UNIL | Université de Lausanne
IDHEAP
Institut de hautes études
en administration publique

Jean-Marc Huguenin

DEA does not like positive discrimination : A comparison of alternative models based on empirical data

Working paper de l'IDHEAP 7/2014
Unité Finances publiques

UNIL | Université de Lausanne
IDHEAP
Institut de hautes études
en administration publique

DEA does not like positive discrimination : A comparison of alternative models based on empirical data

Jean-Marc Huguenin

Working paper de l'IDHEAP 7/2014
Unité Finances publiques/
IDHEAP Working Paper 7/2014
Chair of Public finance
July 2014

Abstract

Due to the existence of free software and pedagogical guides, the use of Data Envelopment Analysis (DEA) has been further democratized in recent years. Nowadays, it is quite usual for practitioners and decision makers with no or little knowledge in operational research to run their own efficiency analysis. Within DEA, several alternative models allow for an environmental adjustment. Five alternative models, each user-friendly and easily accessible to practitioners and decision makers, are performed using empirical data of 90 primary schools in the State of Geneva, Switzerland. As the State of Geneva practices an upstream positive discrimination policy towards schools, this empirical case is particularly appropriate for an environmental adjustment. The majority of alternative DEA models deliver divergent results. It is a matter of concern for applied researchers and a matter of confusion for practitioners and decision makers. From a political standpoint, these diverging results could potentially lead to opposite decisions. As no consensus emerges on the best model to use, practitioners and decision makers may be tempted to select the model that is right for them, in other words, the model that best reflects their own preferences. Further studies should investigate how an appropriate multi-criteria decision analysis method could help decision makers to select the right model.

Grâce à l'existence de logiciels en libre accès et de guides pédagogiques, la méthode Data Envelopment Analysis (DEA) s'est démocratisée ces dernières années. Aujourd'hui, il n'est pas rare que les décideurs avec peu ou pas de connaissances en recherche opérationnelle réalisent eux-mêmes leur propre analyse d'efficience. A l'intérieur de la méthode DEA, plusieurs modèles permettent de tenir compte des conditions plus ou moins favorables de l'environnement. Cinq de ces modèles, facilement accessibles et applicables par les décideurs, sont utilisés pour mesurer l'efficience des 90 écoles primaires du canton de Genève, Suisse. Le canton de Genève pratiquant une politique de discrimination positive envers les écoles défavorisées, ce cas pratique est particulièrement adapté pour un ajustement à l'environnement. La majorité des modèles DEA génèrent des résultats divergents. Ce constat est préoccupant pour les chercheurs appliqués et perturbant pour les décideurs. D'un point de vue politique, ces résultats divergents peuvent conduire à des prises de décision différentes selon le modèle sur lequel elles sont fondées. Dans la mesure où aucun consensus n'émerge sur le meilleur modèle à utiliser, les décideurs peuvent être tentés de choisir le modèle qui reflète au mieux leurs préférences. L'application d'une méthode d'aide à la décision multi-critères, pour aider les décideurs à choisir le « bon » modèle, devrait êtes investiguée dans de futures recherches.

CONTENTS

1. Context

The use of Data Envelopment Analysis (DEA) is experiencing rapid and continuous growth. In 2002, Tavares (2002) identified 3203 DEA publications (journal articles, research articles, event articles, books and dissertations). In 2008, Emrouznejad, Parker and Tavares (2008) inventoried more than 7000 publications. This growth reflects the need for user-friendly performance measurement methods. In recent years, the use of DEA has been further democratized due to *(1)* the existence of free software, such as Win4DEAP, Efficiency Measurement System or DEA Solver, *(2)* the publication of pedagogical guides (Coelli, 1996; Coelli, Prasada Rao, O'Donnell & Battese, 2005, pp. 161-206; Huguenin, 2012; Huguenin, 2013a; Huguenin, 2013b) and *(3)* the teaching of DEA in under- and postgraduate programs[1]. The Google Scholar search engine returned 51 400 documents after a search on 'data', 'envelopment' and 'analysis' on 15 March 2013. Nowadays, it is quite usual for practitioners and decision makers with little or no background in operational research and economics to run their own efficiency analysis[2]. For instance, a web-based platform integrating DEA has been developed in Portugal for secondary schools' headteachers (Portela, Camanho & Borges, 2001). Users are able to perform their own efficiency analysis by selecting the schools to be included in the dataset and the variables to be included in the analysis[3].

The external environment could influence the ability of management to convert inputs into outputs and, as a result, impact entities' technical efficiency. Following Coelli *et al.* (2005, p. 190), an environmental variable is defined as a factor that could influence the efficiency of an entity, where such a factor is not a traditional input and is assumed to be outside of the manager's control. Because it is not under the control of managers, such a factor is also called a non-discretionary variable[4]. It cannot be varied at the discretion of an individual manager but nevertheless needs to be taken into account to measure efficiency (Cooper, Seiford & Tone, 2007, p. 215). This paper considers traditional inputs as those covered by the OECD KLEMS model, which considers five categories of inputs: capital (K) labour (L), energy (E), materials (M) and services (S) (OECD, 2001).

Examples of environmental variables include ownership differences (such as public versus private), location characteristics, labour relations (such as conflictual versus peaceful relationships between trade unions and employers' organizations) and government regulations (Fried, Schmidt & Yaisawarng, 1999). Location characteristics consist of the environmental variables which are specific to the location of an entity, such as a supermarket influenced by population density.

In the education sector, three main generic drivers can be considered as environmental variables. They influence pupil performance but are outside of the control of headteachers (Soteriou, Karahanna, Papanastasiou & Diakourakis, 1998, p. 68, based on Thanassoulis, 1996, p. 883). They consist of *(1)* pupil characteristics, such as intelligence, willingness or effort propensity, *(2)* family and the external

---

1 For instance, DEA is taught at the University of Lausanne, Switzerland, in three different courses: *(1)* Public Sector Performance Measurement (Master of Science in Public Policy and Management), *(2)* Public Sector Financial Management (Master of Advanced Studies in Public Administration) and *(3)* Benchmarking (Certificate of Advanced Studies in Administration and Management of Educational Establishments). About 90 decision makers in the public sector are trained annually in the use of DEA.

2 The author of this study regularly meets Swiss headteachers who use DEA to assess individual teachers, classes or schools.

3 Note that this plateform represents an example of "ascending" benchmarking (Viger, 2007), where the starting point of the analysis comes from the base (i.e. the headteachers).

4 Non-controllable variables and exogenous variables are used as synonyms to non-discretionary variables in the DEA literature.

environment, such as the socioeconomic status of pupils and *(3)* school related factors (which are outside of the control of headteachers). In this latter category, school size (as measured by the number of pupils) is, for instance, outside of the control of headteachers in Switzerland, as they have to register every single pupil residing in the catchment area defined by school authorities.

Environmental variables in school efficiency measurement using Data Envelopment Analysis (DEA) include (non exhaustive list):

- school location in a particular region (Agasisti, 2013; Burney, Johnes, Al-Enezi & Al-Musallam, 2013);
- types of school, such as private schools (Agasisti, 2013; Kirjavainen & Loikkanen, 1998; Lovell, Walters and Wood, 1994; Ramanathan, 2001), all-girls schools (Alexander & Jaforullah, 2004; Alexander, Haug & Jaforullah, 2010; Bradley, Johnes & Millington, 2001), urban and rural schools (Agasisti, 2013; Alexander & Jaforullah, 2004; Alexander, Haug & Jaforullah, 2010; Denaux, Lipscomb & Plumly, 2011; Kantabutra & Tang, 2006; Kirjavainen & Loikkanen, 1998);
- socioeconomic status of pupils (Alexander & Jaforullah, 2004; Alexander, Haug & Jaforullah, 2010; Borge & Naper, 2006; Bradley, Johnes & Little, 2010; Denaux *et al.*, 2011; McCarty & Yaisawarng, 1993; Ouellette & Vierstraete, 2005; Rassouli-Currier, 2007);
- school size (Agasisti, 2013; Borge & Naper, 2006; Bradley *et al.*, 2001; Duncombe, Miner & Ruggiero, 1997; Kantabutra & Tang, 2006; Kirjavainen & Loikkanen, 1998);
- political factors (Borge & Naper, 2006; Waldo, 2007);
- teacher characteristics (Alexander & Jaforullah, 2004; Alexander, Haug & Jaforullah, 2010; Bradley, *et al.*, 2001; Burney, Johnes, Al-Enezi & Al-Musallam, 2013; Diagne, 2006; Duncombe, Miner & Ruggiero, 1997; Lovell, Walters & Wood, 1994; Ruggiero & Vitaliano, 1999).

Positive discrimination policies are implemented by public authorities to adjust for the environment[5]. They aim to compensate the negative impact of environmental variables (mainly socioeconomic status of pupils) on school performance. In Europe, these priority education policies are defined as

> policies designed to have an effect on educational disadvantaged through systems or programs of focused action (whether the focus be determined according to socioeconomic, ethnic, linguistic, religious, geographic, or educational criteria) by offering something more ('better' or 'different') to designated populations (Frandji, 2008, p. 12).

Within DEA, several models allow for an environmental adjustment. Following Muñiz (2002), they can be grouped in three categories: *(1)* one-stage models (Banker & Morey, 1986a; Banker & Morey, 1986b; Ruggiero, 1996[6]; Yang and Paradi in Muñiz, Ruggiero, Paradi and Yang, 2006), *(2)* multi-stage models including two-stage (Ray, 1988; Ray, 1991), three-stage (Ruggiero, 1998; Fried, Lovell, Schmidt & Yaisawarng, 2002; Muñiz, 2002) and four-stage models (Fried, Schmidt & Yaisawarng,

[5] Some of these policies are essentially built on an ideological basis (Demeuse & Friant, 2012).

[6] Ruggiero (1996) develops an additional one-stage model. However, this model seems to be rather an extension of the Banker and Morey (1986a) model that allows for categorical variables. As it allows continous environmental variables, it is comparable to the Banker and Morey (1986b) model (Ruggiero, 1996, p. 555).

1999) and *(3)* program analysis models (Charnes, Cooper & Rhodes, 1981)[7]. There are few published studies which compare these models with one another.

The empirical field of this study considers the case of public primary schools in the State of Geneva, Switzerland. As the State of Geneva has implemented upstream positive discrimination measures since 2008, this empirical case is particularly appropriate for an environmental adjustment. The Geneva public school system is described in the next section.

2. Geneva public school system

In the State of Geneva, education is compulsory at early childhood (corresponding to the international standard classification of education ISCED # 0) for a duration of 2 years, primary (ISCED # 1) for a duration of 6 years and lower secondary education (ISCED # 2) for a duration of 3 years.

In 2010-2011, the State of Geneva registered 90 public primary schools. These schools are funded by the State government (chiefly for staff salary) and by local authorities – municipalities (chiefly for school infrastructure). Pupil competences are assessed with the use of standardized tests at three different times in two or three subjects. At the end of the second grade, French (mother tongue) and mathematics are assessed; at the end of the fourth and sixth grade, French, German (first foreign language) and mathematics are assessed.

Primary schools are managed by headteachers assisted by one or several teachers working part time as headteachers' assistants. Staff consists of teachers, secretaries and schoolkeepers (maintenance). In some schools, educators are also active.

In order to adjust to local environment, partial autonomy in management is granted to schools. For instance, headteachers define job profiles and recruit teachers; they are responsible for school quality (and hence pupil performance); and they also chair the school board.

Every school has a board composed by representatives of the school staff, parents and city civil-servants and is chaired by the headteacher. The board demonstrates instances of democracy where stakeholders are informed and consulted. Whilst they only have limited authority about school management, they can make propositions about day-to-day school life. School boards aim to develop better relationships between school, families and local communities.

The main characteristics of primary schools are as follows:

- A school can be located on one or several sites (up to five); which implies that school buildings can be spread over several locations (or sites);
- Special education is only available in a limited number of schools (21 schools out of 90); which means that pupils with special needs are grouped in the schools where special education is available;
- Special reception classes for immigrant pupils are only available in a limited number of schools (35 schools out of 90).

[7] Note that Yang and Pollitt (2009) propose the following categories: separative models (in which Charnes, Cooper & Rhodes (1981) and Banker & Morey (1986a) would be classified), one-stage models, two-stage models, three-stage models and four-stage models.

The State of Geneva practices a policy of positive discrimination towards schools. Additional teaching resources are allocated to disadvantaged schools. Five school categories (A to E) are defined according to the percentage of pupils (per school) whose parents are blue-collar workers or unqualified workers – category # 9 of the International Standard Classification of Occupations (Observatory on Primary Education, 2010). This variable, SOCIO, reflects the socioeconomic status of pupils. For instance, schools with a SOCIO proportion of more than 50% are considered as the most disadvantaged schools and are classified in the E category. Table 1 describes the quantity of additional teaching staff per pupil that schools receive.

Table 1
Positive discrimination in Geneva: more teaching staff for disadvantaged schools

Category (# of schools)	Pupils in the lowest socioeconomic category (%)	Pupil/teacher target ratio	Teacher/pupil target ratio	Additional teaching staff per pupil (%)
A (15)	0.00-19.99	18.55	0.0539	0.00
B (20)	20.00-29.99	18.15	0.0551	2.20
C (20)	30.00-39.99	17.45	0.0573	6.30
D (15)	40.00-49.99	16.65	0.0601	11.41
E (20)	50.00-100.00	15.25	0.0656	21.64

Source: General Direction of Primary Schools, Education Department, State of Geneva.

A school in category A has a target teacher/pupil ratio of 0.054 (i.e. 18.55 pupils per teacher). This target is defined by the State authority. Such a school does not receive any additional resources as it is in the most advantaged category. A school in category B has a target teacher/pupil ratio of 0.0551 (i.e. 18.15 pupils per teacher). It receives 2.2% additional teaching staff (i.e. 0.0539 + (0.022 x 0.0539)). A school in category C has a target teacher/pupil ratio of 0.0573 (i.e. 17.45 pupils per teacher). It receives 6.3% additional teaching staff (i.e. 0.0539 + (0.063 x 0.0539)). A school in category D has a target teacher/pupil ratio of 0.0601 (i.e. 16.65 pupils per teacher). It receives 11.41% additional teaching staff (i.e. 0.0539 + (0.1141 x 0.0539)). Finally, a school in category E has a target teacher/pupil ratio of 0.0656 (i.e. 15.25 pupils per teacher). It receives 21.64% additional teaching staff (i.e. 0.0539 + (0.2164 x 0.0539))[8].

[8] As the detrimental condition of the environment is compensated by additional resources, the relevance to actually use a model which allows for an environmental adjustment is open to debate. Consider two schools with one pupil each. Both of them obtain a test's results of 6. The first school faces a detrimental environment and receives 20% additional teaching staff. Instead of having one teacher, it thus has 1.2 teachers. The second school faces a favourable environment. It does not receive additional resources, and stays with one teacher. With a classical DEA model, with no environmental adjustment, the first school obtains an efficiency score of 83.3% and the second one a score of 100%. The first school is penalized for having received additional resources. In order to be 100% efficient, its pupils should obtain a test's result 20% higher than the pupil attending the second school. However, one cannot expect from the disadvantaged pupil to become 20% better than the advantaged pupil. One can probably only expect that the disadvantaged pupil becomes as good as the advantaged pupil. Another point to take into consideration is that the test's results are bounded to a maximum number of points. Consider that 6 is the best grade possible. If both pupils obtain a 6, the first school will always be less efficient, because it is not possible for its pupil to score higher than 6. As a result, it seems appropriate to use a model which allows for an environmental adjustment. With such a model, both schools obtain an efficiency score of 100% in the above mentioned example.

3. Objectives

The aim of this study is to test how several alternative DEA models, each of which measure efficiency, can deliver diverging results. Unlike studies using simulated data, this study intentionally uses empirical data. As a result, the comparison is made between the estimates of the alternative models, without knowing whether these estimates approximate the 'true' efficiency measure (which could be estimated with a simulation analysis)[9]. By using empirical data, this study addresses the issue faced by practitioners and decision makers who perform their own efficiency analysis. It seeks to determine whether the alternative models produce convergent results (consistent efficiency scores and rankings of entities). If the alternative models do produce convergent results, practitioners and decision makers may confidently select any model. If they produce divergent results, the choice of model becomes a strategic issue.

The alternative models tested in this study are all user-friendly and easily accessible to practitioners and decision makers. The empirical case is the 90 primary schools of the State of Geneva, Switzerland. It is particularly well suited to test several alternative models, as *(1)* the State of Geneva practices positive discrimination towards disadvantaged schools and *(2)* schools are grouped in five categories defined by one continuous variable (percentage of pupils whose parents are blue-collar workers or unqualified workers). According to their respective category, schools receive additional teaching staff.

4. Adjusting for the environment in DEA

Within DEA, several models allow for an environmental adjustment. Following Muñiz (2002), they can be grouped into three categories: *(1)* one-stage models (Banker & Morey, 1986a; Banker & Morey, 1986b; Ruggiero, 1996[10]; Yang & Paradi model in Muñiz, Ruggiero, Paradi & Yang, 2006, p. 1176), *(2)* multi-stage models including two-stage (Ray, 1991), three-stage (Ruggiero, 1998; Fried, Lovell, Schmidt & Yaisawarng, 2002; Muñiz, 2002) and four-stage models (Fried, Schmidt & Yaisawarng, 1999) and *(3)* program analysis models (Charnes, Cooper & Rhodes, 1981).

The models which allow for an environmental adjustment are shortly introduced hereafter, alongside their main advantages and drawbacks (Thanassoulis, Portela & Despic, 2008). The basic variable returns to scale DEA model (VRS) is first recalled (Banker, Charnes & Cooper, 1984). This basic model does not allow for an environmental adjustment.

---

[9] Another research question, not treated in this study, would be to determine whether the estimates of alternative models converge or diverge with the 'true' efficiency. This question cannot be answered by using empirical data, as the 'true' efficiency is unknown. The only way to calculate the 'true' efficiency would consist of *(1)* defining a production function, *(2)* generating inputs from a random distribution and *(3)* deriving outputs. Note that existing studies using simulated data provide mixed results about the convergence of alternative models with the 'true' efficiency (see Section 5 about it).

[10] Ruggiero (1996) develops an additional one-stage model. However, this model seems to be an extension of the Banker and Morey (1986a) model that allows for categorical variables. As it allows continous environmental variables, it is comparable to the Banker and Morey (1986b) model (Ruggiero, 1996, p. 555).

Banker, Charnes and Cooper (1984) – No environmental adjustment

The basic VRS model measures entities' technical efficiency under the assumption of variable returns to scale.

Following the notation adopted by Johnes (2004, pp. 630-637), there are data on s outputs and m inputs for each of n primary schools to be evaluated ($n = 90$ in the current study). y_{rk} is the quantity of output r produced by school k. x_{ik} is the quantity of input i consumed by school k. θ_k represents the VRS efficiency of school k (i.e. 'pure' technical efficiency free from any scale inefficiency). λ_j represents the associated weighting of outputs and inputs of entity j.

The VRS efficiency of the k[th] school is calculated by solving the following linear problem:

Minimize θ_k (1)

Subject to $y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} \leq 0 \qquad r = 1, \ldots, s$

$$\theta_k x_{ik} - \sum_{j=1}^{n} \lambda_j x_{ij} \geq 0 \qquad i = 1, \ldots, m$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \quad \forall j = 1, \ldots, n$$

Banker and Morey (1986a) – One-stage model

The Banker and Morey (1986a) model, also called the categorical model, can be applied when:

- DMUs are grouped into different categories according to the condition of the environment;
- And the environmental variable can be ordered from the least to the most detrimental upon efficiency.

For instance, the 90 primary schools in the State of Geneva are divided into five hierarchical categories (A to E). Schools in category A face the most advantageous environment. Schools in category E face the most detrimental environment. If the measurement of efficiency did not take into account the fact that schools face different environments (i.e. it considered each school to be in the same category), the evaluation would be unfair on the schools facing a difficult environment and too indulgent on the schools facing an advantageous environment.

In the Banker and Morey (1986a) model, 'E' schools are classified as category 1, 'D' schools as category 2, 'C' schools as category 3, 'B' schools as category 4 and 'A' schools as category 5. School efficiency is then evaluated in the following way, using the basic VRS (or constant returns to scale) model:

- Schools in category 1 are only evaluated against schools within this group;
- Schools in category 2 are evaluated with reference to schools in category 1 and 2;
- Schools in category 3 are evaluated with reference to schools in category 1, 2 and 3;

- Schools in category 4 are evaluated with reference to schools in category 1, 2, 3 and 4;
- Finally, schools in category 5 are evaluated with reference to schools in category 1, 2, 3, 4 and 5.

The Banker and Morey (1986a) model evaluates schools under operating handicaps which take into account their particular environments. This ensures that no school is compared to another with a more favourable environment. The VRS formulation of the categorical model is presented in Appendix 1.

Garrett and Kwak (2011) apply the Banker and Morey (1986a) model in the case of 447 school districts in the State of Missouri, USA. They use relative district wealth as the categorical variable with three categorical levels (rich, average and poor).

The main advantage of the Banker and Morey (1986a) model is that it is appropriate for dealing with non-discretionary variables that are qualitative or categorical. Moreover, it is easy to calculate. The method is simple and therefore transparent. There are at least two disadvantages to this approach. First, the various categories have to be ordered hierarchically (from the least to the most favourable). This ordering is not always natural. Second, the Banker and Morey (1986a) model reduces the discriminating power of DEA which depends on the number of entities relative to the number of variables included in the model. As the Banker and Morey (1986a) model considers various sub-samples according to the number of categories, the smaller the sub-sample, the lower the discriminating power between entities that is achieved by DEA (all other things being equal).

Banker and Morey (1986b) – One-stage model

The Banker and Morey (1986b) model directly includes environmental variable(s) as continuous non-discretionary input or output variables in the linear programming formulation. This model takes into account the fact that environmental variables are outside of the control of management and cannot be treated as discretionary factors. As a result, the constraints on the environmental variable are modified. Assuming an input-orientation with variable returns to scale, the inputs are divided into discretionary (x^D) and non-discretionary (x^{ND}) sets. The VRS formulation of the categorical model is presented in Appendix 1.

The environmental variable has to be included as a non-discretionary input or output variable. This implies it is first necessary to decide upon the direction of influence of the environmental variable. Following Coelli *et al.* (2005):

> If the variable is believed to have a positive effect upon efficiency then it should be included in the linear program in the same way as a non-discretionary input would be included. (…). On the other hand, if instead we have a set of 'negative-effect' environmental variables to add to the model then they should be included in the linear program in the same way as a non-discretionary output would be included (p. 192).

Muñiz (2002) strictly applies the Banker and Morey (1986b) model in the context of the education sector. He tests several models considering different non-discretionary variables: percentage of students who usually study more than 10 hours a week; percentage of students who believe that both their parents and teachers have high prospects with regard to their academic future; percentage of students whose annual family income exceeds two and a half million pesetas; percentage of students

who did not change teaching centres in that academic year or in the previous; percentage of students who are only children.

