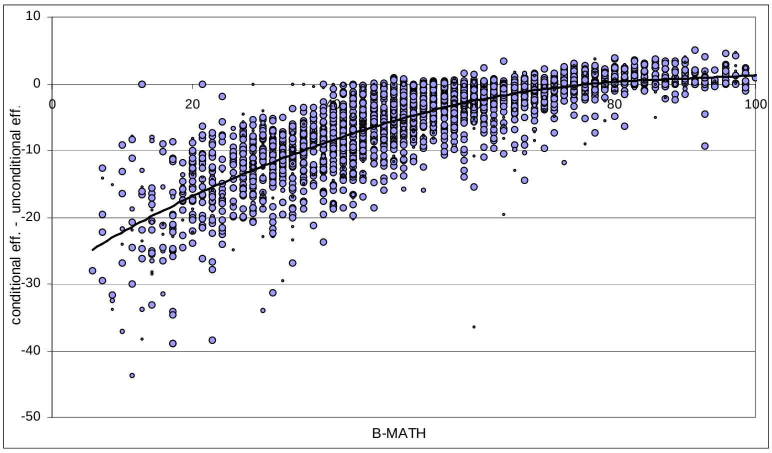
[7] We recall that negative inefficiency values are possible for super-efficient observations because we focus on outlier-robust inefficiency measures.

in terms of explaining patterns of educational inefficiency. The same general conclusions hold for all three school types (private schools, local public schools and Flemish public schools). Remark, finally, that for all specifications of the conditioning variables that we consider, private schools are, on average, more efficient than both types of public schools, and that local public schools outperform Flemish public schools.

4.2 Environmental effects

To visualize environmental effects and, consequently, to detect whether an environmental variable is favorable or unfavorable, we adapt Daraio and Simar (2007)'s methodology to our setting. If z_E^{-j} denotes the vector of all conditioning variables, except for the j-th entry, and z_E^j is the j-th entry, then we can nonparametrically regress the differences $\widetilde{\theta}_E^m(z_E) - \widetilde{\theta}_E^m\left(z_E^{-j}\right)$ on the observed values for z_E^j. If, for a certain range, the regression is decreasing, the j-th environmental variable is unfavorable to output, behaving as a 'substitutive' output in the educational process. Conversely, an increasing curve indicates a favorable variable that plays the role of a 'substitutive' input in the educational process. Finally, a flat curve suggests that there is no output effect of the environmental variable.





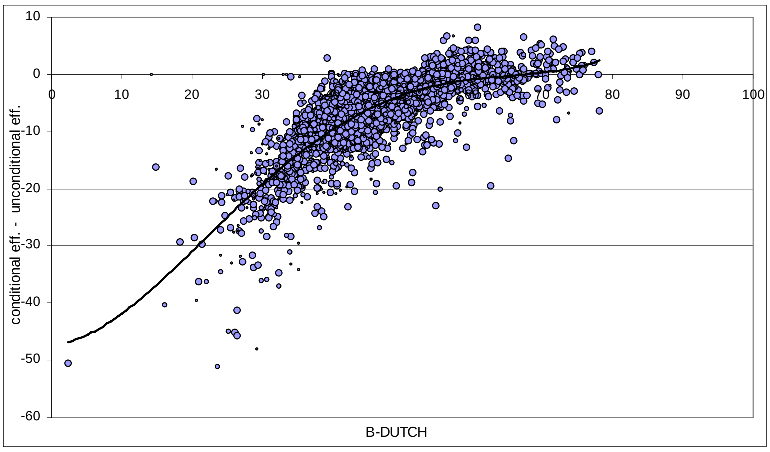


Figure 1: (Environmental impact) *SES*, *B-MATH* and *B-DUTCH.*

Figure 1 portrays the environmental effects. We first consider the variable *SES.* Generally, we find a positive effect of *SES* on the educational output for low *SES* values, and a negative effect for high *SES* values; generally, the effect of *SES* on output gradually decreases for higher *SES* values. We infer that, while *SES* admittedly has some (positive) effect on educational efficiency, much of this effect is already captured by the other two variables *B-MATH* and *B-DUTCH*, which causes the residual impact of *SES* to be rather low.

Let us then regard the variable *B-MATH.* Figure 1 reveals a positive impact of *B-MATH* on output, which tends to decrease for higher *B-MATH* values. Compared to the *SES* picture, the positive effect is more pronounced, which provides more convincing support for this residual *B-MATH* effect.

Finally, we consider the variable *B-DUTCH.* The general conclusions drawn from Figure 1 are similar to those for *B-MATH*: there is a clearly positive effect, wich again decreases for higher *B-DUTCH* values. In this case, the positive effect is very clearly marked, and the observation points are narrowly scattered around the full line, which provides strong support for this conclusion.

Overall, we believe that our results provide sufficiently strong support for simultaneous conditioning on all three variables when comparing the educational efficiency for different pupils. Therefore, our aggregation exercise in the next section will mainly focus on such fully conditioned educational efficiency values.

5 Aggregation: efficiency versus equity

This section aims to compare the aggregate efficiency and equity performance of private schools, local public schools and Flemish public schools. Specifically, we start with the pupils' inefficiency values and the corresponding optimal weights that underlie the results presented in the previous section. Using these pupil-specific weights to aggregate *DUTCH* and *MATH*, we obtain what we call the 'actual score' s_a, with $0 \leq s_a \leq 100$. Adding the inefficiency score θ to it, we get the so-called 'potential score' $s_p = s_a + \theta$, with $0 \leq s_p \leq 100$. It follows from our previous discussion that these potential scores correct for input differences (in terms of *REG* and *EEO* instruction units), and avoid extreme specialization in *DUTCH* or *MATH* (through weight bounds). In addition, given that we focus on robust inefficiency measures, it also accounts for possible outlier behavior. Finally, if we use the conditional inefficiency measures, we also correct for environmental differences (in terms of *SES*, *B-MATH* and *B-DUTCH*).[8]

We want to investigate whether one school type is 'better' than another in a 'robust'

[8] We note that, when conditioning on *B-MATH* and *B-DUTCH*, we also correct for the pupils' starting level of output performance in the aggregation exercise.

way, which here means that we impose minimal normative assumptions when aggregating outcomes. To do so, we focus on two multidimensional stochastic dominance criteria developed in Atkinson and Bourguignon (AB; 1982). The first criterion (FAB) only cares about efficiency, whereas the second criterion (SAB) also takes equity into account.

5.1 Only efficiency matters: FAB

We assume that the overall performance of a school can be measured by the average performance of its pupils, where the pupil performance is measured via a function $P : \mathbb{R}^2 \rightarrow \mathbb{R}$ which maps a pupil's actual and potential score (s_a, s_b) into a scalar $P(s_a, s_b)$. If only educational efficiency matters, then (1) an increase in the actual score s_a of a pupil, *ceteris paribus* (i.e., for a given potential score s_p), must increase the school type's overall performance, and (2) the same holds for a decrease in a pupil's potential score s_p, *ceteris paribus* (i.e., for a given actual score s_a). Formally, one school type, say type A, is better than another school type, say B, according to FAB, denoted $A \succsim_1 B$, if and only if the average performance is higher in A than in B for all (differentiable) performance functions $P : \mathbb{R}^2 \rightarrow \mathbb{R}$ in $\mathbb{P}_1 = \{P \,|P_1' \geq 0 \text{ and } P_2' \leq 0\}$.[9] A special case of interest is $P(s_a, s_p) = s_a - s_p = -\theta$, in which case one school type would be judged better if the average inefficiency is lower. We focus on the more general criterion

$$A \succsim_1 B \Leftrightarrow \int_0^{100} \int_0^{100} PdF_A - \int_0^{100} \int_0^{100} PdF_B \geq 0, \text{ for all } P \text{ in } \mathbb{P}_1,$$

with F_A and F_B the bidimensional distribution functions of the actual and potential scores for both school types. Atkinson and Bourguignon (1982) provide an equivalent implementable condition for $A \succsim_1 B$, requiring that (1) the proportion of 'better' pupils is higher in A

[9] In this formulation, P_1' stands for the first derivative of P with respect to the j-th argument $(j = 1, 2)$. Similarly, P_{12}'' will stand for the corresponding cross-derivative in our discussion of SAB.