Several studies include non-discretionary variables as inputs or outputs but perform a standard DEA model (i.e. a constant returns to scale or a variable returns to scale model) instead of a Banker and Morey (1986b) model. This leads to biased (if not invalid) results. Examples are found in Garrett and Kwak (2011) or in Diagne (2006).

The main advantage of the Banker and Morey (1986b) model is that it is able to accommodate multiple and continuous non-discretionary variables. However, this approach presents various disadvantages:

- Ruggiero (1996) shows that Banker and Morey's (1986b) model formulation leads to referent points that are not feasible. See Ruggiero (2004, pp. 330-331) for a numerical example.
- The Banker and Morey (1986b) model requires a prior understanding and specification of the direction of influence of the non-discretionary variables.

Assuming that the direction of influence of the non-discretionary variables is understood, the Banker and Morey (1986b) model is easy to calculate. The method is simple and therefore transparent.

Ruggiero (1996) – One-stage model

The Banker and Morey (1986b) model defines efficiency with respect to discretionary variables only. Ruggiero (1996) shows that it leads to referent points that are not feasible. He provides a one-stage model to correct this problem by excluding all entities with a more favorable environment from the evaluation of each entity. The Ruggiero (1996) model is quite similar to the Banker and Morey (1986a) model, with the difference that it allows for continuous environmental variables[11].

As the Ruggiero (1996) model is not applied in this study, its formulation is not presented. See Ruggiero (1996, pp.559-560) for the model specification.

Ruggiero (1996) provides an application of the model to the case of school districts in the State of New York. He uses the percentage of adults with college education as an environmental input.

The main advantages of the Ruggiero (1996) model are that *(1)* it is able to accommodate multiple and continuous non-discretionary variables and that *(2)* it does not lead to non-feasible referent points. However, this approach suffers from various drawbacks:

- Similar to the Banker and Morey (1986b) model, the Ruggiero (1996) model requires a prior understanding and specification of the direction of influence of the non-discretionary variables.
- The Ruggiero (1996) model is not able to consistently handle many non-discretionary variables. As Ruggiero (2004, p. 332) points out

> A potentially more serious problem is the inability to handle many non-discretionary factors. As the number of non-discretionary inputs increases, the probability of overestimating efficiency increases. As a result, inefficient DMUs could be identified as efficient by default. This model does not recognize tradeoffs that exist between the non-discretionary variables; a given DMU under analysis could have a

[11] The Ruggiero (1996) model is an extension of the Banker and Morey (1986a) model that allows for categorical variables. As it allows continous environmental variables, it is comparable to the Banker and Morey (1986b) model (Ruggiero, 1996, p. 555).

> favourable environment because it has favourable levels of most non-discretionary factors but have a limited referent set only because it has an unfavourable level of at least one non-discretionary input (p. 332).

As the Ruggiero (1996) model is not included in a DEA software package, it is not easy to calculate, although the method appears simple and therefore transparent.

Yang and Paradi in Muñiz, Ruggiero, Paradi and Yang (2006, p. 1176) **– One-stage model**

The Yang and Paradi model applies a handicapping measure based on the levels of the non-discretionary variables. Entities with a favourable environment are penalized by the handicapping measure. In such a case, inputs are adjusted with a higher handicap (i.e. they are augmented) and/or outputs are adjusted with a lower handicap (i.e. they are reduced). As a result, adjusted inputs have a higher value than original inputs and adjusted outputs have a lower value than original outputs. The VRS formulation of the Yang and Paradi model is presented in Appendix 1.

Muñiz *et al.* (2006) provide an application of the Yang and Paradi model using simulated data. The decision to adjust data before running a DEA model is supported by Barnum and Gleason (2008).

The main advantage of the Yang and Paradi model is that it does not lessen the discriminating power of DEA, as it does not categorize the entities. The use of handicapping measures presents two disadvantages. First, the direction of influence has to be understood prior to the variables' adjustment. Second, the values of the handicapping measures have to be defined. In most cases, the extent to which the variables have to be augmented or lowered is unclear. In the context of this study, it makes sense to apply the Yang and Paradi model as the handicapping values are known.

Assuming that the handicapping measures h_j and $\hat{h}_j$ have been defined, the Yang and Paradi model is moderately easy to calculate[12]. The method is simple and therefore transparent.

Ray (1991) – Two-stage model

The two-stage model is first introduced by Ray (1988) and further developed by Ray (1991). In the first stage, a basic DEA model (1) is performed using only discretionary variables. After obtaining the technical efficiency scores (TE) from the first stage, Ray (1991) uses an OLS model to regress these scores upon non-discretionary variables in the second stage. The second stage regression is specified as follows:

$$TE_k = \alpha_0 + \beta_1 X_1 + \ldots \beta_v X_v + e_k \qquad (2)$$

[12] Priority education policies or "PEPs" (also known as positive discrimination policies) aim to compensate for the negative impact of environmental variables (mainly socioeconomic status of pupils) on school performance. Such policies have been introduced in the US, England, Belgium, France, Greece, Portugal, Czech Republic, Romania or Sweden (Demeuse, Frandji, Greger & Rochex, 2012). Additional funding allocated to disadvantaged schools is an example of PEPs focused on the institutions. When the additional funding is known, the value of the handicapping measure of the Yang and Paradi model can easily be calculated. This is the case in the context of this study.

The error term represents the efficiency. Since Ray (1991), other types of regression have been used in the second stage. For instance, McCarty and Yaisawarng (1993) are the first to use a Tobit regression.

Applications of the two-stage models in the education sector include Agasisti (2013), Borge and Naper (2006), Burney, Johnes, Al-Enezy and Al-Musallam (2013), Denaux *et al.* (2011), Rassouli-Currier (2007) or Waldo (2007).

According to Coelli (2005, pp. 194-195), the two-stage model presents the advantages of being able to accommodate *(1)* more than one variable and *(2)* both categorical and continuous variables. Moreover, it does not require a prior understanding of the direction of influence of the non-discretionary variables. It is also easy to calculate. The method is simple and therefore transparent. As the second stage introduces a regression analysis, the Ray (1991) model presents the disadvantages inherent to such techniques. Mainly, it requires the specification of a functional form to the regression model. Any misspecification may distort the results. Cordero *et al.* (2009) also point out that the adjustment of efficiency scores takes into account (only) the radial component of inefficiency and not the potential inefficiency derived from slacks[13].

Ruggiero (1998) – Three-stage model

The first two stages of the Ruggiero (1998) model are identical to those used in Ray (1991). In the third stage, the parameters estimated from the second stage regression are used to construct an index for the non-discretionary variables. The following index x^{ND} is considered: $x^{ND} = \sum_{u=1}^{v} \beta_u x_u^{ND}$, where v is the number of non-discretionary variables. The DEA model is run again in the third stage by using the index for the non-discretionary variables to exclude all entities with a more favourable environment from the evaluation of each entity[14].

As the Ruggiero (1998) model is not applied in this study; its formulation is not presented. See Ruggiero (2004, pp. 333-334) for the model specification.

Ruggiero (1998; 2004) provides an application of his three-stage model using simulated data.

The advantages and disadvantages of the Ray (1991) model apply to stage one and two of the Ruggiero (1998) model. An additional disadvantage arises in the third stage, as the efficient entities (on the frontier) are the same as those which would be computed by using a DEA model in which all variables were discretionary. This is the case because the efficiency frontier is the same in both situations. As a result, only the scores of the inefficient entities are modified by the Ruggiero (1998) model. This approach is difficult to calculate. It is sophisticated and therefore not transparent.

---

[13] Efficiency scores generated by DEA are similar with or without the calculation of slacks. In the two-stage method, the coefficients of the regression are calculated towards the efficiency scores as a dependent variable. Their values will be identical whether these scores belong to entities whose inefficiency is composed by only a radial factor or a radial and a slack factor.

[14] The specification of the Ruggiero (1998) three-stage model is therefore similar to the specification of the Ruggiero (1996) one-stage model. It only replaces the original values of the environmental variables in the one-stage model by the index of environmental variables in the third-stage model.

Muñiz (2002) – Three-stage model

The first stage of the Muñiz (2002) model uses model # 1 (with only discretionary variables) to compute technical efficiency scores. Muñiz's (2002) following approach focuses on the slacks, which are added in model # 3 hereafter. Considering output slacks, s_r, and input slacks, s_i, the model can be described:

$$\text{Minimize} \quad \theta_k - \varepsilon \sum_{r=1}^{s} s_r - \varepsilon \sum_{i=1}^{m} s_i \qquad (3)$$

$$\text{Subject to} \quad y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} + s_r = 0 \qquad r = 1, \ldots, s$$

$$\theta_k x_{ik} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i = 0 \qquad i = 1, \ldots, m$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j, s_r, s_i \geq 0 \quad \forall j = 1,...,n; r = 1,...,s; i = 1,...,m$$

Here, ε is a non-Archimedean value defined to be smaller than any positive number, but greater than 0.

In the Muñiz (2002) model, the total slack values in each variable, defined as the sum of the radial and the non-radial movements, are used[15]. By taking into account the total slack, the model avoids losing information from the non-radial movement.

The slacks computed by model # 3 are confounded by the influence of the non-discretionary variables (i.e. the non-discretionary inputs in the input-oriented model), since they have not been included in the first stage.

The objective of the second stage is to distinguish between the slacks associated with *(1)* the real technical inefficiency of the entity and *(2)* the non-discretionary variables. A separate DEA analysis is performed for the slacks of each (discretionary) variable. The model must therefore be run once for each discretionary variable. The slack detected for every entity in a specific variable is used as a variable itself (to be minimized) in the respective DEA models. The objective of this second stage is to minimize the slacks in a discretionary variable subject to the non-discretionary variables. In other words, the second stage determines the minimum amount of slacks achievable by an entity for a specific variable subject to the value of the non-discretionary variables.

To perform the third stage, original data of each entity are adjusted by removing the slack values associated with the non-discretionary variables. The third stage consists of a DEA model which uses the adjusted data values of the discretionary variables. The technical efficiency scores are not

[15] By contrast, the Ruggiero (1998) model only takes the radial movement into account in the second stage.

confounded by the influence of non-discretionary variables anymore, as the slacks calculated in the third stage are due exclusively to the inefficient performance of the entity.

As the Muñiz (2002) model is not applied in this study; its formulation is not presented. See Muñiz (2002, pp. 628-631) for the model specification.

Muñiz (2002) applies his model to 62 high schools in the State of Asturias, Spain. In the second stage, he uses the following variables as non-discretionary inputs: percentage of students who usually study more than 10 hours a week; percentage of students who believe that both their parents and teachers have high prospects with regard to their academic future; percentage of students whose annual family income exceeds two and a half million pesetas; percentage of students who did not change teaching centres in that academic year or in the previous; percentage of students who are only children.

The main advantage of the Muñiz (2002) model is the use of non-parametric techniques in every stage[16]. As a result, no functional form has to be specified. This is useful when the productive process of entities under analysis is unknown. The Muñiz (2002) model also captures information included in the slacks. High cost of time and calculation are the main disadvantages of this approach, which is sophisticated and therefore not transparent.

Fried, Lovell, Schmidt and Yaisawarng (2002) – Three-stage model

The first stage of the Fried *et al.* (2002) model uses model # 3 (with only discretionary variables) to compute technical efficiency scores. The slacks are broadly interpreted as being composed of three effects: the influence of the environment (first effect), inefficiencies due to management (second effect) and statistical noise arising from measurement errors (third effect). The second stage aims to decompose the slacks into these three effects using stochastic frontier analysis (SFA).

In the second stage, the dependent variables are the total input slacks (radial and non-radial movements). They are regressed against the non-discretionary variables (first effect). SFA separates residual into two parts: managerial inefficiencies (second effect) and statistical noise (third effect).

In the third stage, discretionary variables data are adjusted in a manner that accounts for the influence of the environment and statistical noise. The first stage is then repeated using the adjusted data, providing technical efficiency scores devoid of environmental effects and statistical noise.

As the Fried *et al.* (2002) model is not applied in this study; its formulation is not presented. See Fried *et al.* (2002, pp. 160-164) for the model specification.

As far as the author is aware, the Fried *et al.* (2002) model has not been applied to the education sector. Applications of this model are found in Yanyan (2012) with respect to commercial banks; Shang, Hung, Lo and Wang (2008) with respect to hotels or Lee (2008) with respect to paper companies.

The Fried *et al.* (2002) model presents the advantages of being able to accommodate the following into the second stage: *(1)* more than one variable and *(2)* both categorical and continuous variables. Moreover, it does not require a prior understanding of the direction of influence of the non-discretionary variables and it captures the information included in the slacks. As the second stage introduces a SFA, the Fried *et al.* (2002) model presents the disadvantages inherent to such technique.

[16] The use of non-parametric techniques in every stage has also a drawback, as it is sensitive to outliers.

As the residual is separated into an error component and an inefficiency component in SFA, it requires specification of the distributional form of the efficiency component. Any misspecification may distort the results. The Fried *et al.* (2002) model is difficult to calculate and time-consuming. The method is sophisticated and therefore not transparent.

Fried, Schmidt and Yaisawarng (1999) – Four-stage model

The first stage uses model # 3 (with only discretionary variables) to compute technical efficiency scores. In the second stage, the total slacks are regressed upon the environmental variables.In the third stage, the parameters estimated from the second stage regression are used to predict the total input slacks (if the model is input-oriented) or the total output surplus (if the model is output-oriented). These predicted values are used to calculate adjusted values of the original inputs or outputs. In the fourth stage, the DEA model is run again using the adjusted data. It provides technical inefficiency scores devoid of environmental influence.

As the Fried *et al.* (1999) model is not applied in this study; its formulation is not presented. See Fried *et al.* (1999, pp. 252-255) for the model specification.

Sav (2013) and Cordero-Ferrara, Pedraja-Chaparro and Salinas-Jiménez (2008) provide the only two existing applications of the Fried *et al.* (1999) model in the education sector. Sav (2013) measures technical efficiency of 227 universities. Three environmental variables are used in the second stage: the state and local government contribution to public university operating expenses per full-time equivalent student; the number of high school students per 1000 that score at the 80th percentile and above on either the SAT or ACT tests; the number of college freshmen imported from other states relative to the number of resident freshman attending college out-of-state. Cordero-Ferrara *et al.* (2008) measure the efficiency of 80 high-schools in the State of Extramaduria, Spain. They use three non-discretionary components in the second stage. These components are derived from eleven non-discretionary variables using Principal Component Analysis.

The Fried *et al.* (1999) four-stage model presents the advantages of being able to accommodate in the second stage *(1)* more than one variable and *(2)* both categorical and continuous variables. Moreover, it does not require a prior understanding of the direction of influence of the non-discretionary variables. It captures the information included in the slacks. As the second stage introduces a regression analysis, the Fried *et al.* (1999) model presents the disadvantages inherent to such techniques. Mainly, it requires the specification of a functional form to the regression model. Moreover, a significant relationship between the slacks and the environmental variable has to be identified in order to apply this approach. The Fried *et al.* (1999) model is moderately complicated to calculate. The method is sophisticated and therefore not transparent.

Charnes, Cooper and Rhodes (1981) – Program analysis model

The program analysis model developed by Charnes *et al.* (1981) is an alternative approach to the previous ones. Its objective is not to adjust the efficiency scores to the environment but to reveal potential efficiency differences between several 'programs'. The Charnes *et al.* (1981) model consists of three steps.

In the first step, the entire sample is divided into sub-samples of entities facing the same environment (or operating the same 'program'). DEA models are solved for each sub-sample separately. In the second step, all observed data points are projected onto their respective frontiers to 'artificially'

eliminate inefficiency attributed to management. Finally, a single DEA model is run using the data projected values. Note that remaining technical inefficiency can be attributed, in this model, to environmental variables[17].

The first application of the program analysis model in the education sector was produced by Charnes *et al.* (1981). Schools running under the 'Program Follow Through' are compared to schools not running under this program[18]. Other applications include Portela and Thanassoulis (2001), Soteriou, Karahanna, Papanastasiou and Diakourakis (1998) or Diamond and Medewitz (1990). Portela and Thanassoulis (2001) use the program analysis model to assess pupils within schools of the same type and within schools of all types. Soteriou *et al.* (1998) assess the efficiency of secondary schools in Cyprus. They separate schools into two groups operating in an urban or a rural environment. Diamond and Medewitz (1990) assess the efficiency of high-school classes. They consider two categories of classes: in the first one, the Developmental Economic Education Program is applied; in the second one, it is not.

The main advantage of the Charnes *et al.* (1981) model is that it is appropriate for dealing with non-discretionary variables that are qualitative or categorical. Moreover, it can be applied even when there is no natural ordering of the environmental variable. This means that the direction of influence does not need to be specified. The model is easy to calculate. It is simple and therefore transparent. The main disadvantage of the Charnes *et al.* (1981) model is that it lessens the discriminating power of DEA, which depends on the number of entities relative to the number of variables included in the model. As the Charnes *et al.* (1981) model considers various sub-samples, the smaller the sub-sample, the lower the discriminating power between entities that is achieved by DEA (all other things being equal).

5. Comparing the models: a literature review

Various studies have conducted benchmark analysis of alternative methods to measure efficiency (such as COLS, SFA, DEA or Free Disposal Hull). Evidence suggests that the choice of technique affects efficiency scores and rankings of entities. See Johnes (2004, pp. 661-662) for a short review. For instance, Farsi and Filippini (2005) assess the electricity distribution utilities in Switzerland. They study the sensitivity of three benchmarking methods, one being non-parametric and two being parametric: DEA, COLS and SFA. Their results indicate that both efficiency scores and rankings of entities are significantly different across methods. Another example is provided by Badillo and Paradi (1999, p 76-100), who show that diverging results are observed when only non-parametric methods are used, such as DEA and Free Disposal Hull (FDH).

Alternative models to measure efficiency, within DEA, can also lead to diverging results but this has been far less investigated. Whilst few studies address this issue, interest seems to have been growing in recent years.

Some studies (Cordero, Pedraja & Santin, 2009; Estelle, Johnson & Ruggiero, 2010; Harrison, Rouse & Armstrong, 2012; Muñiz *et al.*, 2006; Ruggiero, 1996; Ruggiero, 1998; Ruggiero, 2004) use

[17] This is a major difference between the program analysis model and other models. The remaining technical inefficiency in all other models can be attributed to managerial inefficiency.

[18] The 'Program Follow Through' was launched in 1968 for a period of ten years in the United States as a federally sponsored program providing health, educational and social services to disadvantaged early primary school pupils and their family.

simulated data to compare alternative DEA models to the 'true' efficiency estimates performed by the simulation. However, the objective of these studies is to allow for comparisons between efficiency estimates performed by the alternative models and 'true' efficiency estimates. The objective of these studies is not, therefore, to determine if the efficiency estimates performed by the alternative models are convergent or divergent.

Very few studies (namely Cordero-Ferrara *et al.*, 2008; Muñiz, 2002; Yang and Pollitt, 2009) use empirical data in order to specifically benchmark alternative DEA models[19]. In these studies, comparisons are made between the efficiency estimates of the alternative models.

As practitioners and decision makers tend to perform their own efficiency analysis, the potential issue of diverging results is a matter of concern. If the alternative models do produce convergent results, practitioners and decision makers may confidently select any model. If they produce divergent results, the choice of model becomes a strategic issue.

Studies using simulated and empirical data are presented hereafter.

Simulated data

Cordero, Pedraja and Santín (2009) consider the following models: one-stage by Banker and Morey (1986b), two-stage by Ray (1991) with a Tobit regression, three-stage by Muñiz (2002) and four-stage by Fried *et al.* (1999). Technical efficiency scores of these four methods are compared to a 'true' efficiency measure[20]. The four-stage model obtains the best results, although its Spearman rank's correlation with the 'true' efficiency is moderate (lower than 0.8). Note that the other models have very weak or weak Spearman rank's when the sample of DMUs is small (50). Estelle *et al.* (2010) show that the methodology used for comparison in Cordero *et al.* (2009) is flawed. Results are therefore called into question. Ultimately, Cordero-Ferrara, Pedraja-Chaparro and Salinas-Jiménez (2008) conclude that there is no consensus on the best model to use.

Estelle *et al.* (2010) consider the one-stage Banker and Morey (1986b) model and three variants of the three-stage Ruggiero (1998) model (alternatively using ordinary least squares, fractional logit and non-parametric regression in the second stage). Using simulated data, results are compared to the 'true' efficiency estimates. The three-stage model performs better than the one-stage model according to three indicators: correlation, rank correlation and mean absolute deviation between 'true' and estimated efficiency. The three variants of the Ruggiero (1998) model are very close one to one another. They have a strong correlation and rank correlation with the 'true' efficiency (higher than 0.8).

Harrison, Rouse and Armstrong (2012) use simulated data to compare the standard variable returns to scale (VRS) model (without non-discretionary variables), the one-stage Banker and Morey (1986a) model and the one-stage Banker and Morey (1986b) model with the 'true' efficiency estimates performed by the simulation. Discussing, first, all alternative models which allow for an environmental adjustment, Harrison *et al.* (2012) note that "there is no DEA model that is clearly

[19] Other studies, such as Diagne (2006), McCarty and Yaisawarng (1993) or Garrett and Kwak (2011), perform alternative DEA models using empirical data. However, the objective of these studies is not to compare the models. Moreover, in the three studies cited above, a standard DEA model is performed instead of the Banker and Morey (1986b) model, although non-discretionary variables are included. The results are therefore flawed, rendering invalid comparisons.

[20] 'True' efficiency is determined by a known artificial set of data as the production function, used to simulate data.

superior in controlling for non-discretionary inputs (…)" (p. 263). Considering then the objective of their study, they conclude that *(1)* the Banker and Morey models perform equally well and *(2)* the Banker and Morey models should be used in preference to the standard VRS model when the influence of the environment is moderate to high.

Muñiz *et al.* (2006) use simulated data to compare the one-stage Banker and Morey (1986b) model, the three-stage Muñiz (2002) model, the three-stage Ruggiero (1998) model and the one-stage Yang and Paradi model (in Muñiz, Ruggiero, Paradi and Yang (2006, p. 1176)) with the 'true' efficiency estimates performed by the simulation. Three indicators are used to assess the models' performance: the rank correlation between 'true' and estimated efficiency, the mean absolute deviation and the percentage of entities for which inefficiency is overestimated. The Banker and Morey (1986b) model, using the variable returns to scale assumption, provides a rank correlation close to the Muñiz (2002) model. It does not perform as well as the other models. The Muñiz (2002) model is the second best performer when the number of variables is small, but the results worsen when the number of variables increases. The Ruggiero (1998) model is the best performer in all the cases analyzed except one scenario. Finally, the Yang and Paradi model tends to overestimate inefficiency. It is also negatively affected by an increase in the number of variables.

Ruggiero (1996) uses simulated data to compare the one-stage Ruggiero (1996) model and the one-stage Banker and Morey (1986b) model with the 'true' efficiency estimates performed by the simulation. Based on several indicators (Ruggiero, 1996, p. 561), he shows that the Ruggiero (1996) model performs better than the Banker and Morey (1986b) model. This latter model tends to underestimate efficiency scores. Ruggiero (1996) applies his model to an empirical case of 556 school districts in the State of New York, USA. Unfortunately, he does not run a Banker and Morey (1986b) model with the same data in order to compare the results.