everywhere and (2) the proportion of 'worse' pupils is lower in A everywhere. Formally,

$$A \succsim_1 B \Leftrightarrow \begin{cases} L_A^1\left(s_a, s_p\right) - L_B^1\left(s_a, s_p\right) \geq 0, \text{ for all } \left(s_a, s_p\right) \in [0, 100]^2 \\ L_A^2\left(s_a, s_p\right) - L_B^2\left(s_a, s_p\right) \leq 0, \text{ for all } \left(s_a, s_p\right) \in [0, 100]^2 \end{cases}, \tag{1}$$

with $L_j^1\left(s_a, s_p\right) = \int_{s_a}^{100} \int_0^{s_p} dF_j$ the proportion of pupils in school type $j = A$ or B who perform definitely better compared to $\left(s_a, s_p\right)$ and $L_j^2\left(s_a, s_p\right) = \int_0^{s_a} \int_{s_p}^{100} dF_j$ the proportion of pupils in school type $j = A$ or B who perform definitely worse. Notice that FAB is a robust ranking criterion, since it holds for all specifications of P within $\mathbb{P}_1$. Still, it comes at a cost, since two distributions might turn out to be non-comparable.

Table 5 presents our results for the FAB criterion in equation (1). We consider two extreme cases: the first case (denoted by $Z = \{REG{+}EEO;\emptyset\}$) does account for input differences but *not* for environmental differences by calculating the potential scores on the basis of the unconditional inefficiency measures $\widetilde{\theta}_E^m$ (which coincide with $\widetilde{\theta}_E^m\left(z_E\right)$ for z_E empty); the second case (denoted by $Z = \{REG{+}EEO;SES,B\text{-}MATH,B\text{-}DUTCH\}$) simultaneously takes account of input and *all* three environmental variables (i.e., it is based on the measure $\widetilde{\theta}_E^m\left(z_E\right)$, with z_E capturing *SES*, *B-MATH* and *B-DUTCH*). For each case, Table 5 reports the dominance relation between the row school type and the column school type: either the row school type 'dominates', 'is dominated by', or is not comparable to ('not comp. to') the column type. Two remarks are in order. Firstly, following the usual practice, dominance is checked at a finite number of points $\left(s_a, s_p\right) \in \{0, 25, 50, 75, 100\}^2$. Secondly, we use a naive bootstrap procedure for statistical inference. That is, we calculate the proportion of the total number of bootstraps, i.e., 10000 drawings with replacement from the original sample, in which a certain result ('dominates', 'is dominated by', or 'not comp. to') is found.[10] In Table 5 we mention the empirical result for each comparison, together with the corresponding 'naive' p-value, i.e., the proportion of times this result was found. A large p-value indicates a rather robust empirical result.

Considering only average test scores, we saw in Table 2 that private schools outperform

[10] Notice that, from 5000 bootstrap samples onwards, the results remain stable.

local public schools, while the latter in turn outperform Flemish public schools. However, by including both inputs, outputs and exogenous environmental characteristics, the robust FAB criterion does not support this conclusion anymore. Indeed, the results in Table 5 indicate non-comparabilities between all school types. This result holds for both (extreme) specifications of Z that we consider.

Table 5: FAB dominance results for two extreme cases of environment-correction.

$Z =$	$\{$REG+EEO;$\emptyset\}$		$\{$REG+EEO;SES,B-MATH,B-DUTCH$\}$	
	local public	Flemish public	local public	Flemish public
private catholic	not comp. to (1.0000)	not comp. to (1.0000)	not comp. to (1.0000)	not comp. to (1.0000)
local public		not comp. to (1.0000)		not comp. to (1.0000)

5.2 Equity also matters: SAB

In addition to a preference for efficiency, we next include a preference for equity in our comparisons of the overall performance of different school types. The SAB criterion requires (besides (1) and (2) underlying the FAB criterion) that (3) an increase in a pupil's actual score s_a is valued more (in terms of performance) for pupils with a higher potential score s_p and (4) a decrease in a pupil's potential score s_p increases performance more for pupils with a lower actual score s_a. To put it differently, a higher correlation between the pupils' actual and potential scores results in a better overall performance of the school type. Formally, school type A is better than school type B according to SAB, denoted $A \succsim_2 B$, if and only if the average performance is higher in A than in B for all (twice differentiable) performance functions $P : \mathbb{R}^2 \to \mathbb{R}$ in $\mathbb{P}_2 = \{P \,|P'_1 \geq 0, P'_2 \leq 0, \text{ and } P''_{12} \geq 0\}$. We get

$$A \succsim_2 B \Leftrightarrow \int_0^{100} \int_0^{100} P dF_A - \int_0^{100} \int_0^{100} P dF_B \geq 0, \text{ for all } P \text{ in } \mathbb{P}_2.$$

This is equivalent with the implementable condition that the proportion of 'worse' pupils is lower in A compared to B (Atkinson and Bourguignon, 1982), or, formally,

$$A \succsim_2 B \Leftrightarrow L_A^2\left(s_a, s_p\right) - L_B^2\left(s_a, s_p\right) \leq 0, \text{ for all } \left(s_a, s_p\right) \in [0,100]^2, \tag{2}$$

with $L_j^2\left(s_a, s_p\right) = \int_0^{s_a} \int_{s_p}^{100} dF_j$ for all $j = A, B$.

As before, we consider two extreme cases, depending on whether we only correct for inputs (case $Z = \{REG\text{+}EEO;\emptyset\}$) or for both inputs and the complete environment (case $Z = \{REG\text{+}EEO;SES,B\text{-}MATH,B\text{-}DUTCH\}$). Table 6 presents the results. The interpretation of the different entries is similar to that of Table 5, but now pertains to the SAB criterion in (2). Interestingly, we now do find robust dominance relations. If we do not correct for environmental characteristics, we find that the private catholic and the local public schools (robustly) dominate the Flemish public schools; see the middle column of Table 5. However, this comparison is not 'fair', since it does not correct for school environment at all. Therefore, we consider the right column of Table 5 as the fairest base of comparison; and we conclude that these robust dominance results change into either non-robust dominance results or robust non-comparibility results.

Table 6: SAB dominance results for two extreme cases of environment-correction.

$Z =$	{REG+EEO;∅}		{REG+EEO;SES,B-MATH,B-DUTCH}	
	local public	Flemish public	local public	Flemish public
private catholic	not. comp to	dominates	dominates	not comp. to
	(0.7268)	(0.9996)	(0.6881)	(1.0000)
local public		dominates		not comp. to
		(0.9117)		(1.0000)

6 Conclusion

Focusing on educational efficiency, we have presented a nonparametric approach for analyzing public sector efficiency which also accounts for equity considerations. Firstly, we have designed a nonparametric (DEA) model that is specially tailored for educational efficiency evaluation at the pupil level. It requires minimal *a priori* structure regarding the educational feasibility set and objectives. This is particularly convenient in the current context, which typically involves minimal *a priori* information. Next, we introduced multidimensional stochastic dominance criteria that are particularly well-suited for comparing the aggregate educational performance of different school types; they aggregate pupils' output while adjusting for environment-corrected inefficiency. These nonparametric aggregation criteria naturally complement our nonparametric model for evaluating individual (pupil level) efficiency. The first criterion is the appropriate criterion if only efficiency matters. By contrast, the more powerful second criterion is recommendable when equity is important in addition to efficiency; such equity considerations are usually prevalent in the context of public sector efficiency evaluation. We have shown that our approach directly allows for adapting the methodology of Daraio and Simar (2005, 2007), to account for potential uncertainty in the data and environmental characteristics (*in casu* the pupils' educational environment) in the efficiency assessment. Although our application concentrates on educational efficiency, the presented approach is also more generally useful for efficiency evaluation in the public sector.

To avoid interaction between the funding of schools (private versus public funding) and the ideological background, we consider the Flemish situation where both private and public schools are publicly funded. This particular application demonstrates the practical usefulness of our approach. First, we have investigated the impact of the 'environmental characteristics' socio-economic status (*SES*), begin-level in mathematics (*B-MATH*) and language proficiency (*B-DUTCH*) on the educational output for individual pupils. Generally, we find that all three environmental variables positively impact on the educational output, and that this positive effect prevails in particular for low initial values for *SES,* *B-MATH* and *B-DUTCH* values - it may be even negative for high initial *SES* values. We believe that our results convincingly support that all three environmental variables should simultaneously be accounted for to obtain a fair efficiency evaluation. Next, we have compared the aggregate efficiency of private (but publicly funded) schools, local public schools and Flemish public schools (the latter depending on the level of government initiation). Focusing on the first stochastic dominance criterion, we find that no school type robustly dominates another school type; we conclude 'non-comparability' in all pairwise comparisons. However, the story changes if we focus on the second criterion; in that case, private catholic and local public schools (robustly) dominate the Flemish public schools if we do not account for environmental differences. Still, if we account for the diverging environmental characteristics of the pupil populations, we no longer find robust dominance relations between different school types and, as such, no school type robustly dominates any other school type. This contrasts with the conclusions of e.g. Coleman *et al.* (1982). Given that these aggregate comparisons nonparametrically account for efficiency, equity and environment, we consider them as 'fairest' in the (public sector) evaluation context under study. In turn, the paper demonstrates that school comparisons on the basis of average scores instead of on the basis of environment corrected scores could result in biased conclusions.