Ruggiero (1998) uses simulated data to compare a standard DEA model (without non-discretionary variables), the one-stage Banker and Morey (1986a) model, the one-stage Ruggiero (1996) model, the one-stage Banker and Morey (1986b) model, the two-stage Ray (1991) model (with two variants in the second stage regression analysis – linear and log-linear) and the three-stage Ruggiero (1998) model (with two variants in the second stage regression – linear and log-linear) with the 'true' efficiency estimates performed by the simulation. Four indicators are used to assess the models' performance: the correlation and the rank correlation between 'true' and estimated efficiency, the mean absolute deviation and the percentage of entities for which efficiency is inferior to the 'true' efficiency. The two- and three-stage models perform better than the one-stage models (including the standard DEA). The Ray (1991) model (both linear and log-linear variants) performs better than all other models based on the correlation and rank correlation criteria. These main results are confirmed by Ruggiero (2004).

Empirical data

Cordero-Ferrara *et al.* (2008) discuss various models including Banker and Morey (1986b), Ruggiero (1996)[21], Ray (1991), Fried *et al.* (2002) and Fried *et al.* (1999). They conclude that "an analysis of the different options does not allow us to conclude that any one is better than the others, that is, none of them is free of constraint" (p. 1329). Cordero-Ferrara *et al.* (2008) also apply a second-stage model, using Tobit regression, and the fourth-stage Fried *et al.* (1999) model to an empirical case (80

[21] This model is wrongly cited as Ruggiero (1998) in Cordero-Ferrara *et al.* (2008, p. 1326).

high-schools in the State of Extramaduria, Spain). They compare the fourth-stage model with a standard DEA model containing only discretionary variables. The rank correlation (Spearman) between the two models is 0.714. The number of efficient schools and the mean efficiency increase in the four-stage model.

Muñiz (2002) uses empirical data on 62 high-schools in the State of Asturias (Spain). He tests a standard DEA model without non-discretionary variables, the Banker and Morey (1986b) model and the three-stage Muñiz (2002) model. The most important finding lies in the number of efficient schools: 5 in the standard model, 12 in the three-stage model and 30 in the one-stage model. Based on the comparison between the one- and the three-stage models, Muñiz (2002) also shows that the majority of schools (75%) present less than 10% difference in efficiency scores. Schools facing a difference of more than 10% are usually efficient in the one-stage model, but not in the third-stage model. These results provide support for the Banker and Morey (1986b) model, as "it can be stated that both classifications present similar results except in the case when part of the units are considered efficient in the one-stage model" (Muñiz, 2002, p. 637). Unfortunately, no Pearson and Spearman correlations are run by Muñiz (2002).

Yang and Pollitt (2009) use empirical data on 221 Chinese coal-fired power-plants. They test a standard DEA model (without non-discretionary variables), the one-stage Banker and Morey (1986b) model, the two-stage Ray (1991) model (with two variants in the second stage regression – Tobit and logistic), the three-stage Fried *et al.* (2002) model and the four-stage Fried *et al.* (1999) model. The Yang and Pollitt (2009) study distinguishes itself from other studies comparing models in the sense that they include undesirable outputs. The fact that the number of non-discretionary variables included in the alternative models is not the same (two in the one-stage model, seven in the other models) must also be noted[22]. Based on the correlations and the rank correlations between the efficiency scores of the alternative models, the following comments can be made:

- The standard DEA model and the third-stage model have a perfect correlation (0.98) and a perfect rank correlation (0.988);
- The standard DEA model and the fourth-stage model have a strong correlation (0.885) and a strong rank correlation (0.889);
- The three- and four-stage models have, in general, a higher correlation with the other models; on this basis, Yang and Pollitt (2009) suggest that "it indicates that these two models are able to explain most of the features of the other models, thus suggesting their superiority" (p. 1104).
- Correlations between the models vary from 0.311 to 0.98; rank correlations vary from 0.441 to 0.988. This suggests that alternative models lead to diverging results.

To sum up

The best available evidence, synthetized in Table 2, suggests that:

- There is no consensus on the best model to use (Cordero-Ferrara *et al.*, 2008);
- The one-stage Banker and Morey (1986a; 1986b) models perform equally well (Harrison *et al.*, 2012);
- The one-stage Banker and Morey (1986b) and the three-stage Muñiz (2002) model present similar results for a majority of entities under assessment (Muñiz, 2002);

[22] As the one-stage Banker and Morey (1986b) model cannot accommodate dummy variables, only two remaining non-discretionary variables were included in it.

- The one-stage Ruggiero (1996) model performs better than the Banker and Morey (1986b) model (Ruggiero, 1996);
- The three-stage Ruggiero (1998) model perform better than the Banker and Morey (1986b) model, the Yang and Paradi model and the Muñiz (2002) model (Muñiz *et al.*, 2006);
- Based on the correlation and the rank correlation criteria, the two-stage Ray (1991) model performs better than the one-stage Banker and Morey (1986b) model, the one-stage Ruggiero (1996) model and the three-stage Ruggiero (1998) model (Ruggiero, 1998); this evidence is confirmed in Ruggiero (2004);
- The three-stage Fried *et al.* (2002) and the four-stage Fried *et al.* (1999) are only compared to other models in Yang and Pollitt (2009). As the models include undesirable outputs and as the number of non-discretionary variables varies across the models, the results of this study cannot be generalized.

Table 2
Models' performance

Best available evidence	Reference
Banker and Morey (1986a) = Banker and Morey (1986b)	Harrison *et al.* (2012)
Banker and Morey (1986b) = Muñiz (2002)	Muñiz (2002)
Ray (1991) > Banker and Morey (1986b)	Ruggiero (1998); Ruggiero (2004)
Ray (1991) > Ruggiero (1996)	Ruggiero (1998); Ruggiero (2004)
Ray (1991) > Ruggiero (1998)	Ruggiero (1998); Ruggiero (2004)
Ruggiero (1996) > Banker and Morey (1986b)	Ruggiero (1996)
Ruggiero (1998) > Banker and Morey (1986b)	Ruggiero (1998)
Ruggiero (1998) > Yang and Paradi (2006)	Ruggiero (1998)
Ruggiero (1998) > Muñiz (2002)	Muñiz *et al.* (2006)

Although the rule of transitivity does not apply, the best evidence available suggests that the Ray (1991) model seems superior to alternative models.

The following critical comments conclude this literature review:

- The indicators used to assess the models' performance in the above mentioned studies (with simulated and empirical data) are probably not sufficient. It would have been wise to add statistical hypothesis tests in order to refine the analyszes. Yang and Pollitt (2009) are the only ones to perform a Wilcoxon-Mann-Whitney test. However, this test does not seem appropriate in their context, and should probably have been substituted by a Wilcoxon signed rank sum test (see Section 8 about it).
- Some studies (Muñiz *et al.*, 2006; Ruggiero, 1996; Ruggiero, 1998; Yang & Pollitt, 2009) do not indicate the level of significance of the correlation coefficients (Pearson and/or Spearman) between the results of the alternative models and the 'true' efficiency measure. As a result, it is difficult to validly take into account their conclusion with respect to these indicators.
- Studies using simulated data merely indicate which is the model whose results are the closest to the 'true' efficiency measure. But they do not indicate whether the convergence between the alternative models (even the 'best' one) and the 'true' efficiency measure is acceptable or not. As a result, it is difficult to draw a general conclusion stating that the alternative models produce convergent or divergent results with the 'true' efficiency measure.

6. Methodology

The choice of alternative models later used in this study is made from a practitioner's standpoint according to three criteria: the degree of sophistication of the models, the level of computational skills needed to perform the models and the inclusion of models in DEA software[23]. Three commercial (*PIM-DEA* ®, *DEA-Solver PRO* ® and *DEAFrontier* ®) and two free (*Win4DEAP* and *Efficiency Measurement System* – EMS –) software packages are considered.

The degree of sophistication is considered as:

- Low for one-stage models which can be performed in existing software;
- Moderate for one-stage models which are not included in existing software and which need, as a result, coding from the practitioners;
- Moderate for two-stage models;
- High for three- and four-stage models.

The level of computational skills is considered as:

- Low if the model can be performed using an existing software;
- Moderate if it requires two different software packages but can easily be performed;
- High if it requires coding or two different software packages and a good command of these packages.

To be retained, a model has to show a low or moderate degree of sophistication, a low or moderate level of computational skills and be included in existing software. Table 3 presents the alternative models according to the three criteria.

[23] Note that Yang and Pollitt (2009, p. 1098) consider the easiness to understand, to apply and to interpret the models as advantages. The simplicity of calculation is considered by Muñiz (2002, p. 632) as another advantage.

Table 3
Ten models are assessed according to their sophistication, the computational skills needed to perform them and their inclusion in existing software

Model	Degree of sophistication	Level of computational skills	Software	Comments
One-stage				
Banker and Morey (1986a)	Low	Low	PIM-DEA ®, DEA-Solver PRO ® Win4DEAP, EMS	The categorical model can be performed in Win4DEAP and EMS by running a separate model for each category
Banker and Morey (1986b)	Low	Low	PIM-DEA ®, DEA-Solver PRO ®, DEAFrontier ®, EMS	
Ruggiero (1996)	Moderate	High		
Yang and Paradi in Muniz *et al.* (2006)	Moderate	Moderate	PIM-DEA ®, DEA-Solver PRO ®, DEAFrontier ®, Win4DEAP, EMS	The data needs first to be adjusted
Two-stage				
Ray (1991)	Moderate	Moderate	First stage in all software packages	The second stage can easily be performed in software like STATA ®, SPSS ® or R; simple and multiple regression can also be run in Excel ®
Three-stage				
Ruggiero (1998)	High	High	First stage in all software packages	The second stage can easily be performed in software like STATA ®, SPSS ® or R; the construction of the index to be used in the third stage requires a good command of these software packages; the third stage cannot be performed in existing software packages
Fried *et al.* (2002)	High	High	First and third stage in all software packages	The second stage can be performed in software like STATA ®, SPSS ® or R; it requires a very good command of these software packages; data need to be adjusted in order to run the third stage
Muniz (2002)	High	High	PIM-DEA ®, DEA-Solver PRO ®, DEAFrontier ®, Win4DEAP, EMS	The DEA model can be run separately at all stages in all existing software; however, the number of models to be run and the necessary adjustement to the data in the second stage means that the Fried *et al.* (2002) model not user-friendly
Four-stage				
Fried et al. (1999)	High	High	First and fourth stage in all software packages	The second and third stages can be performed in software like STATA ®, SPSS © or R; it requires a good command of these sofware packages
Program analysis				
Charnes *et al.* (1981)	Low	Low	PIM-DEA ®, DEA-Solver PRO ®, DEAFrontier ®, Win4DEAP, EMS	A DEA model for each subset has to be run separately; data need to be adjusted before running the last model containing all firms

Five models are retained: one-stage Banker and Morey (1986a) – BM1986a; one-stage Banker and Morey (1986b) – BM1986b; one-stage Yang and Paradi – YP2006; two-stage Ray (1991) – R1991;

program analysis Charnes *et al.* (1981) – C1981. With the exception of YP2006, these models coincidentally correspond to those recommended by Coelli *et al.* (2005, pp. 191-194).

According to the best evidence available, the two-stage model performs better than the one-stage Ruggiero (1996) model and the three-stage Ruggiero (1998) model (Ruggiero, 1998, 2004). As the three-stage Ruggiero (1998) model performs better than the three-stage Muñiz (2002) model, it appears logical to retain the two-stage model.

Although they have been criticized, the one-stage Banker and Morey (1986a) and Banker and Morey (1986b) models are supported by Harrison *et al.* (2012) who note that these models are widely used by researchers. They have generated at least 239 different publications (Löber & Staat, 2010, p. 810). Harrison *et al.* (2010, p. 263) stress that it suggests that many researchers have found these models appropriate for their particular context. They also mention that "given there is no DEA model that is clearly superior in controlling for non-discretionary inputs, researchers continue to refer to the work of Banker and Morey (1986a, b)" (p. 263).

The three-stage Fried *et al.* (2002) model and the four-stage Fried *et al.* (1999) model suffer from a lack of comparison with other models. Yang and Pollitt (2009, p. 1097) clearly considered the three-stage model as the most sophisticated. As the comparison performed by Yang and Pollitt (2009) includes undesirable outputs, it is not clear if the conclusion of their study would have remained the same had the undesirable outputs been excluded. Further comparative studies featuring theses models are therefore needed.

The Charnes *et al.* (1981) program analysis model is retained. Unlike the other models, it estimates a technical efficiency devoid of managerial efficiency. As a result, it cannot be direcly compared to the other models. However, this model is retained to test if its results are, somehow unexpectedly, convergent to the estimates of other models.

Finally, a standard VRS model (without non-discretionary variables) is also retained as a base case – VRS.

The models are compared with one another on the basis of several indicators: mean efficiency, median efficiency, minimum efficiency, maximum efficiency, number of efficient schools, Pearson correlation and Spearman rank correlation. These are standard indicators commonly used in studies comparing models, such Cordero *et al.* (2009), Estelle *et al.* (2010), Harrison *et al.* (2012), Muñiz *et al.* (2006), Ruggiero (1996), Ruggiero (1998), Cordero-Ferrara *et al.* (2008), Muñiz (2002) and Yang and Pollitt (2009). In addition to these indicators, Yang and Pollitt (2009) perform the Wilcoxon-Mann-Whitney test. However, as developed in Section 8, the Wilcoxon-Mann-Whitney test does not seem appropriate in the case described in Yang and Pollitt (2009). As a result, the Wilcoxon signed rank sum test is preferred and retained in this study.

7. Data and models

Database

At the State of Geneva level, information about school input and output are atomized into various databases belonging to different administrative units. Public access to these databases is denied, making information about school production process unknown and opaque. However, cross-sectional data concerning the 2010-2011 school year and the 90 public primary schools has been secured for

this study[24]. It includes pupils' results at standardized tests (aggregated at schools level), the number of full-time equivalent staff and various environmental variables. Useful data had to first be gathered from the different administrative units and second be organized into a workable order.

Discretionary and non-discretionary variables

Three discretionary outputs and three discretionary inputs are considered. These variables are all under the control of headteachers and are aggregated over schools.

Discretionary outputs include three composite scores (on a standardized scale with a maximum of 100) purely reflecting the quality of the education process. The first one is composed of pupils' results in French and mathematics standardized tests at the end of the second grade (SCORE2). The second one is composed of pupils' results in French, German and mathematics standardized tests at the end of the fourth grade (SCORE4). Finally, the third one is composed of pupils' results in French, German and mathematics standardized tests at the end of the sixth grade (SCORE6).

A large part of the studies focus specifically on standardized test scores as outputs[25]. Among those are Bessent and Bessent (1980), Bessent, Bessent, Kennington and Reagan (1982), Bradley *et al.* (2001), Chalos and Cherian (1995), Chalos (1997), Demir and Depren (2010), Kirjavainen and Loikkanen (1998), Mizala, Romaguera and Farren (2002), Ray (1991), Ruggiero (1996, 2000) or Sengupta (1990). Agasisti *et al.* (2014, p. 123) note that "such choice represents today the standard for analyzing school efficiency".

Discretionary inputs include *(1)* the number of full-time equivalent (FTE) teaching staff (TEACHER), *(2)* the number of FTE administrative and technical staff (ADMIN) and *(3)* the school budget in Swiss francs – excluding staff salaries and capital expenditure (BUDGET)[26]. The three inputs are expressed by pupils to be coherent with the formulation of the outputs. Note that BUDGET consists of a (relatively) small financial amount received by schools according to the number and the types of classes it runs. It can be used to finance teachers conducting supplementary tasks (i.e. tasks which do not appear in their contracts) or to buy school materials, support cultural activities, etc.

In 2010, according to the Swiss Federal Statistical Office, the first two inputs (TEACHER and ADMIN) correspond to 94.9% of the public education operating expenses of the State of Geneva (State and local authorities – municipalities)[27]. They are formulated in FTE as opposed to monetary terms given that schools are not responsible for the age pyramid of their teachers and other staff.

---

[24] The 2010-2011 school year is a typical school year. Nothing makes the author think that the results of this study would have been different if another school year was used.

[25] The fact to include variables reflecting other aspects of human capability (and not only test scores) is open to debate. For instance, David Broddy, chairman of the Society of Heads, made the following statement at the Society of Heads' annual meeting in 2013 (Paton, 2013): "What part have we played in allowing that only academic success is a measure of human capability? That a definition of a "good" school is one that rises to the top of exam league tables and the definition of a "bright" pupil is one that gets A* grades?" Unfortunately, in the State of Geneva, such other aspects are either not defined or, if defined, not measured.

[26] Note that test scores from previous school year could be used as an input when longitudinal data is available.

[27] These statistics are available at : http://www.bfs.admin.ch/bfs/portal/fr/index/themen/15/02/data/blank/01.html.

Taking into account the wages of the employees (which automatically grow higher alongside seniority) would unfairly alter the efficiency of a school with a greater proportion of senior staff[28].

The inputs are very similar to those used by Arcelus and Coleman (1997) – FTE teachers, FTE support staff, operating expenses and library expenses – although BUDGET is a feature in this study. The number of teachers and the number of administrative staff are classical inputs (Abbott & Doucouliagos, 2003; Avkiran, 2001; Grosskopf & Moutray, 2001) as are the overhead expenses (Ahn & Seiford, 1993; Beasley, 1990; Chalos & Cherian, 1995; Engert, 1996).

In the State of Geneva, schools are grouped into five categories according to a single non-discretionary variable: the percentage of pupils (per school) whose parents are blue-collar workers or unqualified workers (SOCIO). Note that the positive discrimination policy impacts only TEACHER (see Table 1 in Section 2) and not ADMIN or BUDGET.

Descriptive statistics of the variables are reported in Table 4. For instance, schools in category C have 0.0566 teachers per pupil, 0.0034 administrative staff per pupil and CHF 18.8496 per pupil. Pupils in category C schools obtain 77.9059 points at the end of grade 2, 77.751 at the end of grade 4 and 76.8070 points at the end of grade 6. The SOCIO variable has an average value of 38.15% in category C.

Note that 34 schools out of 90 (37.8%) are not grouped according to their theoretical category. For instance, 34% of pupils at school # 74 are classified as disadvantaged. This school should be in category C, but is actually categorized in B. Several assumptions can explain this observation:

- The State authority has the discretionary power to group schools in other categories despite the value of SOCIO. Out of the 34 schools which are not grouped according to their theoretical category, 26 are grouped in a more advantaged category than the one in which they should be included[29]. For example, 23% of pupils at school # 79 are classified as disadvantaged, indicating that it should be in category B, but is actually categorized in A.
- Headteachers use their negotiation power in order to move their school to a more disadvantaged category than the one in which they should be included. Out of the 34 schools which are not grouped according to their theoretical category, 8 are grouped in a more disadvantaged category[30].

The fact that some schools are not grouped in their theoretical category has an impact on the Banker and Morey (1986a) model and the Charnes *et al.* (1981) model, as these two models are based on entities' categories. In this study, two alternatives are therefore considered:

- In the first one, the Banker and Morey (1986a) model and the Charnes *et al.* (1981) model are based on the observed schools' categorization (BM1986a-O and C1981-O);
- In the second one, the Banker and Morey (1986a) model and the Charnes *et al.* (1981) model are based on the theoretical schools' categorization (BM1986a-T and C1981-T).

---

[28] The question to include wages as an input instead of FTE is open to debate. It would probably be appropriate in a context where schools can freely set teachers' salary. But in a context where teachers' salary is set by public authority and grow automatically alongside seniority, higher wages are not a good proxy of teaching quality. For instance, Woessmann (2003) shows that the teachers' age influences negatively pupil's performance.

[29] As TEACHER depends on the category, this could reflect the State's willingness to minimize expenses. But considerations other than financial could also explain the fact that some schools are not grouped in their theoretical category. For instance, the State authority may have considered that, for other reasons than the socioeconomic statu, some particular schools should be moved to another category.

[30] As TEACHER depends on the category, headteachers have an interest to be in a more disadvantaged category in order to receive more resources.

Table 4
Statistical summary of output and input variables included in the first stage DEA model -Observed category- (sample size = 90 primary schools)

	OBSERVED CATEGORY					
	A	B	C	D	E	Total
Number of schools	15	20	20	15	20	90
OUTPUTS						
SCORE2 (points/pupil)						
Mean	81.1284	80.6277	77.9059	77.9345	76.8063	78.8082
SD	2.3604	4.2687	4.0426	4.6632	5.1538	4.4956
Minimum	76.1504	71.9674	68.9075	69.0868	64.9589	64.9589
Maximum	84.0542	91.9591	83.2465	85.6571	88.8975	91.9591
SCORE4 (points/pupil)						
Mean	80.0865	79.1950	77.7510	75.9859	73.7298	77.2733
SD	2.2735	3.6067	2.8951	3.9605	2.9263	3.8718
Minimum	75.8127	68.9830	72.7422	68.0930	68.9577	68.0930
Maximum	83.4049	87.3654	81.5557	81.3806	78.5661	87.3654
SCORE6 (points/pupil)						
Mean	80.2470	78.6407	76.8070	76.2867	72.4740	76.7382
SD	2.6391	3.8218	3.6879	4.8185	3.6197	4.5361
Minimum	75.6189	70.2255	66.1693	66.2378	64.7010	64.7010
Maximum	85.1323	84.5935	81.4771	85.5275	78.5470	85.5275
INPUTS						
TEACHER (FTE/pupil)						
Mean	0.0558	0.0550	0.0566	0.0581	0.0648	0.0582
SD	0.0018	0.0017	0.0013	0.0018	0.0035	0.0043
Minimum	0.0532	0.0520	0.0546	0.0559	0.0572	0.0520
Maximum	0.0599	0.0583	0.0596	0.0618	0.0689	0.0689
ADMIN (FTE/pupil)						
Mean	0.0035	0.0034	0.0034	0.0035	0.0037	0.0035
SD	0.0005	0.0007	0.0005	0.0004	0.0005	0.0005
Minimum	0.0027	0.0026	0.0026	0.0029	0.0032	0.0026
Maximum	0.0045	0.0052	0.0044	0.0041	0.0050	0.0052
BUDGET (CHF/pupil)						
Mean	22.3694	19.8652	18.8496	19.1546	20.8817	20.1643
SD	6.2819	7.5281	4.0493	5.1942	5.4515	5.8233
Minimum	13.8019	13.2040	8.8186	13.3897	13.6034	8.8186
Maximum	32.1989	48.2835	27.6211	31.3439	33.3575	48.2835
NON-DISC. VARIABLE						
SOCIO						
Mean	19.6000	26.0500	38.1500	46.2000	54.9000	37.4333
SD	3.6801	6.9998	3.7455	3.7455	5.4086	13.7253
Minimum	15.0000	11.0000	29.0000	39.0000	45.0000	11.0000
Maximum	28.0000	37.0000	46.0000	54.0000	64.0000	64.0000

Source: General Direction of Primary Schools, Education Department, State of Geneva.

An unexpected observation emerges from Table 4. The average teacher/pupil ratio is lower in category B (0.055) than in category A (0.0558). Theoretically, it should be higher. This is partially explained by the fact that:

- Eight schools grouped in category A should actually belong to category B as they present a value of SOCIO higher than 19.99%. This implies that the teacher/pupil ratio of category A is pushed upwards.

- Three schools grouped in category B should actually belong to category A as they present a value of SOCIO lower than 20%. This implies that the teacher/pupil ratio of category B is pushed downwards.

Descriptive statistics of the variables based on their theoretically categories are reported in Table 5. The average teacher/pupil ratio in category B (0.0554) is still lower than in category A (0.0557). This means that even when the categories are theoretically (re)composed, other unknown factors influence the allocated quantity of teaching staff[31].

[31] This could simply be due to the fact that the number of teachers cannot be easily adjusted – up- or down – from one school year to the next. For instance, assume that the ratio of teachers to pupils has to be reduced in a school. As the number of pupils is non-discretionary, the State authority has to reduce the number of teachers in this school. However, teachers benefit from the guarantee of employment. Except under exceptional circumstances, they cannot be fired. Neither can they be forced to move to another school. Consequently, if teachers refuse to relocate to another school, the ratio of teachers to pupils cannot be reduced and would thus remain 'artificially high'. This could be the case in category A schools.

Table 5
Statistical summary of output and input variables included in the first stage DEA model -Theoretical category- (sample size = 90 primary schools)

	THEORETICAL CATEGORY					
	A	B	C	D	E	Total
Number of schools	10	18	20	23	19	90
OUTPUTS						
SCORE2 (points/pupil)						
Mean	81.3409	79.8970	79.9912	77.9780	76.2036	78.8082
SD	2.5378	2.9255	4.2877	4.6426	5.3413	4.4956
Minimum	75.7948	74.4785	71.9674	68.9075	64.9589	64.9589
Maximum	84.0542	85.0372	91.9591	85.6571	88.8975	91.9591
SCORE4 (points/pupil)						
Mean	81.5598	79.8333	77.8690	76.0102	73.4941	77.2733
SD	3.2380	1.6527	2.7511	3.2162	3.3763	3.8718
Minimum	76.0190	75.8127	68.9830	69.5880	68.0930	68.0930
Maximum	87.3654	81.8875	80.8669	81.3806	80.2393	87.3654
SCORE6 (points/pupil)						
Mean	81.1046	78.6428	78.0296	76.5137	71.5482	76.7382
SD	2.8699	3.7172	2.4987	4.2160	3.4189	4.5361
Minimum	75.6189	70.2255	73.5402	66.1693	64.7010	64.7010
Maximum	85.1323	82.5055	84.5935	85.5275	78.3456	85.5275
INPUTS						
TEACHER (FTE/pupil)						
Mean	0.0557	0.0554	0.0563	0.0579	0.0646	0.0582
SD	0.0011	0.0020	0.0023	0.0023	0.0039	0.0043
Minimum	0.0543	0.0520	0.0526	0.0551	0.0562	0.0520
Maximum	0.0583	0.0599	0.0616	0.0661	0.0689	0.0689
ADMIN (FTE/pupil)						
Mean	0.0035	0.0036	0.0032	0.0035	0.0036	0.0035
SD	0.0003	0.0008	0.0004	0.0005	0.0004	0.0005
Minimum	0.0031	0.0026	0.0026	0.0027	0.0032	0.0026
Maximum	0.0040	0.0052	0.0041	0.0044	0.0050	0.0052
BUDGET (CHF/pupil)						
Mean	22.6474	20.9626	18.0419	19.7753	20.8063	20.1643
SD	6.6455	8.1456	3.6295	4.7841	5.5989	5.8233
Minimum	13.8019	13.7500	8.8186	13.3897	13.6034	8.8186
Maximum	32.1989	48.2835	23.1899	31.3439	33.3575	48.2835
NON-DISC. VARIABLE						
SOCIO						
Mean	16.1000	23.3889	35.0500	44.4348	56.0000	37.4333
SD	2.3310	3.1086	2.8741	2.8095	4.2817	13.7253
Minimum	11.0000	20.0000	30.0000	40.0000	50.0000	11.0000
Maximum	19.0000	29.0000	39.0000	49.0000	64.0000	64.0000

Source: General Direction of Primary Schools, Education Department, State of Geneva, and own calculation.

Table 6 compares the teacher/pupil target ratio (second column) with the teacher/pupil observed ratio (fourth column) in the two alternatives considered: observed (upper part of the table) versus theoretical (lower part of the table) categorization. The third column of the table recalls the percentage of targeted additional teaching staff that each category should have when compared with category A. For instance, schools in category D should have 11.41% more teaching staff than schools in category A. As these values are target values, they are the same in both alternatives. Finally, the fifth column displays the percentage of real additional teaching staff that each category gets when compared with category A. For instance, schools in category C have 1.49% more teaching staff than

schools in category A when the categorization is observed, but only 1.08% when the categorization is theoretically-based.

Table 6
Teacher/pupil ratio in the observed and in the theoretical categorization

Observed category (# of schools)	Teacher/pupil target ratio	Additional teaching staff per pupil (%) (target)	Teacher/pupil effective ratio	Additional teaching staff per pupil (%) (effective)
A (15)	0.0539	0.00	0.0558	0.00
B (20)	0.0551	2.20	0.0550	-1.34
C (20)	0.0573	6.30	0.0566	1.49
D (15)	0.0601	11.41	0.0581	4.18
E (20)	0.0656	21.64	0.0648	16.15
Theoritical category (# of schools)	**Teacher/pupil target ratio**	**Additional teaching staff per pupil (%) (target)**	**Teacher/pupil effective ratio**	**Additional teaching staff per pupil (%) (effective)**
A (10)	0.0539	0.00	0.0557	0.00
B (18)	0.0551	2.20	0.0554	-0.54
C (20)	0.0573	6.30	0.0563	1.08
D (23)	0.0601	11.41	0.0579	4.01
E (19)	0.0656	21.64	0.0646	15.99

Source: General Direction of Primary Schools, Education Department, State of Geneva, and own calculation.

Except for schools in category B, the observed value of additional teaching staff is higher than the theoretically-based categorization. As a result, the values of the observed categorization are closer to the targeted additional teaching staff values than the theoretical categorization. This could also explain why the observed categorization of schools differs from the theoretical one (i.e. it has been adjusted from the theoretical categorization in order to better reduce the gap towards the targeted values)[32].

The correlation matrix of the input and output variables is presented in Table 7.

Table 7
Correlation Matrix for the variables

	TEACHER	ADMIN	BUDGET	SCORE2	SCORE4	SCORE6	SOCIO
TEACHER	1.00						
ADMIN	0.29 **	1.00					
BUDGET	0.08	-0.10	1.00				
SCORE2	-0.22 *	-0.09	0.07	1.00			
SCORE4	-0.46 **	-0.01	-0.07	0.33 **	1.00		
SCORE6	-0.49 **	-0.09	0.05	0.30 **	0.49 **	1.00	
SOCIO	0.75 **	0.07	-0.04	-0.36 **	-0.67 **	-0.61 **	1.00

** Significant at the 1% level; * Significant at the 5% level

Statistically significant correlations are discussed hereafter. On the input side, the correlation between TEACHER and ADMIN is positive but very weak[33]. On the output side, correlations are positive but

[32] To sum up, the separation between categories is not as complete as might be desired. In addition to the human resources factor, other possible contaminating effects could emerge from pupils' mobility from one school to another during the school year.

[33] Correlation coefficients are considered as perfect between 1 and 0.98 (or -1 and-0.98), strong between 0.97 and 0.8 (or -0.97 and -0.8), moderate between 0.79 and 0.6 (or -0.79 and -0.6), weak between 0.59 and 0.35 (or -0.59 and -0.35) and very weak between 0.34 and 0 (or -0.34 and 0).

very weak between SCORE2 and SCORE 4 (0.33) and between SCORE2 and SCORE6 (0.3) and weak between SCORE4 and SCORE6 (0.49).

Correlations between TEACHER and the discretionary output variables are negative and very weak (TEACHER and SCORE2) or weak (TEACHER and SCORE4, TEACHER and SCORE6). This finding is coherent with Hanushek (2006). Based on a meta-analysis, he shows that school resources are weakly associated with school performance. The fact that the value of the correlation is increasing (or worsening) betweenTEACHER and SCORE2 (–0.22), SCORE4 (–0.46) and SCORE6 (–0.49) is intriguing. A possible interpretation of this result is that the number of teachers matters more in the early grades than in the later grades.

The correlation between the non-discretionary variable SOCIO and TEACHER is positive but only moderate (0.75). This reflects the fact that, despite the positive discrimination policy, the State of Geneva retains discretionary power in the allocation of resources, or that the rigidity in terms of human resource management does not always allow the State authority to increase or reduce the teacher/pupil ratio as desired. Unsurprisingly, correlations between SOCIO and SCORE2, 4 and 6 is negative.

Models

All DEA models are run using a variable returns to scale (VRS) assumption and an input orientation. The free software package Win4DEAP is used to perform all models except the Banker and Morey (1986b) model[34]. For this model, the free package EMS is used[35]. The software STATA ® is used to perform the second stage of the two-stage model.

The standard VRS model is performed without SOCIO. The Banker and Morey (1986a) model, the Charnes *et al.* (1981) model and the Yang and Paradi model are performed according to *(1)* the five observed school categories and *(2)* the five theoretical school categories. SOCIO is included as a continuous non-discretionary variable in the Banker and Morey (1986b) model. In order to allow a coherent comparison, SOCIO is also the only environmental variable considered in the two-stage Ray (1991) model[36]. Finally, note that no bootrstapping procedure is applied[37].

---

[34] As DEAP is a DOS program, a user friendly Windows interface has been developed for it (Win4DEAP). These 'twin' software packages have to be both downloaded and extracted to the same folder. Win4DEAP cannot work without DEAP.
DEAP Version 2.1: http://www.uq.edu.au/economics/cepa/deap.htm
Win4DEAP Version 1.1.3: http://www8.umoncton.ca/umcm-deslierres_michel/dea/install.html

[35] The Banker and Morey (1986b) model is not included in Win4DEAP, but is included in EMS or commercial software packages such as PIM-DEA or DEA-Solver PRO.
EMS: http://www.holger-scheel.de/ems/

[36] Other environmental variables than SOCIO could have been added in the BM1986a, BM1986b and R1991 models. The decision to include only SOCIO is justified by the fact that the State of Geneva uses SOCIO as the only variable in order to model its positive discrimination policy. The results of the models are therefore influenced by a single environmental variable. They could have been different if additional environmental variables had been considered.

[37] Applying a bootstrapping procedure could make sense in the case of the Banker and Morey (1986a) and the Charnes *et al.* (1981) models, as the E category contains a limited number of schools. This option has not been retained because it introduces a supplementary difficulty and sophistication for practitioners and decision makers. Bootstrapping procedures are not included in the basic version of existing software packages, and therefore need coding skills from the practitioners.

Two alternative variants of the Yang and Paradi model are performed. The first variant applies the values of h_j and $\hat{h}_j$ displayed in Table 8 to all inputs and outputs (YP2006-I&O). For instance, schools' inputs in category C are multiplied by a factor of 0.9407; offsetting the additional the 6.3% of resources received by schools C according to the teacher/pupil target ratio (see Table 6); outputs in category C are multiplied by a factor of 1.063; allowing for the 6.3% augmenting of outputs. The second variant applies the values of h_j to all inputs but does not adjust the outputs (YP2006-I)[38].

Table 8
Inputs are multiplied by the h_j factor and outputs by the $\hat{h}_j$ factor

Category	h_j	$\hat{h}_j$
A	1	1
B	0.9784	1.0220
C	0.9407	1.0630
D	0.8976	1.1141
E	0.8221	1.2164

As the positive discrimination policy of the State of Geneva concerns only the number of teaching staff, the handicapping measure in the Yang and Paradi model could be modified in order to be exclusively oriented towards it. This modified model corresponds to an extension of YP2006 and is customized for cases of positive discrimination concerning specific variables in the model. As the original YP2006 model adjusts all discretionary inputs and outputs, the modified model restricts the adjustment exclusively to the variables impacted by the positive discrimination policy (one in the Geneva case). Other discretionary inputs and outputs are not adjusted. As a result, inputs are divided into two categories: inputs impacted by the positive discrimination policy (x_{uj}^{WithPD}) and inputs not impacted by the positive discrimination policy (x_{ij}^{NoPD}). Assume h_j is the handicapping measure to adjust input variables impacted by the positive discrimination policy. The adjusted inputs are $h_j x_{uj}^{WithPD}$. There are data on s outputs, m inputs not impacted by the positive discrimination policy and v inputs impacted by the positive discrimination policy for each of n primary schools to be evaluated. y_{rk} is the quantity of output r produced by school k. x_{uj}^{WithPD} is the quantity of input u consumed by school k. x_{ij}^{NoPD} is the quantity of input i consumed by school k. λ_j represents the associated weighting of outputs and inputs of entity j. θ_k represents the VRS efficiency of school k (i.e. 'pure' technical efficiency free from any scale inefficiency).

[38] The Yang and Paradi model formulation specifies that the handicapping measure is applied to all inputs and/or outputs. In this study, the handicap measure is applied *(1)* to all inputs and outputs and *(2)* to all inputs only, as the positive discrimination policy in the State of Geneva is oriented towards inputs.

This modified model, named Huguenin (H2014), is specified as follows:

$$\text{Minimize} \quad \theta_k \qquad (4)$$

$$\text{Subject to} \quad y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} \leq 0 \qquad r = 1, \ldots, s$$

$$\theta_k x^{NoPD}{}_{ik} - \sum_{j=1}^{n} \lambda_j x_{ij}^{NoPD} \geq 0 \qquad i = 1, \ldots, m$$

$$\theta_k h_k x^{WithPD}{}_{uk} - \sum_{j=1}^{n} \lambda_j h_j x_{uj}^{WithPD} \geq 0 \qquad u = 1, \ldots, v$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \quad \forall j = 1, \ldots, n$$

The handicapping measure h_j takes the values displayed in Table 8 (second column). These values allow TEACHER to be adjusted for the additional staff allocated under the positive discrimination policy. For instance, schools in category D obtain 11.41% additional teaching staff. The handicapping measure of 0.8976 applied for schools in category D allows for the actual number of teaching staff that these schools would have obtained if the positive discrimination policy had not been implemented[39]. As BM1986a, C1981 and YP2006, the Huguenin model is performed according to *(1)* the five observed school categories and *(2)* the five theoretical school categories.

Two-stage model

As only one environmental variable is used as an explanatory variable in the second stage of the Ray (1981) model, no risk of multicollinearity arises. The OLS model takes the following form[40]:

$TE_k = \alpha_0 + \alpha_1 SOCIO + e_k$

TE_k is the gross efficiency score, derived from the first stage analysis, of the k^{th} school and e_k is an error term satisfying the usual conditions for ordinary least squares estimation.

---

[39] It could be criticised that *(1)* the handicapping measure is only applied to inputs impacted by the positive discrimination policy and *(2)* the choice of using the target additional teaching staff values as handicapping measures are questionable. However, these decisions are justified by the fact that the State of Geneva estimates that its positive discrimination policy only impacts one discretionary input (TEACHER) and that the targeted values adequately reflect the environmental influence. It is not the aim of this study to assess the relevancy of these political decisions.

[40] Recent studies have shown that Ordinary Least Squares (OLS) regression is sufficient or even more appropriate to model the efficiency scores (Hoff, 2007; McDonald, 2009). OLS is, therefore, the method of choice in the ensuing study.

The potential presence of heteroskedasticity in the second stage is considered. A Breusch-Pagan / Cook-Weisberg test for heteroskedasticity is performed. It tests the null hypothesis (Ho) that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. If Ho is accepted, it indicates homoskedasticity; if it is rejected, it indicates heteroskedasticity.

The χ^2 of the Breusch-Pagan / Cook-Weisberg test is equal to 7.83 with a p-value of 0.0051. As the p-value is smaller than 0.05, the null hypothesis is rejected indicating that there is significant evidence of heteroskedasticity. Following this result, the model is corrected for heteroskedasticity by running an OLS regression with robust standard errors.

In the second-stage regression, it could be argued that the proportion of disadvantaged pupils increases where pupil performance is poor, and therefore SOCIO is endogenous to school efficiency. Pupil performance (measured by standardized tests) is used as an output in the first stage. All other things being equal, poor performance reduces efficiency. Priviledged parents will move to other neighbourhoods in order to remove their children from low performing schools. As the State of Geneva is facing a housing crisis, with limited housing available and high rental rates, only privileged parents could afford to move into these areas. This move consequently increases the proportion of remaining disadvantaged pupils. As a result, a risk of simultaneity could occur between SOCIO and SCORE2, 4 and 6.

Endogeneity is solved by using instrumental variables. Instruments are identified following the procedure used by Waldo (2007): first, the instruments have to correlate with the potential endogenous variables; second, they must not have any explanatory power on efficiency scores if they are to be used as independent variables alongside the potential endogenous variables.

27 variables are tested in order to identify instruments. These variables are all measured at the municipality level in which schools are located[41]. The potentially endogenous SOCIO variable presents a correlation coefficient above |0.5| with only one variable: social assistance rate (BENEFIT), with a positive correlation of 0.6.

To measure the explanatory power of BENEFIT, an additional model is run. It includes BENEFIT alongside SOCIO. BENEFIT has a coefficient value of minus 0.0002792 (t value of – 0.08) and is not statistically significant. As a result, BENEFIT can be considered as an instrumental variable.

The model tests SOCIO as a potentially endogenous variable, using BENEFIT as an instrument. A Durbin-Wu-Hausman test is performed. The null hypothesis (Ho) states that endogeneity is not present in the model. If Ho is accepted, it indicates the absence of endogeneity; if it is rejected, it indicates that endogeneity exists within the model. The χ^2 of the Durbin-Wu-Hausman test is equal

[41] The 27 variables are as follows: population (2011), population density per km^2,(2011), proportions of the population (2011) between *(1)* 0 and 19 years old, *(2)* 20 and 64 years old, *(3)* over 64 years old, area in km^2 (1992/1997, latest data available), habitat and infrastructure area (%), agricultural area (%), wooded area (%), unproductive area (%), total number of jobs (2008, latest data available), number of jobs in the primary sector, number of jobs in the secondary sector, number of jobs in the tertiary sector, total number of companies (2008, latest data available), number of companies in the primary sector, number of companies in the secondary sector, number of companies in the tertiary sector, number of newly built apartments (2010), social assistance rate (2011), share of votes in the last federal election for left parties (2011), tax burden for married people with two children and an annual revenue of 100'000 CHF (State, municipal and religious tax, in % of gross labour income) (2011), budget surplus (excess revenue) (2011), gross debt (2011), taxable wealth of natural persons (2008, latest data available), taxable income of natural persons (2008, latest data available), taxable profit of corporations (2009, latest data available).

to 0.00638 with a p-value of 0.93635. As the p-value is larger than 0.05, the null hypothesis is accepted. No endogeneity is found.

This is not surprising. In this study, SOCIO is assumed to be the cause of SCORE2, 4 and 6. If information about pupil performance (measured by standardized tests) was public knowledge, it could potentially encourage parents to move into catchment areas of better schools. However, in a principal-agent approach of educational production (Wössmann, 2005), asymmetric information about school data between the principal (i.e. the parents) and the agent (i.e. the headteacher) appears to be strong in the State of Geneva. Information about school quality (pupil performance) and resource consumption are computed at the State level and is unknown by parents. Therefore, parents cannot base their move on rational data and it is unlikely that SOCIO is endogenous.

The procedure of Gasparini and Ramos (2003), applied in De Witte and Moesen (2010) or in Agasisti, Bonomi and Sibiano (2014), is used to derive adjusted net efficiency scores for each school:

$$\theta_k^{Net} = e_k + (1 - \max_{i=1,\ldots,n} e_k)$$

where θ_k^{Net} is the adjusted net efficiency score of of the kth school and e_k is the residual for each school obtained from the OLS estimation.

8. Results

Descriptive statistics

Table 9 displays the descriptive statistics of:

- The standard VRS model;
- The five models which allow for an environmental adjustment (BM1986a; BM1986b; R1991; YP2006; H2014);
- The C1981 model; noting that this model shows efficiency scores devoid of managerial inefficiency but does not adjust for the environment.

The upper part of Table 9 displays results for the observed categories; the lower part displays results for the theoretical categories.

For instance, the YP2006-I&O model has, considering the theoretical categories, a mean efficiency score of 0.9345 with a standard deviation of 0.056. The median efficiency score is 0.9452. This means that half the schools have a score higher than 0.9452 and half the schools have a score lower than 0.9452. In this model, the minimum efficiency score obtained by a school is 0.7976 and the maximum score is 1. 19 schools (row 'Number of efficient schools') are fully efficient.

Table 9
Descriptive statistics of the five models which allow for an environmental adjustment and the two models without environmental adjustment (VRS and C1981)

					I&O			
	VRS	BM1986a	BM1986b	R1991	I&O	I	H2014	C1981
Observed category								
Mean	0.9321	0.9787	0.9793	0.9009	0.9340	0.9654	0.9650	0.9517
SD	0.0671	0.0342	0.0339	0.0450	0.0516	0.0392	0.0401	0.0537
Min.	0.7604	0.8572	0.8415	0.7939	0.7981	0.8556	0.8544	0.7976
Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Median	0.9481	1.0000	1.0000	0.9041	0.9344	0.9763	0.9779	0.9657
Number of efficient schools	20.0000	51.0000	46.0000	1.0000	17.0000	31.0000	31.0000	25.0000
Theoretical category								
Mean	0.9321	0.9751	0.9793	0.9009	0.9345	0.9604	0.9587	0.9553
SD	0.0671	0.0394	0.0339	0.0450	0.0560	0.0455	0.0463	0.0537
Min.	0.7604	0.8482	0.8415	0.7939	0.7976	0.8344	0.8338	0.7813
Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Median	0.9481	1.0000	1.0000	0.9041	0.9452	0.9769	0.9747	0.9774
Number of efficient schools	20.0000	50.0000	46.0000	1.0000	19.0000	29.0000	29.0000	28.0000

Based on Table 9, the following facts are established:

- No obvious difference emerges from the descriptive statistics between results of the observed and the theoretical categories.
- BM1986a and BM1986b have a lower discriminating power than the other models; more than half of the schools are efficient in these models. They have the highest mean efficiency scores amongst all models.
- VRS and YP2006-I&O have close mean efficiency scores.
- YP-I and H2014 have close mean efficiency scores and a similar number of efficients schools.
- H2014 and C1981 have close mean efficiency scores.

Comparison between the observed and the theoretical categorizations

Table 10 displays the Pearson and Spearman correlation coefficients between the same models in the observed and the theoretical categorizations. For instance, the Pearson correlation between the efficiency scores of YP2006-I&O in the observed categorization and the efficiency scores of YP2006-I&O in the theoretical categorization is equal to 0.8022.