References

[1] Altonji, J.G., T.E. Elder and C.R. Taber (2005a), Selection on observed and unobserved variables: assessing the effectiveness of catholic schools, *Journal of Political Economy* 113 (1), 151-184.

[2] Altonji, J.G., T.E. Elder and C.R. Taber (2005b), An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling, *Journal of Human Resources* 40 (4), 791-821.

[3] Atkinson, A.B. and F. Bourguignon (1982), The comparison of multidimensioned distributions of economic status, *Review of Economic Studies* 49, 183-201.

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

[5] Cain, G.G., A.S. Goldberger (1983), Public and private schools revisited, *Sociology of Education* 56 (4), 208-218.

[6] Chambers, R., R. Färe and Y. Chung (1998), Profit, directional distance functions, and nerlovian efficiency, *Journal of Optimization Theory and Applications* 98, 351-364.

[7] Coleman, J., T. Hoffer and S. Kilgore (1982), Cognitive Outcomes in Public and Private Schools, *Sociology of Education* 55 (2), 65-76.

[8] Daraio, C. and L. Simar (2005), Introducing environmental variables in nonparametric frontier models: a probabilistic approach, *Journal of Productivity Analysis* 24, 93–121.

[9] Daraio, C. and L. Simar (2007), *Advanced robust and nonparametric methods in efficiency analysis. Methodology and applications*, Series: Studies in Productivity and Efficiency, Springer.

[10] Deprins, D., L. Simar and H. Tulkens (1984), Measuring labor efficiency in post offices, *The Performance of Public Enterprises: Concepts and Measurements*, M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243–267.

[11] De Witte, K., E. Thanassoulis, G. Simpson, G. Battisti and A. Charlesworth-May (2008), Assessing pupil and school performance by non-parametric and parametric techniques, *Aston University Working Paper.*

[12] Goldstein, H. (2003), *Multilevel Statistical Models – Third edition',* Arnold Publisher. London.

[13] Grosskopf, S., K. Hayes, L. Taylor and W. Weber (1997), Budget-constrained frontier measures of fiscal equality and efficiency in schooling, *Review of Economics and Statistics* 79, 116-124.

[14] Grosskopf, S., K. Hayes, L. Taylor and W. Weber (1999), Anticipating the consequences of school reform: a new use of DEA, *Management Science* 45, 608-620.

[15] Hanushek, E.A. and M.E. Raymond (2005), Does school accountability lead to improved student performance?, *Journal of Policy Analysis and Management* 24 (2), 297-327.

[16] Mardia, K.V., J.T. Kent and J.M. Bibby (1979), Multivariate analysis, Academic Press New York.

[17] Portela, M. and E. Thanassoulis (2001), Decomposing school and school type efficiencies, *European Journal of Operational Research* 132, 357-373.

[18] Reynders, T., I. Nicaise and J. Van Damme (2005), Longitudinaal onderzoek in het basisonderwijs. De constructie van een SES-variabele voor het SIBO-onderzoek, *LOA-rapport* 31.

[19] Sandström, F.M. and F. Bergström (2005), School vouchers in practice: competition will not hurt you, *Journal of Public Economics* 89 (2-3), 351-380.

[20] Simar, L. (2003), Detecting outliers in frontier models: a simple approach, *Journal of Productivity Analysis* 20, 391–424.

[21] Toma, E.F. (1996), Public Funding and Private Schooling across Countries, *Journal of Law and Economics* 39 (1), 121-148.

[22] Tulkens, H. (1993), On FDH efficiency analysis: some methodological issues and applications to retail banking, courts and urban transit, *Journal of Productivity Analysis* 4 (1/2), 183-210.

[23] Varian, H.R. (1984), The nonparametric approach to production analysis, *Econometrica* 52, 579-598.

[24] Worthington, A.C. (2001), An empirical survey of frontier efficiency measurement techniques in education, *Education Economics* 9, 245-268.