Table 10
Pearson and Spearman correlation coefficients between the same models in the observed and the theoretical categorizations

	BM1986a	BM1986b	R1991	YP2006 I&O	YP2006 I	H2014	VRS	C1981
Pearson	0.7696**	1.0000**	1.0000**	0.8022**	0.8343**	0.8736**	1.0000**	0.8834**
Spearman	0.6872**	1.0000**	1.0000**	0.7810**	0.8192**	0.8606**	1.0000**	0.8955**

** Significant at the 1% level; * Significant at the 5% level

As the efficiency scores of BM1986b, R1991 and VRS do not differ in the two categorizations, the Pearson and Spearman correlations are perfect. The efficiency scores of the other models are impacted by the type of categorization. YP2006-I, H2014 and C1981 present strong correlations in both Pearson and Spearman. YP2006-I&O has a strong Pearson correlation and a moderate Spearman correlation. BM1986a have moderate correlations in both Pearson and Spearman.

The difference between the two categorizations is further tested in order to determine whether differences occur by chance or are statistically significant.

Cooper, Seiford and Tone (2007) show that "since the theoretical distribution of the efficiency score in DEA is usually unknown, we are forced to deal with nonparametric statistics for which the distribution of the DEA scores are statistically independent" (p. 233). They use a Wilcoxon-Mann-Whitney test in order to identify whether the differences between two different groups (for instance entities located in an urban environment versus entities located in a rural environment) are significant[42].

Yang and Pollitt (2009, p 1103) use the Wilcoxon-Mann-Whitney test in order to identify whether the difference between the efficiency scores of different models containing the same entities is significant. As the Wilcoxon-Mann-Whitney test seems appropriate in the case described in Cooper *et al.* (2007), it does not seem appropriate in the case described in Yang and Pollitt (2009). For this latter case, the Wilcoxon signed rank sum test seems better suited and is therefore appropriate to test repeated measurements on a single sample (or two related samples or matched samples) in order to assess whether their population mean ranks differ[43]. It therefore seems appropriate to compare efficiency scores of different models containing the same entities. A Wilcoxon signed rank sum test is thus performed between each model in the two categorizations For instance, the efficiency scores of BM1986a in the observed categorization are tested against the efficiency scores of BM1986a in the theoretical categorization.

Table 11 displays results of the Wilcoxon signed rank sum test for the three models which allow for an environmental adjustment and are impacted by the type of categorization (BM1986a, YP2006 and H2014) and for C1981 which is also impacted by the type of categorization but does not allow for an environmental adjustment[44]. The null hypothesis states that there is no statistically significant difference between the efficiency scores of the same model in the two categorization alternatives. The null hypothesis is accepted for BM1986a, YP2006-I&O and C1981 but is rejected for YP2006-I and H2014 at the 1% level[45]. For YP2006-I and H2014, there is a statistically significant difference between the efficiency scores in the observed categorization and the efficiency scores in the theoretical categorization[46].

---

42 The Wilcoxon-Mann-Whitney test is the non-parametric version of the independent samples t-test.

43 The Wilcoxon signed rank sum test is the non-parametric version of the paired samples t-test.

44 The BM1986b and the R1991 models allow for an environmental adjustement but are not impacted by the type of categorization.

45 A Wilcoxon-Mann-Whitney test has also been performed. The null hypothesis is accepted for all models.

46 Note that the Kolmogorov-Smirnov test is another non-parametric hypothesis test used in DEA (Banker, Zheng and Natarajan, 2010). As it compares the distribution of two independent samples (and not repeated measurements on a single sample), it does not seem appropriate in the context of this study. The null hypothesis of the Kolmogorov-Smirnov test states that there is no statistically significant difference between the distribution functions of the same model in the two categorizazions alternatives. For the record, a Kolmogorov-Smirnov test has been performed between each model in the two categorizations. The null hypothesis is accepted for all models.

Table 11
Wilcoxon signed rank sum test

		YP2006			
	BM1986a	I&O	I	H2014	C1981
z-statistic	0.8370	0.4520	2.8200	3.5140	-1.5710
p-value	0.4024	0.6515	0.0048	0.0004	0.1161

Based on Table 10 and Table 11, the following facts are established:

- All correlations between the same models in the two types of categorization are positive and strong, with the exception of BM1986a (Pearson and Spearman correlations) and YP2006-I&O (Spearman correlations), which are also positive but only moderate.
- The null hypothesis of the Wilcoxon signed rank sum test cannot be accepted for all models which allow for an environmental adjustment. As a result, the distinction between the two types of categorization will be kept in the upcoming analysis.

Pearson correlation

Table 12 displays the Pearson correlation coefficients between each pair of models in the **observed** categorization alternative.

Focusing on the five models which allow for an environmental adjustment (the cells in the first six rows and columns shaded in light grey), the correlation coefficients are positive and vary from 0.3301 (BM1986b and YP2006-I&O) to 0.9499 (YP2006-I and H2014). Every single correlation is significant at the 1% level. Three correlations are higher than 0.8 and can be described as strong (BM1986a and H2014; BM1986a and YP2006-I; YP2006-I and H2014). Seven correlations are moderate (0.6787 between BM1986a and R1991; 0.7007 between BM1986b and R1991; 0.6572 between BM1986b and YP2006-I; 0.6434 between BM1986b and H2014; 0.7454 between R1991 and YP2006-I; 0.7837 between R1991 and H2014; 0.6002 between YP2006-I&O and YP2006-I). Finally, five correlations are weak (0.5811 between BM1986a and BM1986b; 0.5762 between BM1986a and YP2006-I&O; 0.3301 between BM1986b and YP2006-I&O; 0.5158 between R1991 and YP2006-I&O; 0.4931 between YP2006-I&O and H2014).

Table 12
Pearson correlation coefficients (observed categorization)

				YP2006				
	BM1986a	BM1986b	R1991	I&O	I	H2014	VRS	C1981
BM1986a	1.0000							
BM1986b	0.5811**	1.0000						
R1991	0.6787**	0.7007**	1.0000					
YP2006-I&O	0.5762**	0.3301**	0.5158**	1.0000				
YP2006-I	0.9201**	0.6572**	0.7454**	0.6002**	1.0000			
H2014	0.8983**	0.6434**	0.7837**	0.4931**	0.9499**	1.0000		
VRS	0.6144**	0.4864**	0.6682**	-0.0003	0.6162**	0.7529**	1.0000	
C1981	0.1679	0.2627*	0.4388**	-0.3321**	0.2214*	0.4036**	0.8778**	1.0000

** Significant at the 1% level; * Significant at the 5% level

The Pearson correlation coefficients between the standard VRS model and the five models which allow for an environmental adjustment are positive and weak (BM1986b), positive and moderate (BM1986a; R1991; YP2006-I; H2014) or negative and very weak (YP2006-I&O). Note that this latter correlation is not statistically significant. VRS and C1981 have a strong positive correlation.

The Pearson correlation coefficients between the C1981 model and the five models which allow for an environmental adjustment are positive and very weak (BM1986a; BM1986b; YP2006-I), positive and weak (R1991; H2014) or negative and very weak (YP2006-I&O). Note that the correlation between C1981 and BM1986a is not statistically significant.

Table 13 displays the Pearson correlation coefficients between each pair of models in the **theoretical** categorization alternative.

Focusing on the five models which allow for an environmental adjustment (the cells in the first six rows and columns shaded in light grey), the correlation coefficients are positive and vary from 0.5050 (YP2006 and H2014) to 0.9652 (YP2006-I and H2014). Every single correlation is significant at the 1% level. Six correlations are higher than 0.8 and can be described as strong (BM1986a and BM1986b; BM1986a and YP2006-I; BM1986a and H2014; R1991 and YP2006-I; R1991 and H2014; YP2006-I and H2014). Seven correlations are moderate (0.7218 between BM1986a and R1991; 0.6238 between BM1986a and YP2006-IO&O; 0.7007 between BM1986b and R1991; 0.7641 between BM1986b and YP2006-I; 0.7343 between BM1986b and H2014; 0.6329 between R1991 and YP2006-I&O; 0.6108 between YP2006-I&O and YP2006-I). Finally, two correlations are weak (0.5414 between BM1986b and YP2006-I&O; 0.5050 between YP2006-I&O and H2014).

Table 13

Pearson correlation coefficients (theoretical categorization)

				YP2006				
	BM1986a	BM1986b	R1991	I&O	I	H2014	VRS	C1981
BM1986a								
BM1986b	0.8145**	1.0000						
R1991	0.7218**	0.7007**	1.0000					
YP2006-I&O	0.6238**	0.5414**	0.6329**	1.0000				
YP2006-I	0.9071**	0.7641**	0.8394**	0.6108**	1.0000			
H2014	0.8693**	0.7343**	0.8306**	0.5050**	0.9652**	1.0000		
VRS	0.5958**	0.4864**	0.6682**	0.0490	0.6935**	0.7914**	1.0000	
C1981	0.0421	0.0392	0.3328**	-0.3584**	0.2303*	0.3765**	0.8238**	1.0000

** Significant at the 1% level; * Significant at the 5% level

The Pearson correlation coefficients between the standard VRS model and the five models which allow for an environmental adjustment are positive and moderate (R1991; YP2006-I; H2014), positive and weak (BM1986a; BM1986b) or positive and very weak (YP2006-I&O). Note that this latter correlation is not statistically significant. VRS and C1981 have a strong positive correlation.

The Pearson correlation coefficients between the C1981 model and the five models which allow for environmental adjustment are positive and very weak (BM1986a; BM1986b; R1991; YP2006-I), positive and weak (H2014) or negative and weak (YP2006-I&O). Note that the correlations between C1981 and BM1986a or BM1986b are not statistically significant.

Based on Table 12 and Table 13, the following facts are established:

- The Pearson correlation between C1981 and the five models which allow for an environmental adjustment is either weak or very weak. This is not a surprise, as C1981 does not adjust for the environment.
- In some cases, the Pearson correlation between VRS and models which allow for an environmental adjustment is moderate and statistically significant. This was not expected, as VRS does not adjust for the environment.
- The Pearson correlations among the five models which allow for an environmental adjustment are positive. However, they are mainly moderate.
- Overall, nine correlations are strong (30%), fourteen are moderate (47%) and seven are weak (23%). The nine strong correlations link the following models: BM1986a and H2014, BM1986a and YP2006-I, YP2006-I and H2014 in the observed categorization; BM1986a and BM1986b, BM1986a and YP2006-I, BM1986a and H2014, R1991 and YP2006-I, R1991 and H2014, YP2006-I and H2014 in the theoretical categorization. Note that BM1986a appears five times in these nine strong correlations.
- The Pearson correlation coefficient analysis is the first indication that the results of a majority of models which allow for an environmental adjustment are divergent. To be considered as convergent, a strong or a perfect correlation would be needed.

Spearman correlation

Table 14 displays the Spearman's rank correlation coefficients between each pair of models in the **observed** categorization alternative.

Focusing on the five models which allow for an environmental adjustment (the cells in the first six rows and columns shaded in light grey), the correlation coefficients are positive and vary from 0.3710 (BM1986b and YP2006-I&O) to 0.9072 (YP2006-I and H2014). Every single correlation is significant at the 1% level. Three correlations are higher than 0.8 and can be described as strong (BM1986a and YP2006-I; BM1986a and H2014; YP2006-I and H2014)[47]. Seven correlations are moderate (0.6022 between BM1986a and R1991; 0.6081 between BM1986a and YP2006-I&O; 0.6010 between BM1986b and R1991; 0.6482 between BM1986b and YP2006-I; 0.6163 between BM1986b and H2014; 0.6638 between R1991 and YP2006-I; 0.7008 between R1991 and H2014). Finally, five correlations are weak (0.5964 between BM1986a and BM1986b; 0.3710 between BM1986b and YP2006-I&O; 0.5178 between R1991 and YP2006-I&O; 0.5848 between YP2006-I&O and YP2006-I; 0.4704 between YP2006-I&O and H2014).

[47] These three pairs of models are also associated with a strong Pearson correlation.

Table 14

Spearman correlation coefficients (observed categorization)

				YP2006				
	BM1986a	BM1986b	R1991	I&O	I	H2014	VRS	C1981
BM1986a	1.0000							
BM1986b	0.5964**	1.0000						
R1991	0.6022**	0.6010**	1.0000					
YP2006-I&O	0.6081**	0.3710**	0.5178**	1.0000				
YP2006-I	0.8508**	0.6482**	0.6638**	0.5848**	1.0000			
H2014	0.8092**	0.6163**	0.7008**	0.4704**	0.9072**	1.0000		
VRS	0.5033**	0.4363**	0.5580**	-0.0183	0.5467	0.7192**	1.0000	
C1981	0.1498	0.1925	0.3377**	-0.2587*	0.2587*	0.4493**	0.8756**	1.0000

** Significant at the 1% level; * Significant at the 5% level

The Spearman's rank correlation coefficients between the standard VRS model and the five models which allow for an environmental adjustment are positive and weak (BM1986a; BM1986b; R1991; YP2006-I), positive and moderate (H2014) or negative and very weak (YP2006-I&O). VRS and C1981 have a strong positive correlation. Note that the correlations between VRS and YP2006-I&O or YP2006-I are not statistically significant.

The Spearman correlation coefficients between the C1981 model and the five models which allow for an environmental adjustment are positive and very weak (BM1986a; BM1986b; R1991; YP2006-I), positive and weak (H2014) or negative and very weak (YP2006-I&O). Note that the correlations between C1981 and BM1986a or BM1986b are not statistically significant.

Table 15 displays the Spearman's rank correlation coefficients between each pair of models in the **theoretical** categorization alternative.

Focusing on the five models which allow for an environmental adjustment (the cells in the first six rows and columns shaded in light grey), the correlation coefficients are positive and vary from 0.4594 (YP2006 and H2014) to 0.9179 (YP2006-I and H2014). Every single correlation is significant at the 1% level. Two correlations are higher than 0.8 and can be described as strong (BM1986a and YP2006-I; YP2006-I and H2014). Eleven correlations are moderate (0.7207 between BM1986a and BM1986b; 0.6576 between BM1986a and R1991; 0.6649 between BM1986a and YP2006-I&O; 0.7986 between BM1986a and H2014; 0.6010 between BM1986b and R1991; 0.6804 between BM1986b and H2014; 0.6353 between R1991 and YP2006-I&O; 0.7752 between R1991 and YP2006-I; 0.7643 between R1991 and H2014; 0.6035 between YP2006-I&O and YP2006-I). Finally, two correlations are weak (0.5257 between BM1986b and YP2006-I&O; 0.4594 between YP2006-I&O and H2014)[48].

[48] These two pairs of models are also associated with a weak Pearson correlation.

Table 15

Spearman correlation coefficients (theoretical categorization)

				YP2006				
	BM1986a	BM1986b	R1991	I&O	I	H2014	VRS	C1981
BM1986a	1.0000							
BM1986b	0.7207**	1.0000						
R1991	0.6576**	0.6010**	1.0000					
YP2006-I&O	0.6649**	0.5257**	0.6353**	1.0000				
YP2006-I	0.8449**	0.7000**	0.7752**	0.6035**	1.0000			
H2014	0.7986**	0.6804**	0.7643**	0.4594**	0.9179**	1.0000		
VRS	0.5169**	0.4363**	0.5580**	0.0244	0.6326**	0.7872**	1.0000	
C1981	0.0817	0.0970	0.2842**	-0.2640*	0.2843**	0.4586**	0.8250**	1.0000

** Significant at the 1% level; * Significant at the 5% level

The Spearman's rank correlation coefficients between the standard VRS model and the five models which allow for an environmental adjustment are positive and moderate (YP2006-I; H2014), positive and weak (BM1986a; BM1986b, R1991) or positive and very weak (YP2006-I&O). Note that this latter correlation is not statistically significant. VRS and C1981 have a strong positive correlation.

The Spearman correlation coefficients between the C1981 model and the five models which allow for an environmental adjustment are positive and very weak (BM1986a; BM1986b; R1991; YP2006-I; H2014) or negative and very weak (YP2006-I&O). Note that the correlations between C1981 and BM1986a or BM1986b are not statistically significant.

Based on Table 14 and Table 15, the following facts are established:

- The Spearman correlation between C1981 and each of the five models which allow for an environmental adjustment is either weak or very weak. This is not a surprise, as C1981 does not adjust for the environment.
- The Spearman correlation between VRS and models which allow for an environmental adjustment is either weak or very weak. It is moderate in only two cases (YP2006-I; H2014).
- The Spearman correlations among the five models which allow for an environmental adjustment are positive. However, they are mainly moderate.
- Overall, eighteen correlations are moderate (60%), seven are weak (23%) and five are strong (17%). The five strong correlations link the following models: BM1986a and YP2006-I, BM1986a and H2014, YP2006-I and H2014 in the observed categorization; BM1986a and YP2006-I, YP2006-I and H2014 in the theoretical categorization. Note that BM1986a appears three times in these five strong correlations.
- After the Pearson correlation analysis, the Spearman's rank correlation coefficient analysis is the second indication that the results of the majority of models which allow for an environmental adjustment are divergent. To be considered convergent, a strong or a perfect correlation would be needed.

Comparison between the models in the observed categorization

The Wilcoxon signed rank sum test is used to assess the difference between each pair of models in the **observed** categorization alternative. For example, the test is performed between the efficiency scores of BM1986a and the efficiency scores of R1991. Results are displayed in Table 16. The first number appearing in a given cell is the z-statistic and the second number is the p-value. For instance, the

Wilcoxon signed rank sum test between BM1986b and H2014 has a z-statistic of 5.124 and a p-value of 0.

Table 16
Wilcoxon signed rank sum test between each pair of models in the observed categorization

				YP2006				
	BM1986a	BM1986b	R1991	I&O	I	H2014	VRS	C1981
BM1986a		-0.1330	-8.2370	-7.8620	-7.3980	-7.3980	-7.2450	-4.2670
		0.8944	0.0000	0.0000	0.000	0.0000	0.0000	0.0000
BM1986b	0.1330		-8.2370	-6.7670	-5.3430	-5.1240	-7.4470	-4.4870
	0.8944		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
R1991	8.2370	8.2370		6.1420	8.2370	8.2370	4.8730	6.8490
	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000
YP2006 I&O	7.8620	6.7670	-6.1420		6.1400	5.5100	0.2800	2.2530
	0.0000	0.0000	0.0000		0.0000	0.0000	0.7797	0.0242
YP2006 I	7.3980	5.3430	-8.2370	-6.1400		1.6220	-4.5950	-1.5930
	0.0000	0.0000	0.0000	0.0000		0.1047	0.0000	0.1111
H2014	7.3980	5.1240	-8.2370	-5.5100	-1.6220		-5.5420	-2.3880
	0.0000	0.0000	0.0000	0.0000	0.1047		0.0000	0.0169
VRS	7.2450	7.4470	-4.8730	-0.2800	4.5950	5.5420		5.6140
	0.0000	0.0000	0.0000	0.7797	0.0000	0.0000		0.0000
C1981	4.2670	4.4870	-6.8490	-2.2530	1.5930	2.3880	-5.6140	
	0.0000	0.0000	0.0000	0.0242	0.1111	0.0169	0.0000	

The null hypothesis is rejected at the 5% level for all but four pairs of models (BM1986a and BM1986b; YP2006-I&O and VRS; YP2006-I and H2014; YP2006-I and C1981)[49]. These four pairs of models appear in light grey cells. For the following pairs of models, there is a statistically significant difference between the efficiency scores in the first model mentioned and the efficiency scores in the second model mentioned: BM1986a and R1991, BM1986a and YP2006-I&O, BM1986a and YP2006-I, BM1986a and H2014, BM1986a and VRS, BM1986a and C1981, BM1986b and R1991, BM1986b and YP2006-I&O, BM1986b and YP2006-I, BM1986b and H2014, BM1986b and VRS, BM1986b and C1981, R1991 and YP2006-I&O, R1991 and YP2006-I, R1991 and H2014, R1991 and VRS, R1991 and C1981, YP2006-I&O and YP2006-I, YP2006-I&O and H2014, YP2006-I&O and C1981, YP2006-I and VRS, H2014 and VRS, H2014 and C1981, VRS and C1981.

Among the four pairs for which the null hypothesis is accepted, only two pairs concern models which exclusively allow for an environmental adjustment (BM1986a and BM1986b; YP2006-I and H2014).

Based on Table 16, the following facts are established:

- Two of the pairs of models which allow for an environmental adjustment do not have a statistically significant difference between their efficiency scores (BM1986a and BM1986b; YP2006-I and H2014). The other pairs of models which allow for an environmental adjustment have a statistically significant difference between their efficiency scores (BM1986a and R1991; BM1986a and YP2006-I&O; BM1986a and YP2006-I; BM1986a and H2014; BM1986b and R1991; BM1986b and YP2006-I&O; BM1986b and YP2006-I; BM1986b and H2014; R1991 and YP2006-I&O; R1991 and YP2006-I; R1991 and H2014; YP2006-I&O and H2014).

[49] Two additional tests have also been performed (a Wilcoxon-Mann-Whitney test and a Kolgomorov-Smirnov test). The results are similar to the Wilcoxon signed rank sum test, except that the null hypothesis is accepted by two additional pairs of models (H2014 and C1981; VRS and C1981).

- The Wilcoxon signed rank sum test performed on every pair of models in the observed categorization is the third indication that the results for the majority of models which allow for an environmental adjustment are divergent.

Comparison between the models in the theoretical categorization

The Wilcoxon signed rank sum test is used to assess the difference between each pair of models in the **theoretical** categorization alternative. For example, the test is performed between the efficiency scores of BM1986b and the efficiency scores of R1991. Results are displayed in Table 17. The first number appearing in a given cell is the z-statistic and the second number is the p-value. For instance, the Wilcoxon signed rank sum test between R1991 and YP2006-I&O has a z-statistic of –6.307 and a p-value of 0.

Table 17
Wilcoxon signed rank sum test between every pair of models in the theoretical categorization

				YP2006				
	BM1986a	BM1986b	R1991	I&O	I	H2014	VRS	C1981
BM1986a		0.4070	-8.2370	-7.8260	-7.4940	-7.3630	-7.1290	-2.7510
		0.6838	0.0000	0.0000	0.0000	0.0000	0.0000	0.0059
BM1986b	-0.4070		-8.2370	-7.1400	-6.4570	-6.4010	-7.3130	-3.3610
	0.6838		0.0000	0.0000	0.0000	0.0000	0.0000	0.0008
R1991	8.2370	8.2370		6.307	8.2370	8.2370	4.885	6.7220
	0.0000	0.0000		0.000	0.0000	0.0000	0.000	0.0000
YP2006 I&O	7.8260	7.1400	-6.307		5.4660	4.5340	0.0720	2.3000
	0.0000	0.0000	0.000		0.0000	0.0000	0.9423	0.0215
YP2006 I	7.4940	6.4570	-8.2370	-5.4660		-0.5250	-4.2400	-0.6920
	0.0000	0.0000	0.0000	0.0000		0.5996	0.0000	0.4888
H2014	7.3630	6.4010	-8.2370	-4.5340	0.5250		-4.9300	-0.8720
	0.0000	0.0000	0.0000	0.0000	0.5996		0.0000	0.3831
VRS	7.1290	7.3130	-4.885	-0.0720	4.2400	4.9300		5.0910
	0.0000	0.0000	0.000	0.9423	0.0000	0.0000		0.0000
C1981	2.7510	3.3610	-6.7220	-2.3000	0.6920	0.8720	-5.0910	
	0.0059	0.0008	0.0000	0.0215	0.4888	0.3831	0.0000	

The null hypothesis is rejected at the 5% level for all but five pairs of models (BM1986a and BM1986b; YP2006-I&O and VRS; YP2006-I and H2014; YP2006-I and C1981; H2014 and C1981)[50]. These five pairs of models appear in light grey cells. Compared to the observed categorization, the null hypothesis is accepted for an additional pair (H2014 and C1981). For the following pairs of models, there is a statistically significant difference between the efficiency scores in the first model mentioned and the efficiency scores in the second model mentioned: BM1986a and R1991, BM1986a and YP2006-I&O, BM1986a and YP2006-I, BM1986a and H2014, BM1986a and VRS, BM1986a and C1981, BM1986b and R1991, BM1986b and YP2006-I&O, BM1986b and YP2006-I, BM1986b and H2014, BM1986b and VRS, BM1986b and C1981, R1991 and YP2006-I&O, R1991 and YP2006-I, R1991 and H2014, R1991 and VRS, R1991 and C1981,

[50] Two additional tests have also been performed (a Wilcoxon-Mann-Whitney test and a Kolgomorov-Smirnov test). The results of the Wilcoxon-Mann-Whitney test are similar to the results of the Wilcoxon signed rank sum test. The results of the Kolgomorov-Smirnov test are similar to the results of the Wilcoxon signed rank sum test, except that the null hypothesis is accepted by two additional pairs of models (H2014 and VRS; VRS and C1981).

YP2006-I&O and YP2006-I, YP2006-I&O and H2014, YP2006-I&O and C1981, YP2006-I and VRS, H2014 and VRS, VRS and C1981.

Among the five pairs for which the null hypothesis is accepted, only two pairs concern models which exclusively allow for an environmental adjustment (BM1986a and BM1986b; YP2006-I and H2014). These pairs are the same identified by the Wilcoxon signed rank sum test in the observed categorization.

Based on Table 17, the following facts are established:

- Two of the pairs of models which allow for an environmental adjustment do not have a statistically significant difference between their efficiency scores (BM1986a and BM1986b; YP2006-I and H2014).
- The other pairs of models which allow for an environmental adjustment have a statistically significant difference between their efficiency scores (BM1986a and R1991; BM1986a and YP2006-I&O; BM1986a and YP2006-I&O; BM1986a and H2014; BM1986b and R1991; BM1986b and YP2006-I&O; BM1986b and YP2006-I; BM1986b and H2014; R1991 and YP2006-I&O; R1991 and YP2006-I; R1991 and H2014; YP2006-I&O and YP2006-I; YP2006-I&O and H2014).
- The Wilcoxon signed rank sum test performed on each pair of models in the theoretical categorization is the fourth indication that the results for the majority of models which allow for an environmental adjustment are divergent.

To sum up

Table 18 sums up the Pearson, Spearman and Wilcoxon signed rank sum analysis. Among the five models which allow for an environmental adjustment (BM1986a, BM1986b, R1991, YP2006 and H2014):

- The results of BM1986a seem to diverge with R1991, YP2006 and H2014 based on the Wilcoxon test. Consequently, the choice of the model (made by politicians or decision makers) impacts school management in terms of schools' input targets and rankings. According to the model selected, the technical efficiency score and the ranking of a particular school are divergent.

  The results of BM1986a and BM1986b seem to converge based on the Wilcoxon test. This finding is in line with Harrison *et al.* (2012) who conclude that both models perform equally well with small or medium sample sizes. However, the Pearson and the Spearman correlations are weak in the observed categorization. From a managerial perspective, the choice of the model is therefore not meaningless in terms of schools' efficiency scores and rankings.

  Figure 1 shows the efficiency scores (in the observed categorization) of BM1986a and BM1986b for each school[51]. Eight schools out of 90 have an efficiency score which differs by more than 5% between the two models. These schools are assigned by their respective numbers on the figure. For instance, school # 11 has an efficiency score of 1 and is equally ranked # 1 ex aequo with the other efficient schools in the BM1986a model[52]; however, it has an efficiency score of 0.8415 and is ranked # 90 in the BM1986b model. For such a school, the choice of the model implies serious managerial consequences. In the BM1986a model, school # 11 is considered efficient and

[51] In order to facilitate comparison, schools are arranged in the figure according to the efficiency scores obtained by the BM1986a model. Note that the Y-axis is truncated at the value of 0.7.

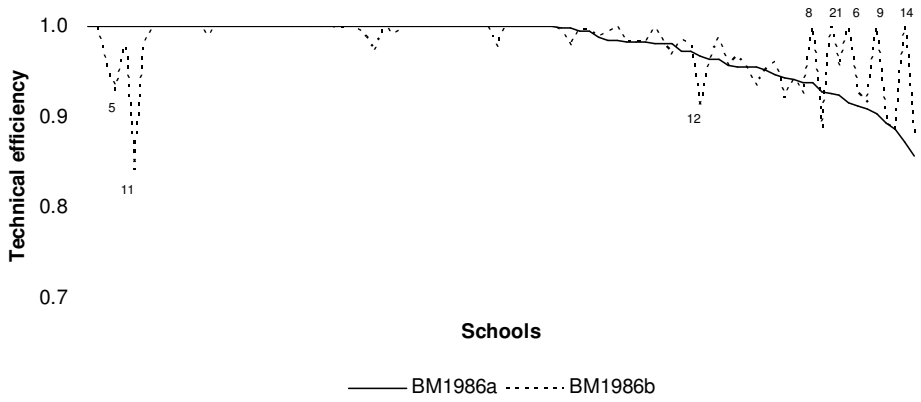
[52] According to the Spearman method of calculating the ranks, school # 11 is ranked # 26 (compared to all the other efficient schools).

is top-ranked. In the BM1986b model, school # 11 should reduce its inputs by 15.85% in order to become efficient and is ranked last.

Among these eight schools, seven are in category E and one in category D (school # 21). Schools # 5, 11 and 12 have a SOCIO value of under 50%. The other five schools have a SOCIO value higher than 60%. Two interpretations can be made. First, it seems that the difference of efficiency scores between BM1986a and BM1986b grows alongside the value of SOCIO. Second, it seems that, among the eight schools, BM1986b tends to allocate a smaller efficiency score, compared to BM1986a, to schools with a relatively small value of SOCIO, and a higher efficiency score to schools with a relatively high value of SOCIO. This is not surprising, as BM1986a does not discriminate among schools in the same category, as opposed to BM1986b, which actually takes into consideration the individual value of SOCIO for each school.

Note that when the eight schools mentioned above are taken out of the sample, the Pearson and the Spearman correlations of the 82 remaining schools have a value of 0.9261 and of 0.8191 respectively. Both correlations are considered as strong and are significant at the 1% level.

Figure 1
Efficiency scores provided by BM1986a and BM1986b for each school

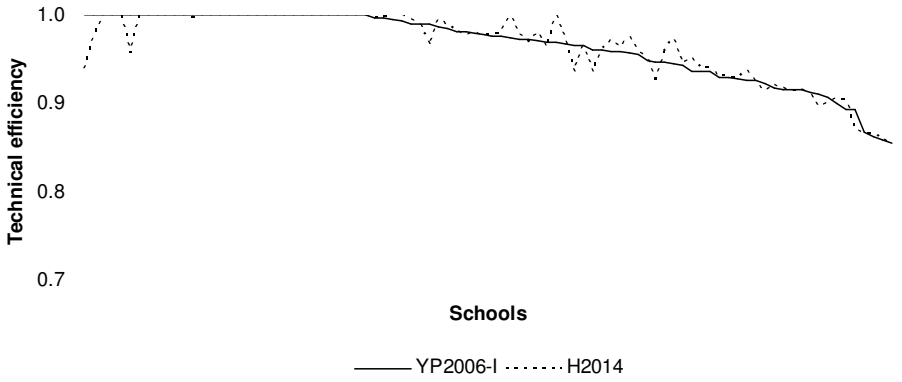


- The results of BM1986b seem to diverge with R1991, YP2006-I&O, YP2006-I and H2014 based on the Wilcoxon test.
- The results of R1991 seem to diverge with BM1986a, BM1986b, YP2006-I&O, YP2006-I and H2014 based on based on the Wilcoxon test.
- The results of YP2006-I&O seem to diverge with BM1986a, BM1986b, R1991, YP2006-I and H2014 based on the Wilcoxon test. The fact that YP2006-I&O and YP2006-I diverge is problematic. It shows that, within the same model, the choice of adjusting inputs and/or outputs lead to different results.
- The results of YP2006-I seem to diverge with BM1986a, BM1986b, R1991 and YP2006-I&O based on the Wilcoxon test. However, they seem to converge with H2014 (see Figure 2[53]). The

[53] In order to facilitate comparison, schools are arranged in the figure according to the efficiency scores obtained by the YP2006-I model. Note that the Y-axis is truncated at the value of 0.7.

converging results between YP2006-I and H2014 are easily understandable, as these two models are very similar. Recall that when H2014 adjusts only the input impacted by the positive discrimination policy (TEACHER), YP2006-I adjusts all inputs, impacted or not by the above mentioned policy (TEACHER, ADMIN and BUDGET).

Figure 2
Efficiency scores provided by YP2006-I and H2014 for each school



- The results of H2014 seem to diverge with BM1986a, BM1986b, R1991 and YP2006-I&O based on the Wilcoxon test. However, they seem to converge with YP2006-I.

In cases of divergence, the choice of the model (made by politicians or decision makers) impacts school management in terms of schools' input targets and rankings. According to the model selected, the technical efficiency score and the ranking of a particular school are divergent.

When the models allowing for an environmental adjustment are compared with the VRS and the C1981 models, the following conclusions are made:

- The results of VRS seem to diverge with BM1986a, BM1986b, R1991, YP2006-I and H2014 based on the Wilcoxon test.

  The VRS results seem to converge with YP2006-I&O based on the Wilcoxon test. As YP2006-I&O adjusts the efficiency scores for the environmental influence and VRS does not, the fact that these two models provide convergent efficiency scores is a counterintuitive result. However, Muñiz *et al.* (2006) show that the YP2006 model tends to overestimate inefficiency (in other words, to underestimate efficiency) when compared to the 'true' efficiency. As a matter of fact, the mean efficiency of the VRS and the YP2006 models is 0.9321 and 0.934 respectively. Among all of the models which allow for an environmental adjustment (except for the R1991 model), the YP2006-I&O model produces the lowest mean efficiency. This could explain why its results are convergent with the VRS results.
- The results of C1981 seem to diverge with BM1986a, BM1986b, R1991 and YP2006-I&O based on the Wilcoxon test. However, they converge with YP2006-I.

  The picture between C1981 and H2014 is not clear. Based on the Wilcoxon test, the results of these two models seem to diverge when the observed categorization is considered; but the results seem to converge when the theoretical categorization is considered. In both cases, the Pearson and Spearman correlations are weak. The convergence in the case of the theoretical categorization is surprising, as H2014 adjusts for the environment and C1981 adjusts for managerial inefficiency. In H2014, efficiency scores are devoid of environmental effects and reveal managerial inefficiency.

In C1981, efficiency scores are devoid of managerial inefficiency and reveal the impact of the environment.

- The results of the Wilcoxon signed rank sum test in the observed and in the theoretical categorizations are convergent for the pairs of models composed exclusively by those allowing for an environmental adjustment.

Interested readers will find a graphical representation of every pair of models in Appendix 2.

Table 18
A diagnostic per model

Model	Pearson	Spearman	Wilcoxon	Diagnostic
BM1986a	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	BM1986b (O+T)	No difference between BM1986a and BM1986b
	Strong: YP2006-I, H2014	Strong: YP2006-I, H2014		Weak Pearson and Spearman correlations with BM1986b (O)
	Moderate: R1991, VRS	Moderate: R1991, YP2006-I&O	**Difference**	Strong Pearson and moderate Spearman correlations with BM1986b (T)
	Weak: BM1986b, YP2006-I&O	Weak: BM1986b, VRS	R1991 (O+T)	
	Very weak: C1981	Very weak: C1981	YP2006-I&O (O+T)	**Conclusion**
			YP2006-I (O+T)	The results of BM1986a and BM1986b seem to converge
	Theoretical categorization (T)	**Theoretical categorization (T)**	H2014 (O+T)	The results of BM1986a and the other models seem to diverge
	Perfect: -	Perfect: -	VRS (O+T)	
	Strong: BM1986b, YP2006-I, H2014	Strong: YP2006-I	C1981 (O+T)	
	Moderate: R1991, YP2006-I&O	Moderate: BM1986b, R1991, YP2006-I&O, H2014		
	Weak: VRS	Weak: VRS		
	Very weak: C1981	Very weak: C1981		
BM1986b	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	BM1986a (O+T)	No difference between BM1986b and BM1986a
	Strong: -	Strong: -		Weak Pearson and Spearman correlations with BM1986a (O)
	Moderate: R1991, YP2006-I, H2014	Moderate: R1991, YP2006-I, H2014	**Difference**	Strong Pearson and moderate Spearman correlations with BM1986a (T)
	Weak: BM1986a, VRS	Weak: BM1986a, YP2006-I&O, VRS	R1991 (O+T)	
	Very weak: YP2006-I&O, C1981	Very weak: C1981	YP2006-I&O (O+T)	**Conclusion**
			YP2006-I (O+T)	The results of BM1986b and BM1986a seem to converge
	Theoretical categorization (T)	**Theoretical categorization (T)**	H2014 (O+T)	The results of BM1986b and the other models seem to diverge
	Perfect: -	Perfect: -	VRS (O+T)	
	Strong: BM1986a	Strong: -	C1981 (O+T)	
	Moderate: R1991, YP2006-I, H2014	Moderate: BM1986a, R1991, YP2006-I, H2014		
	Weak: YP2006-I&O, VRS	Weak: YP2006-I&O, VRS		
	Very weak: C1981	Very weak: C1981		
R1991	**Observed categorization (O)**	**Observed categorization (O)**	**Difference**	**Comments**
	Perfect: -	Perfect: -	BM1986a (O+T)	Difference between R1991 and the other models
	Strong: -	Strong: -	BM1986b (O+T)	
	Moderate: BM1986a, BM1986b, H2014, VRS, YP2006-I	Moderate: BM1986a, BM1986b, YP2006-I, H2014	YP2006-I&O (O+T)	**Conclusion**
	Weak: YP2006-I&O, C1981	Weak: YP2006-I&O, VRS	YP2006-I (O+T)	The results of R1991 and the other models seem to diverge
	Very weak: -	Very weak: C1981	H2014 (O+T)	
			VRS (O+T)	
			C1981 (O+T)	
	Theoretical categorization (T)	**Theoretical categorization (T)**		
	Perfect: -	Perfect: -		
	Strong: YP2006-I, H2014	Strong: -		
	Moderate: BM1986a, BM1986b, YP2006-I&O, VRS	Moderate: BM1986a, BM1986b, YP2006-I&O, YP2006-I, H2014		
	Weak: -	Weak: VRS		
	Very weak: C1981	Very weak: C1981		

Table 18
A diagnostic per model (continued)

Model	Pearson	Spearman	Wilcoxon	Diagnostic
YP2006-I&O	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	VRS (O+T)	No difference between YP2006-I&O and VRS
	Strong: -	Strong: -		Very weak Pearson and Spearman correlations with VRS (O+T)
	Moderate: YP2006-I	Moderate: BM1986a	**Difference**	
	Weak: BM1986a, R1991, H2014	Weak: BM1986b, R1991, YP2006-I, H2014	BM1986a (O+T)	**Conclusion**
	Very weak: BM1986b, VRS, C1981	Very weak: VRS, C1981	BM1986b (O+T)	The results of YP2006-I&O and VRS seem to converge
			R1991 (O+T)	The results of YP2006-I&O and the other models seem to diverge
	Theoretical categorization (T)	**Theoretical categorization (T)**	YP2006-I (O+T)	
	Perfect: -	Perfect: -	H2014 (O+T)	
	Strong: -	Strong: -	C1981 (O+T)	
	Moderate: BM1986a, R1991, YP2006-I	Moderate: BM1986a, R1991, YP2006-I		
	Weak: BM1986b, H2014, C1981	Weak: BM1986b, H2014		
	Very weak: VRS	Very weak: VRS, C1981		
YP2006-I	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	H2014 (O+T)	No difference between YP2006-I and H2014
	Strong: BM1986a, H2014	Strong: BM1986a, H2014	C1981 (O+T)	Strong Pearson and Spearman correlations with H2014 (O+T)
	Moderate: BM1986b, R1991, YP2006-I&O, VRS	Moderate: BM1986b, R1991		
	Weak: -	Weak: YP2006-I&O, VRS	**Difference**	No difference between YP2006-I and C1981
	Very weak: C1981	Very weak: C1981	BM1986a (O+T)	Very weak Pearson and Spearman correlations with C1981 (O+T)
			BM1986b (O+T)	
			R1991 (O+T)	**Conclusion**
	Theoretical categorization (T)	**Theoretical categorization (T)**	YP2006-I&O (O+T)	The results of YP2006-I and H2014 seem to converge
	Perfect: -	Perfect: -	C1981 (O+T)	The results of YP2006-I and C1981 seem to converge
	Strong: BM1986a, R1991, H2014	Strong: BM1986a, H2014		The results of YP2006-I and the other models seem to diverge
	Moderate: BM1986b, YP2006-I&O, VRS	Moderate: BM1986b, R1991, YP2006-I&O, VRS		
	Weak: -	Weak: -		
	Very weak: C1981	Very weak: C1981		
H2014	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	YP2006-I (O+T)	No difference between H2014 and YP2006-I (O+T)
	Strong: BM1986a, YP2006-I	Strong: BM1986a, YP2006-I	C1981 (T)	Strong Pearson and Spearman correlations with H2013 (O+T)
	Moderate: BM1986b, R1991, VRS	Moderate: BM1986b, R1991, VRS		
	Weak: YP2006-I&O, C1981	Weak: YP2006-I&O, C1981	**Difference**	No difference between H2014 and C1981 (T)
	Very weak: -	Very weak: -	BM1986a (O+T)	Weak Pearson and Spearman correlations with C1981 (O+T)
			BM1986b (O+T)	
	Theoretical categorization (T)	**Theoretical categorization (T)**	R1991 (O+T)	**Conclusion**
	Perfect: -	Perfect: -	YP2006-I&O (O+T)	The results of H2014 and YP2006-I seem to converge
	Strong: BM1986a, R1991, YP2006-I	Strong: YP2006-I	VRS (O+T)	The results of H2014 and C1981 seem to converge (T)
	Moderate: BM1986b, VRS	Moderate: BM1986a, BM1986b, R1991, VRS	C1981 (O)	The results of H2014 and C1981 seem to diverge (O)
	Weak: YP2006-I&O, C1981	Weak: YP2006-I&O, C1981		The results of H2014 and the other models seem to diverge
	Very weak: -	Very weak: -		

Table 18
A diagnostic per model (continued)

Model	Pearson	Spearman	Wilcoxon	Diagnostic
VRS	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	YP2006-I&O (O+T)	No difference between VRS and YP2006-I&O
	Strong: C1981	Strong: C1981		Very weak Pearson and Spearman correlations with YP2006-I&O (O+T)
	Moderate: BM1986a, R1991, YP2006-I, H2014	Moderate: H2014	**Difference**	
	Weak: BM1986b	Weak: BM1986a, BM1986b, R1991, YP2006-I	BM1986a (O+T)	**Conclusion**
	Very weak: YP2006-I&O	Very weak: YP2006-I&O	BM1986b (O+T)	The results of VRS and YP2006-I&O seem to converge
			R1991 (O+T)	The results of VRS and the other models seem to diverge
	Theoretical categorization (T)	**Theoretical categorization (T)**	YP2006-I (O+T)	
	Perfect: -	Perfect: -	H2014 (O+T)	
	Strong: C1981	Strong: C1981	C1981 (O+T)	
	Moderate: R1991, YP2006-I, H2014	Moderate: YP2006-I, H2014		
	Weak: BM1986a, BM1986b	Weak: BM1986a, BM1986b, R1991		
	Very weak: YP2006-I&O	Very weak: YP2006-I&O		
C1981	**Observed categorization (O)**	**Observed categorization (O)**	**No difference**	**Comments**
	Perfect: -	Perfect: -	YP2006-I (O+T)	No difference between C1981 and YP2006-I (O+T)
	Strong: VRS	Strong: VRS	H2014 (T)	Very weak Pearson and Spearman correlations with YP2006-I (O+T)
	Moderate: -	Moderate: -		
	Weak: R1991, H2014	Weak: H2014	**Difference**	No difference between C1981 and H2014 (T)
	Very weak: BM1986a, BM1986b, YP2006-I&O, YP2006-I	Very weak: BM1986a, BM1986b, R1991, YP2006-I, YP2006-I&O	BM1986a (O+T)	Weak Pearson and Spearman correlations with H2014 (O+T)
			BM1986b (O+T)	
			R1991 (O+T)	**Conclusion**
	Theoretical categorization (T)	**Theoretical categorization (T)**	YP2006-I&O (O+T)	The results of C1981 and YP2006-I seem to converge
	Perfect: -	Perfect: -	H2014 (O)	The results of C1981 and H2014 seem to converge (T)
	Strong: VRS	Strong: VRS	VRS (O+T)	The results of C1981 and H2014 seem to diverge (O)
	Moderate: -	Moderate: -		The results of C1981 and the other models seem to diverge
	Weak: YP2006-I&O, H2014	Weak: H2014		
	Very weak: BM1986a, BM1986b, R1991, YP2006-I	Very weak: BM1986a, BM1986b, R1991, YP2006-I&O, YP2006-I		

9. Further analysis

This study could be prolonged by several means which are discussed hereafter.

- When dealing with empirical data, the quality of data is a serious concern, especially when a particular variable is used to group entities into different categories. Even when the quality of data seems appropriate, the discretionary power of decision makers could potentially bias the categories. Using different or additional variables to group entities into categories could also potentially modify the results. In the current study, two alternative categorizations have been considered (and tested). 37.8% of schools have been moved from the first categorization (observed) to the second categorization (theoretical). It has been concluded that the results of the models which allow for an environmental adjustment are unaffected by the categorization. Further studies should confirm this conclusion.
- Additional models (three- and four-stage models) could be performed and compared with the models included in the current study. However, as models are compared in pairs, the results of the pairs of models performed in this study would remain the same. It must be noted that this study has positioned itself from the standpoint of practitioners and decision makers. As a result, it has voluntarily omitted some models.
- The Pearson and the Spearman correlations might be influenced by the fact that many schools have efficiency scores equal to one. Table 9 displays the number of efficient schools by model. The BM1986a model identifies the highest number of efficient schools: 51 out of 90 (56.67%) in the observed categorization. The R1991 model identifies the lowest number of efficient schools: 1 out of 90 (1.11%). Table 12 and Table 14 display the Pearson and the Spearman correlations across models in the observed categorization. The variations in correlation coefficients between the models do not seem to be influenced by the number of efficient schools. For instance, the Pearson correlations between BM1986a (51 efficient schools) and the other models are as follows: 0.5811 (BM1986b, 46 efficient schools); 0.6787 (R1991, one efficient school); 0.5762 (YP2006, 17 efficient schools); 0.9201 (YP2006-I, 31 efficient schools); 0.8983 (H2014, 31 efficient schools); 0.6144 (VRS, 20 efficient schools); 0.1679 (C1981, 25 efficient schools)[54].
- In relation to variables, Smith and Mayston (1987) argue the following:

  > The choice and relative importance of outputs is ultimately a political judgement, and no amount of mathematical analysis can reconcile the diversity of views concerning priorities in the public sector. The user of DEA has to recognise this limitation, and at the very least it would seem sensible to test the implications of a variety of output sets (p.188).

- The main findings of the current study indicate that results diverge according to the model performed (with the exception of the BM1986a and BM1986b models and of the YP2006-I and H2014 models). Ultimately, there is no consensus on the best model to use (Cordero-Ferrara *et al.*, 2008). Echoing Smith and Mayston (1987), the choice of model is ultimately a political judgement. Practitioners and decision makers have to select the model which is right *for them*, in other words, the model which best suits their own criteria (not to say the model which best serves their own interests). In this sense, the application of an appropriate multi-criteria decision analysis method to help decision makers select the right model should be investigated in further studies.

[54] In this example, the Pearson correlation between the number of efficient schools and the associated Pearson coefficients is equal to 0.3365.

10. Conclusion

This study tests how several alternative models, within DEA, potentially lead to divergent results. Five models which allow for an environmental adjustment are retained based on their degree of sophistication, their inclusion in existing software and the level of computational skills that they require. These models are the following: Banker and Morey (1986a), Banker and Morey (1986b), Ray (1991), Yang and Paradi in Muñiz *et al.* (2006, p. 1176) and a new model developed in this study called Huguenin (2013). Unlike studies using simulated data to compare efficiency scores from several models, this study uses empirical data concerning 90 primary schools in the State of Geneva, Switzerland. With the exception of Ruggiero (1998), no existing study tests so many models.

The results of the five models are compared on the basis of *(1)* a Pearson and a Spearman correlation analysis and *(2)* a Wilcoxon signed rank sum test analysis. Except for BM1986a and BM1986b and for YP2006-I and H2014, whose results seem to converge, each and every other pair of models (for instance R1991 and BM1986b) provide diverging results. In other words, the efficiency scores generated by the models forming each pair are significantly different. This finding is valid for the specific empirical dataset used in the current study. For this reason, it cannot be generalised to other datasets. However, the fact that the efficiency scores diverge in the current study may suggest that the results obtained from several alternative models may diverge in other cases too.

Applied DEA studies traditionally end with recommendations and policy implications. See for instance McCarty and Yaisawarng (1993, pp. 285-286), Kantabutra and Tang (2006, pp. 370-372) or Jeon and Shields (2008, p. 611). Most of these studies base their recommendations on the efficiency results produced by a particular DEA model. This appears to be problematic. As shown in this study, several alternative models to measure efficiency, within DEA, deliver diverging results. Consequently, recommendations and policy implications may differ according to the model used. From a political standpoint, these diverging results could potentially lead to opposite decisions. From an applied research standpoint, they should represent a serious matter of concern. And from a decision making standpoint, they may lead to opposing managerial choices.

As no consensus emerges on the best model to use, practitioners and decision makers may be tempted to select the model which is right *for them*, in other words, the model which best suits their own criteria and preferences (not to say the model which best serves their own interests). The choice of model thus becomes a strategic issue. Further studies should identify and validate such criteria. Once these criteria are known, the application of an appropriate multi-criteria decision analysis method to help decision makers select the right model should also be investigated.

11. References

Abbott, M. & Doucouliagos, C. (2003). The efficiency of Australian universities: a data envelopment analysis. *Economics of Education Review*, *22*(1), 89-97.

Agasisti, T. (2013). The efficiency of Italian secondary schools and the potential role of competition: a data envelopment analysis using OECD – PISA2006 data. *Education Economics*, *21*(5), 520-544.

Agasisti, T., Bonomi, F. & Sibiano, P. (2014). Measuring the "managerial" efficiency of public schools: a case study in Italy. *International Journal of Educational Management, 28*(2), 120-140.

Ahn, T. & Seiford, L. M. (1993). Sensitivity of data envelopment analysis to models and variable sets in a hypothesis test setting: the efficiency of university operations. In Y. Ijiri (Eds.), *Creative and Innovative Approaches to the Science of Management* (pp. 191-208). Westport: Quorum Books.

Alexander, W. R. J. & Jaforullah, M. (2004). Explaining efficiency differences of New Zealand secondary schools. *Economics Discussion Papers No. 0403*. Dunedin: University of Otago.

Alexander, W. R. J., Haug, A. A. & Jaforullah, M. (2010). A two-stage double-bootstrap data envelopment analysis of efficiency differences of New Zealand secondary schools. *Journal of Productivity Analysis*, *34*(2), 99-110.

Arcelus, F. J. & Coleman, D. F. (1997). An efficiency review of university departments. *International Journal of Systems Science*, *28*(7), 721-729.

Avkiran, N. K. (2001). Investigating technical and scale efficiencies of Australian universities through data envelopment analysis. *Socio-Economic Planning Science*, *35*(1), 57-80.

Badillo, P.-Y. & Paradi, J. C. (1999). *La méthode DEA: analyse des performances*. Paris: HERMES Science Publications.

Banker, R. D., Zheng, Z. & Natarajan R. (2010). DEA-based hypothesis tests for comparing two groups of decision making units. *European Journal of Operational Science*, *206*(1), 231-238.

Banker, R. D., Charnes, A. & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*(9), 1078-1092.

Banker, R. D. & Morey, R. C. (1986a). The Use of Categorical Variables in Data Envelopment Analysis. *Management Science*, *34*(4), 1613-1627.

Banker, R. D. & Morey, R. C. (1986b). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. *Operations Research*, *32*(12), 513-521.

Barnum, D. T. & Gleason, J. M. (2008). Bias and precision in the DEA two-stage method. *Applied Economics*, *40*(18), 2305-2311.

Beasley, J. E. (1990). Comparing university departments. *Omega*, *18*(2), 171-183.

Bessent, A. M. & Bessent, E. W. (1980). Determining the comparative efficiency of schools through data envelopment analysis. *Educational Administration Quarterly*, *16*(2), 57-75.

Bessent, A. M., Bessent, E. W., Kennington, E. W. & Reagan, B. (1982). An application of mathematical programming to assess the productivity in the Houston independent school district. *Management Science*, *28*(12), 1355-1367.

Bifulco, R. & Bretschneider, S. (2001). Estimating school efficiency. A comparison of methods using simulated data. *Economics of Education Review*, *20*(5), 417-422.

Borge, L.-E. & Naper L. R. (2006). Efficiency Potential and Efficiency Variation in Norwegian Lower Secondary Schools. *FinanzArchiv / Public Finance Analysis*, *62*(2), 221-249.

Bradley, S., Johnes, J. & Little, A. (2010). Measurement and determinants of efficiency and productivity in the further education sector in England. *Bulletin of Economic Research*, *62*(1), 1-30.

Bradley, S., Johnes, G. & Millington J. (2001). The effect of competition on the efficiency of secondary schools in England. *European Journal of Operational Research*, *135*(3), 545-568.

Burney, M. A., Johnes, J., Al-Enezi, M. & Al-Mussalam, M. (2013). The efficiency of public schools: the case of Kuwait. *Education Economics*, 21(4), 360-379.

Chalos, P. (1997). An examination of budgetary inefficiency in education using data envelopment analysis. *Financial Accountability and Management*, *13*(1), 55-69.

Chalos, P. & Cherian, J. (1995). An application of data envelopment analysis to public sector performance measurement and accountability. *Journal of Accounting and Public Policy*, *14*(2), 143-160.

Charnes, A., Cooper, W. W., Rhodes, E. (1981). Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through. *Management Science*, *27*(6), 668-697.

Coelli, T. J. (1996). A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program. *CEPA Working Paper 96/08*. Brisbane: Centre for Efficiency and Productivity Analysis, University of Queensland.

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

Cooper, W. W., Seiford, L. M. & Tone, K. (2007). *Data Envelopment Analysis: A comprehensive Text with Models, Applications, References and DEA-Solver Software*. New York: Springer.

Cordero-Ferrara, J. M., Pedraja-Chaparro, F. & Salinas-Jiménez, J. (2008). Measuring efficiency in education: an analysis of different approaches for incorporating non-discretionary inputs. *Applied Economics*, *40*(10), 1323-1339.

Cordero, J. M., Pedraja, F. & Santín, D. (2009). Alternative approaches to include exogenous variables in DEA measures: A comparison using Monte Carlo. *Computers & Operations Research*, *36*(10), 2699-2706.

De Witte, K. & Moesen, W. (2010). Sizing the government. *Public Choice, 145*(1), 39-55.

Demeuse, M., Frandji, D., Greger, D. & Rochex, J.-Y. (2012). *Education Policies and Inequalities in Europe*. New York: Palgrage Macmillan.

Demeuse, M. & Friant, N. (2012). Evaluer les politiques d'éducation prioritaire en Europe : un défi méthodologique. *Revue Suisse des Sciences de l'Education*, *34*(1), 39-55.

Demir, I. & Depren, Ö. (2010). Assessing Turkey's secondary schools performance by different region in 2006. *Procedia – Social and Behavioral Sciences*, *2*(1), 2305-2309.

Denaux, Z. S., Lipscomb, C. A. & Plumly, L. W. (2011). Assessing the technical efficiency of public high schools in the state of Georgia. *Review of Business Research*, *11*(5), 46-57

Diagne, D. (2006). Mesure de l'efficience technique dans le secteur de l'éducation : une application de la méthode DEA. *Revue suisse d'économie et de statistique*, *142*(2), 231-262.

Diamond, A. & Medewitz, J. N. (1990). Use of data envelopment analysis in an evaluation of the efficiency of the DEEP program for economic education. *Journal of Economic Education*, *21*(3), 337-354.

Duncombe, W., Miner, J. & Ruggiero, J. (1997). Empirical evaluation of bureaucratic models of inefficiency. *Public Choice*, *93*(1), 1-18.

Emrouznejad, A., Parker, B. R. & Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences*, *42*(3), 151-157.

Engert, F. (1996). The reporting of school district efficiency: the adequacy of ratio measures'. *Public Budgeting and Financial Management*, *8*(2), 247-271.

Estelle, S. M., Johnson, A. L. & Ruggiero, J. (2010). Three-stage DEA models for incorporating exogenous inputs. *Computers & Operations Research*, *37*(6), 1087-1090.

Farsi, M. & Filippini, M. (2005). A benchmarking analysis of electricity distribution utilities in Switzerland. *CEPE Working Paper No. 43*. Zurich: Center for Energy Policy and Economics, Swiss Federal Institute of Technology.

Frandji, D. (2008). Pour une comparaison des politiques d'éducation prioritaire en Europe. In M. Demeuse, D. Frandji, D. Greger & J.-Y. Rochex (Eds.), *Les politiques d'éducation prioritaire en Europe: Conceptions, mises en oeuvre, débats* (pp. 9-34). Lyon : Institut national de Recherche pédagogique.

Fried, H. O., Lovell, C. A. K., Schmidt, S. S. & Yaisawarng, S. (2002). Accounting for environmental effects and statistical noise in data envelopment analysis. *Journal of Productivity Analysis*, *17*(1/2), 157-174.

Fried, H. O., Schmidt, S. S. & Yaisawarng, S. (1999). Incoporating the Operating Environment Into a Nonparametric Measure of Technical Efficiency. *Journal of Productivity Analysis*, *12*(3), 249-267.

Garrett, W. A. & Kwak, N. K. (2011). Performance comparisons of Missouri public schools using data envelopment analysis. In K. D. Lawrence & G. Kleinman (Eds.), *Applications in Multicriteria Decision Making, Data Envelopment Analysis, and Finance (Applications of Management Science, Volume 14)* (pp. 135-155). Bingley: Emerald Group Publishing Limited.

Gasparini, C. & Ramos, F. (2003). Efetividade e eficiencia no ensino medio brasileiro. *Economia Aplicada, 7*(2), 389-411.

Grosskopf, S. & Moutray, C. (2001). Evaluating performance in Chicago public high schools in the wake of decentralization. *Economics of Education Review*, *20*(1), 1-14.

Hanushek, E. A. (2006). School Resources. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education, Volume 2* (pp. 865-903). Amsterdam: North-Holland.

Harrison, J., Rouse, P. & Armstrong, J. (2012). Categorical and continuous non-discretionary variables in data envelopment analysis: a comparison of two single-stage models. *Journal of Productivity Analysis*, *37*(3), 261-276.

Hoff, A. (2007). Second stage DEA: Comparison of approaches for modelling the DEA score. *European Journal of Operational Research*, *181*(1), 425-435.

Huguenin, J.-M. (2012). Data Envelopment Analysis (DEA): a pedagogical guide for decision makers in the public sector. *Cahier de l'IDHEAP No. 276*. Lausanne: Swiss Graduate School of Public Administration.

Huguenin, J.-M. (2013a). Data Envelopment Analysis (DEA): un guide pédagogique à l'intention des décideurs dans le secteur public. *Cahier de l'IDHEAP No. 278*. Lausanne : Institut de hautes études en administration publique.

Huguenin, J.-M. (2013b). Data Envelopment Analysis (DEA). In A. Ishizaka & P. Nemery (Eds.), *Multi-Criteria Decision Analysis: Methods and Software* (pp. 235-274). Chichester: John Wiley & Sons.

Jeon, Y. and Shields, M. P. (2008). Integration and utilization of public education resources in remote and homogenous areas: a case study of the upper peninsula of Michigan. *Contemporary Economics Policy*, *23*(4), 601-614.

Johnes, J. (2004). Efficiency measurement. In G. Johnes & J. Johnes (Eds.), *International Handbook on the Economics of Education* (pp. 613-742). Cheltenham: Edward Elgar Publishing.

Kantabutra, S. & Tang, J. C. S. (2006). Urban-rural and size effects on school efficiency: The case of Northern Thailand. *Leadership and Policy in Schools*, *5*(4), 355-377.

Kirjavainen, T. & Loikkanen, H. A. (1998). Efficiency Differences of Finnish Senior Secondary Schools: An Application of DEA and Tobit Analysis. *Economics of Education Review*, *17*(4), 377-394.

Lee, J.-Y. (2008). Application of the three-stage DEA in measuring efficiency – an empirical evidence. *Applied Economics Letters*, *15*(1), 49-52.

Löber, G. & Staat, M. (2010). Integrating categorical variables in Data Envelopment Analysis models: A simple solution technique. *European Journal of Operational Research*, *202*(3), 810-818.

Lovell, C. A. K., Walters, L. C. & Wood, L. L. (1994). Stratified models of education production using modified data envelopment analysis and regression analysis. In A. Charnes, W. W. Cooper, A. Y. Lewin & L. M. Seiford (Eds.), *Data envelopment analysis: Theory, methodology and applications* (pp. 329-351). Dordrecht: Kluwer Academic.

McCarty, T. A. & Yaisawarng, S. (1993). Technical efficiency in New Jersey school districts. In H. O. Fried, C. A. K Lovell & S. S. Schmidt (Eds.), *The Measurement of Productive Efficiency: Techniques and Applications* (pp. 271-287). New York: Oxford University Press.

McDonald, J. (2009). Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research*, *197*(2), 792-798.

Mizala, A., Romaguera, P. & Farren, D. (2002). The technical efficiency of schools in Chile. *Applied Economics*, *34*(12), 1533-1552.

Muñiz, M. A. (2002). Separating managerial inefficiency and external conditions in data envelopment analysis. *European Journal of Operational Research*, *143*(3), 625-643.

Muñiz, M., Paradi, J., Ruggiero, J. & Yang, Z. (2006). Evaluating alternative DEA models used to control for non-discretionary inputs. *Computers & Operations Research*, *33*(5), 1173-1183.

Observatory on Primary Education (2010). *Allocation des ressources aux établissements* (Report of December 2010). Geneva: General Direction of Primary Schools, Education Department, State of Geneva.

Ouellette, P. & Vierstraete, V. (2005). An evaluation of the efficiency of Québec's school boards using the Data Envelopment Analysis method. *Applied Economics*, *37*(14), 1643-1653.

Organisation for Economic Co-operation and Development (2001). *Measuring Productivity: Measurement of Aggregate and Industry-level Productivity Growth*. Paris: OECD.

Paton, G. (2013, March 4). Schoolchildren losing the power to concentrate in class. *The Daily Telegraph*.

Portela, M. C. A. S., Camanho, A. S. & Borges, D. N. (2011). BESP – benchmarking of Portuguese secondary schools. *Benchmarking: An International Journal*, *18*(2), 240-260.

Portela, M. C. A. S. & Thanassoulis, E. (2001). Decomposing school and school type efficiency. *European Journal of Operational Research*, *132*(2), 114-130.

Ramanathan, R. (2001). A Data Envelopment Analysis of Comparative Performance of Schools in the Netherlands. *Opsearch*, *38*(2), 160-182.

Rassouli-Currier, S. (2007). Assessing the Efficiency of Oklahoma Public Schools: A Data Envelopment Analysis. *Southwestern Economic Review*, *34*(1), 131-144.

Ray, S. C. (1988). Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation. *Socio-economic Planning Sciences*, *22*(4), 167-176.

Ray, S. C. (1991). Resource-use efficiency in public schools: a study of Connecticut data. *Management Science*, *37*(12), 1620-1628.

Ruggiero, J. (1996). On the measurement of technical efficiency in the public sector. *European Journal of Operational Research*, *90*(3), 553-565.

Ruggiero, J. (1998). Non-discretionary inputs in data envelopment analysis. *European Journal of Operational Research*, *111*(3), 461-469.

Ruggiero, J. (2000). Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. *Journal of the Operational Research Society*, *51*(8), 906-912.

Ruggiero, J. (2004). Performance Evaluation in Education. In W. W. Cooper, L. M. Seiford & J. Zhu (Eds.), *Handbook on Data Envelopment Analysis* (pp. 323-346). Dordrecht: Springer.

Ruggiero, J. & Vitaliano, D. F. (1999). Assessing the efficiency of public schools using data envelopment analysis and frontier regression. *Contemporary Economic Policy*, *17*(3), 321-331.

Sav, G. T. (2013). Four-Stage DEA Efficiency Evaluations: Financial Reforms in Public University Funding. *International Journal of Economics and Finance*, *5*(1), 24-33.

Sengupta, J. K. (1990). Tests of efficiency in DEA. *Computers Operations Research*, *17*(2), 123-132.

Shang, J.-K., Hung, W.-T., Lo, C.-F. & Wang, F.-C. (2008). Ecommerce and hotel performance: three-stage DEA analysis. *The Service Industries Journal*, *28*(4), 529-540.

Smith, P. & Mayston, D. (1987). Measuring efficiency in the public sector. *Omega*, *15*(3), 181-189.

Soteriou, A. C., Karahanna, E., Papanastasiou, C. & Diakourakis, M. S. (1998). Using DEA to evaluate the efficiency of secondary schools: the case of Cyprus. *International Journal of Educational Management*, *12*(2), 65-73.

Syrjänen, M. J. (2004). Non-discretionary and discretionary factors and scale in data envelopment analysis. *European Journal of Operational Research*, *158*(1), 20-33.

Tavares, G. (2002). A bibliography of Data Envelopment Analysis (1978-2001). *Rutcor Research Report No. 01-02*. Piscataway: Rutgers University.

Thanassoulis, E. (1996). Altering the Bias in Differential School Effectiveness Using Data Envelopment Analysis. *The Journal of the Operational Research Society*, *47*(7), 882-894.

Thanassoulis, E., Portela, M. C. S. & Despic, O. (2008). Data Envelopment Analysis: The Mathematical Programming Approach to Efficiency Analysis. In H. O. Fried, C. A. K. Lovell & S. S. Schmidt (Eds.), *The Measurement of Productive Efficiency and Productivity Growth* (pp. 251-420). Oxford: Oxford University Press.

Viger, G. (2007, June). *L'analyse comparative au service de l'amélioration et de la performance.* Note de cadrage presented at the third meeting of the Contrôle de Gestion des Programmes.

Waldo, S. (2007). Efficiency in Swedish Public Education: Competition and Voter Monitoring. *Education Economics*, *15*(2), 231-251.

Woessmann, L. (2003). Schooling Resources, Educational Institutions and Student Performance: the International Evidece. *Oxford Bulletin of Economics and Statistics*, *65*(2), 117-170.

Wössmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, *13*(2), 143-169.

Yang, H. & Pollitt, M. (2009). Incorporating both undesirable outputs and uncontrollable variables into DEA: The performance of Chinese coal-fired power plants. *European Journal of Operational Research*, *197*(3), 1095-1105.

Yanyan, P. (2012). Commercial Bank Branch Efficiencies Based on Three-Stage DEA Model. In L. Zhang & C. Zhang (Eds.), *Engineering Education and Management* (pp. 465-470). Berlin: Springer.

12. Appendix 1

Banker and Morey (1986a) – One-stage model

The VRS formulation of the categorical model is specified as follows:

Minimize θ_k (5)

Subject to $y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} \leq 0 \qquad r = 1, \ldots, s$

$\theta_k x_{ik}^D - \sum_{j=1}^{n} \lambda_j x_{ij}^D \geq 0 \qquad i = 1, \ldots, m$

$\sum_{j=1}^{n} \lambda_j d_{rj}^{(Cr)} \leq d_{rk}^{(Cr)} \qquad r = 1, ..., R$

$C\, r = 1, ..., C-1$

$\sum_{j=1}^{n} \lambda_j = 1$

$\lambda_j \geq 0 \quad \forall j = 1, \ldots, n$

The third set of constraints corresponds to an index of dummy variables d_{rk}^{Cr} representing categories of the environment. C represents the category level (e.g. school category C) and r represents the category variable (where there is more than one category variable). In the case of the State of Geneva, there is only one category variable (SOCIO). For example, if there are five category levels (A to E), this can be coded using four dummy variables where:

- $d^{(1)}$ equals zero for schools in category E and one for schools in category D, C, B and A;
- $d^{(2)}$ equals zero for schools in category E and D and one for schools in category C, B and A;
- $d^{(3)}$ equals zero for schools in category E, D and C and one for schools in category B and A;

$d^{(4)}$ equals zero for schools in category E, D, C and B and one for schools in category A.

Banker and Morey (1986b) – One-stage model

The VRS formulation of the Banker and Morey (1986b) model is specified as follows:

$$\text{Minimize} \quad \theta_k \tag{6}$$

$$\text{Subject to} \quad y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} \leq 0 \qquad r = 1, \ldots, s$$

$$\theta_k x_{ik}^{D} - \sum_{j=1}^{n} \lambda_j x_{ij}^{D} \geq 0 \qquad i = 1, \ldots, m$$

$$x_{uk}^{ND} - \sum_{j=1}^{n} \lambda_j x_{uj}^{ND} \geq 0 \qquad u = 1, \ldots, v$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \quad \forall j = 1, \ldots, n$$

In the above model, an additional constraint is included for each non-discretionary input (x^{ND}). These constraints are similar to the constraints for the discretionary inputs (x^{D}) with the exception that the efficiency component is not included. As a result, the efficiency is defined with respect to the discretionary inputs only.

Yang and Paradi model in Muñiz, Ruggiero, Paradi and Yang (2006, p. 1176) **– One-stage model**

Assume h_j is the handicapping measure to adjust input variables and $\hat{h}_j$ the handicapping measure to adjust output variables. The adjusted input is $h_j x_{ij}$ and the adjusted output is $\hat{h}_j y_{rj}$.The model is specified as follows:

Minimize θ_k (7)

Subject to $\hat{h}_k y_{rk} - \sum_{j=1}^{n} \lambda_j \hat{h}_j y_{rj} \leq 0 \quad r = 1, \ldots, s$

$$\theta_k h_k x_{ik} - \sum_{j=1}^{n} \lambda_j h_j x_{ij} \geq 0 \quad i = 1, \ldots, m$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \quad \forall j = 1, \ldots, n$$

Charnes, Cooper and Rhodes (1981) – Program analysis model

Charnes *et al.* (1981) use a constant returns to scale model to assess efficiency. This model is defined as follows:

Minimize θ_k (8)

Subject to $y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} \leq 0 \qquad r = 1, \ldots, s$

$$\theta_k x_{ik} - \sum_{j=1}^{n} \lambda_j x_{ij} \geq 0 \qquad i = 1, \ldots, m$$

$$\lambda_j \geq 0 \quad \forall j = 1, \ldots, n$$

13. Appendix 2

This appendix presents two graphical representations for each pair of models in the observed categorization. Both graphs are built with the same data. Note that the Y-axis of all graphs is truncated at the value of 0.7.

The first graph arranges schools in the figure according to the five school categories (category E to category A from left to right). Among a category (for instance school category E), schools are listed by alphabetical order. This graphical representation allows identifying visually where the divergence is mostly concentrated. For instance, Figure 3 displays the efficiency scores of BM1986a and BM1986b. The gap between the two curves is more important on the left of the graph, meaning that the divergence occurs mostly in the disadvantaged schools.

The second graph arranges schools in the figure according to the efficiency scores obtained by one of the two models contained in the graph. For instance, Figure 4 arranges schools in the figure according to the efficiency scores obtained by the BM1986a model.

Figure 3
Efficiency scores provided by BM1986a and BM1986b for each school – first graph

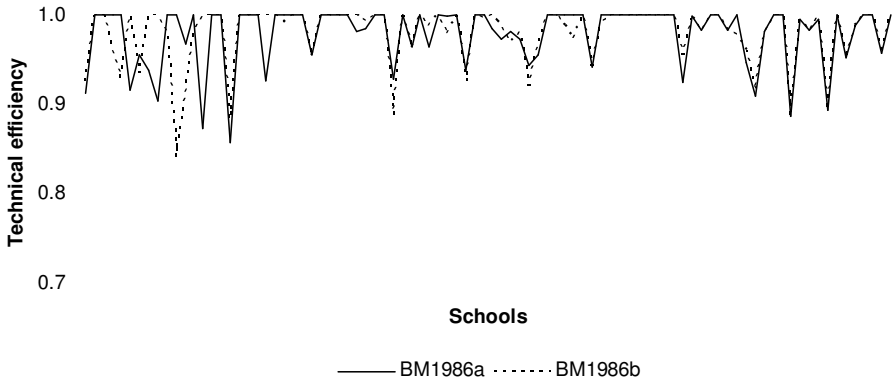


Figure 4
Efficiency scores provided by BM1986a and BM1986b for each school – second graph

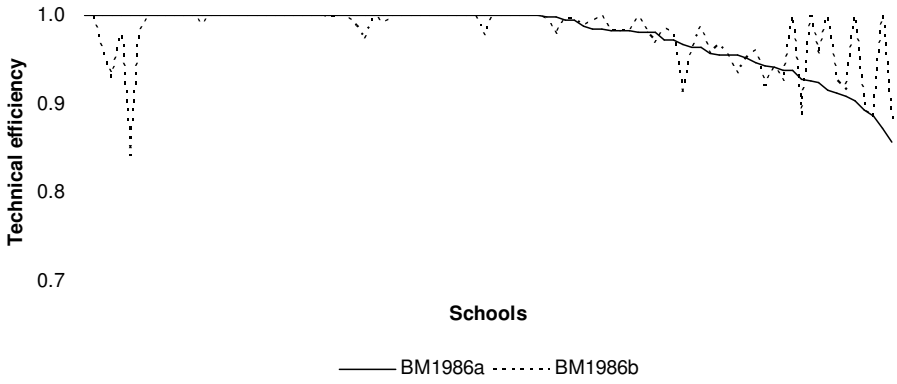


Figure 5
Efficiency scores provided by BM1986a and R1991 for each school – first graph

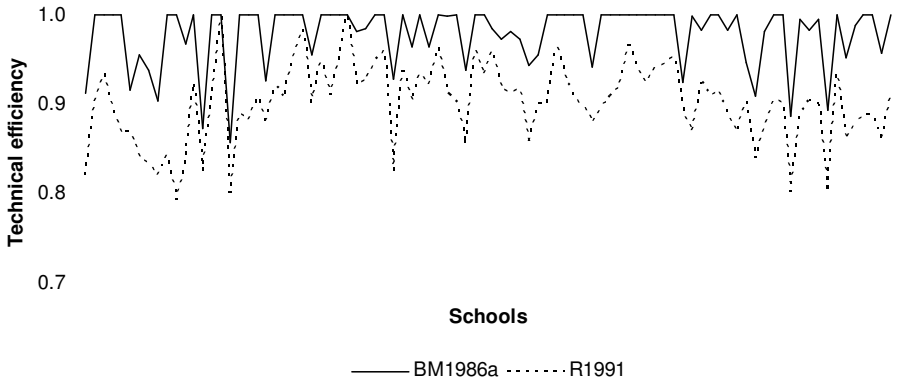


Figure 6
Efficiency scores provided by BM1986a and R1991 for each school – second graph

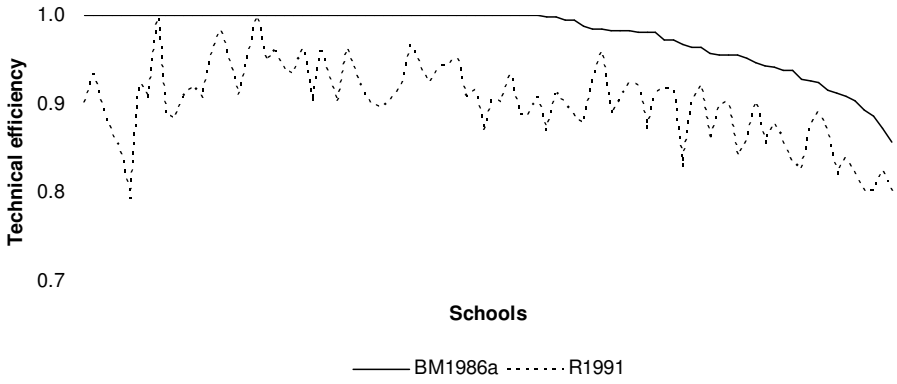


Figure 7
Efficiency scores provided by BM1986a and YP2006-I&O for each school – first graph

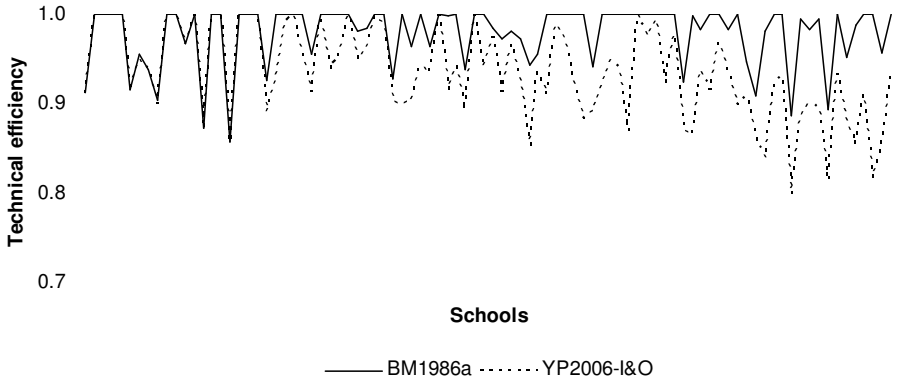


Figure 8
Efficiency scores provided by BM1986a and YP2006-I&O for each school – second graph

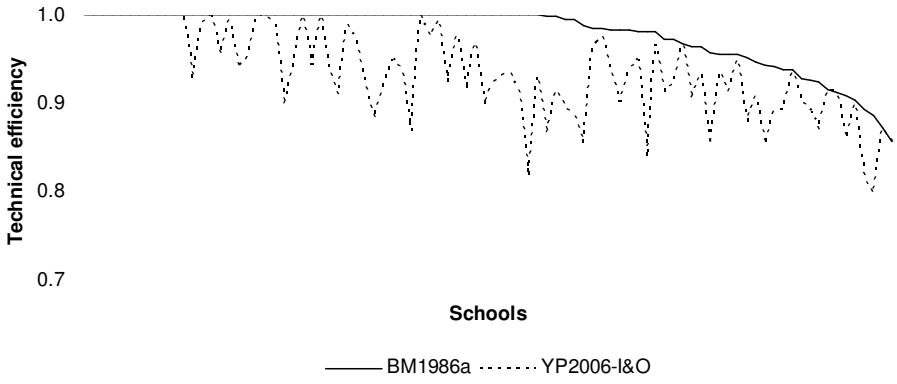


Figure 9
Efficiency scores provided by BM1986a and YP2006-I for each school – first graph

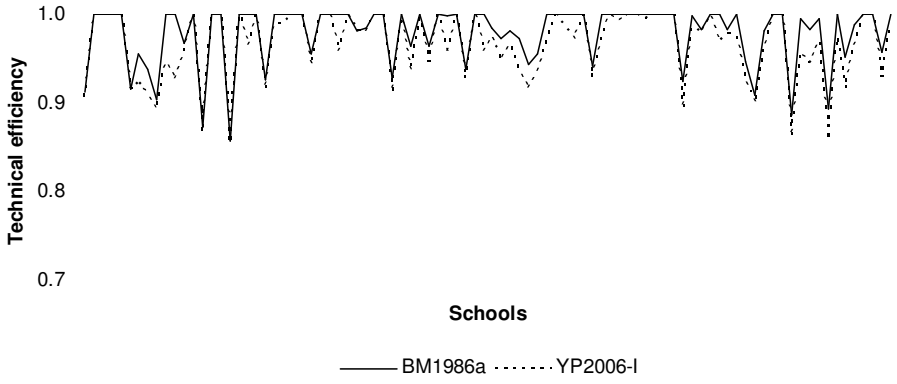


Figure 10
Efficiency scores provided by BM1986a and YP2006-I for each school – second graph

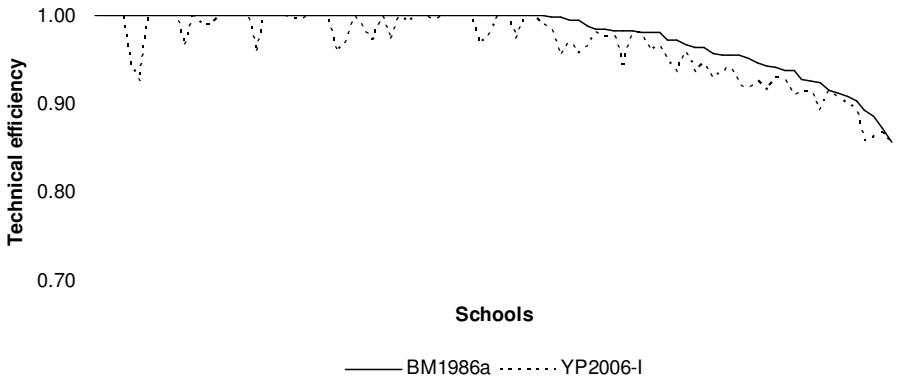


Figure 11
Efficiency scores provided by BM1986a and H2014 for each school – first graph

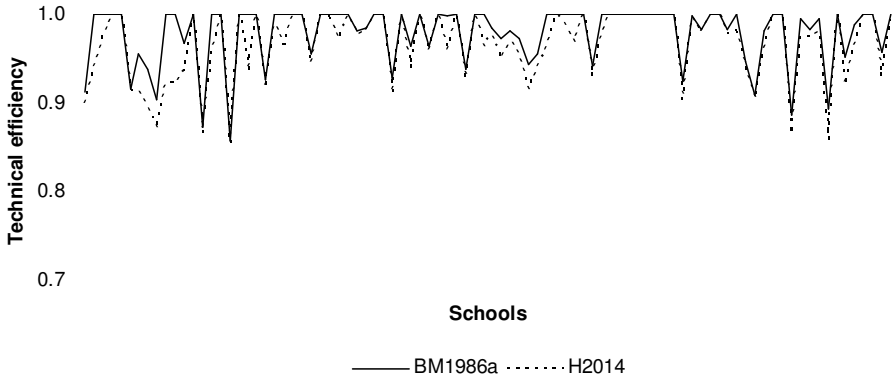


Figure 12
Efficiency scores provided by BM1986a and H2014 for each school – second graph

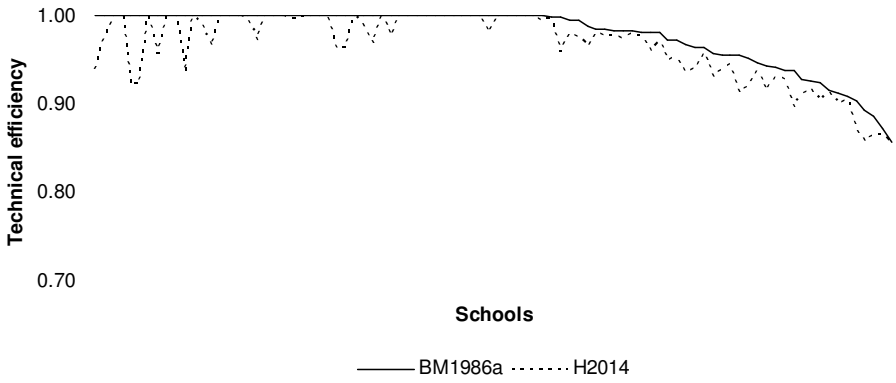


Figure 13
Efficiency scores provided by BM1986a and VRS for each school – first graph

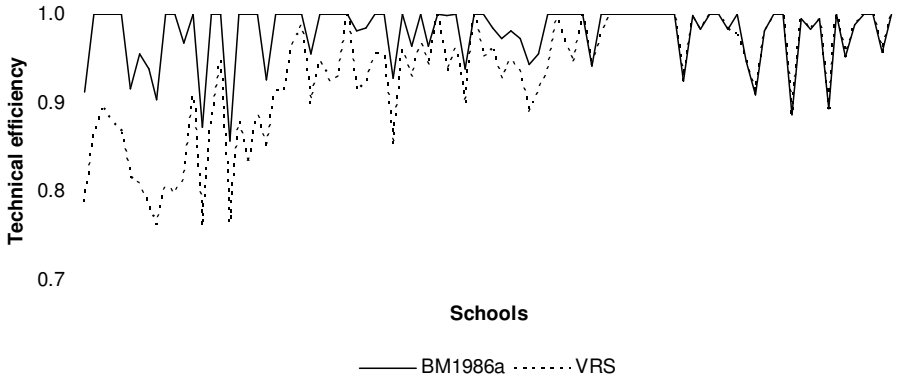


Figure 14
Efficiency scores provided by BM1986a and VRS for each school – second graph

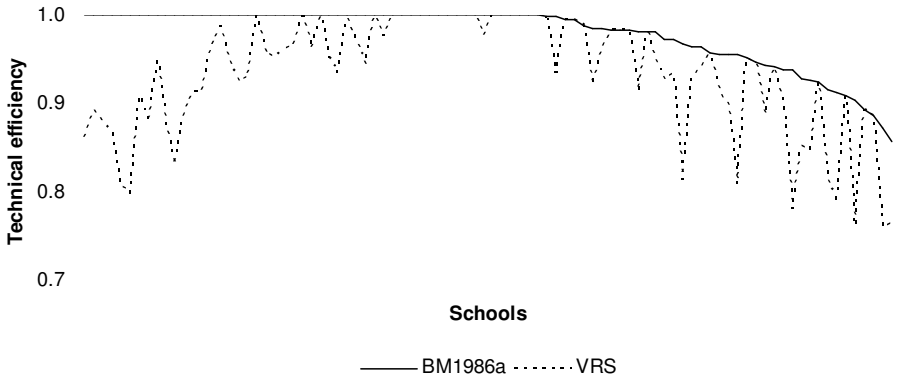


Figure 15
Efficiency scores provided by BM1986a and C1981 for each school – first graph

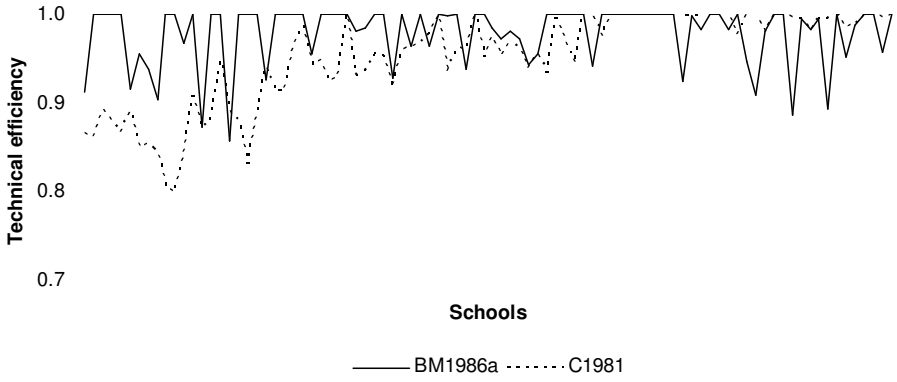


Figure 16
Efficiency scores provided by BM1986a and C1981 for each school – second graph

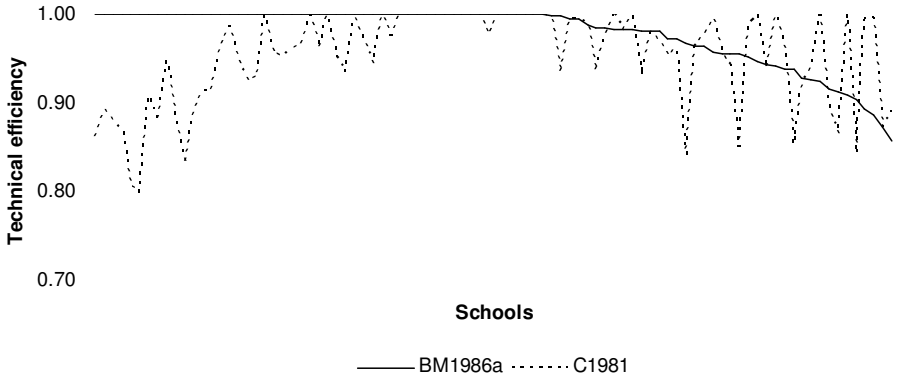


Figure 17
Efficiency scores provided by BM1986b and R1991 for each school – first graph

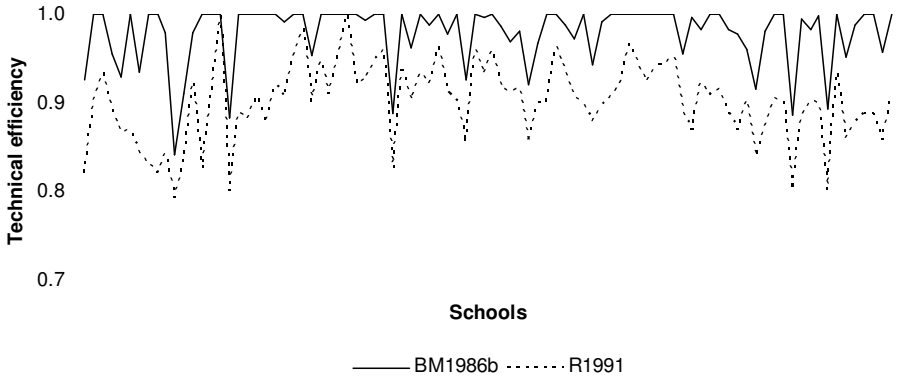


Figure 18
Efficiency scores provided by BM1986b and R1991 for each school – second graph

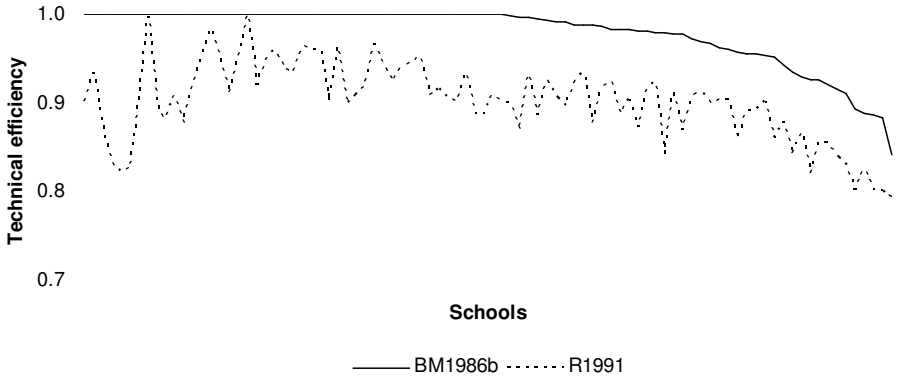


Figure 19
Efficiency scores provided by BM1986b and YP2006-I&O for each school – first graph

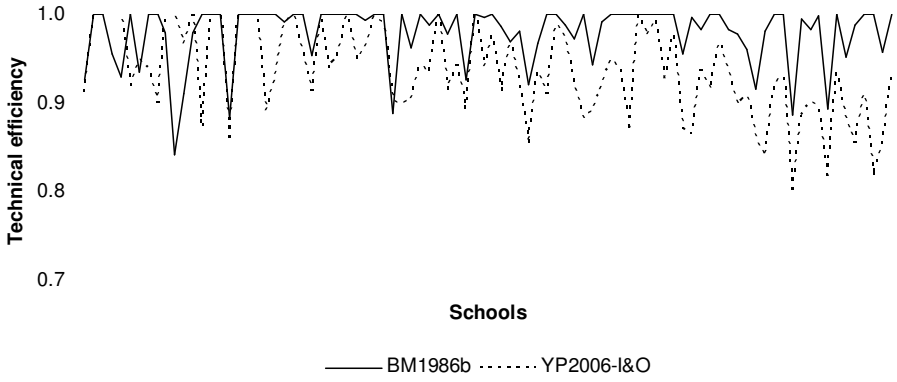


Figure 20
Efficiency scores provided by BM1986b and YP2006-I&O for each school – second graph

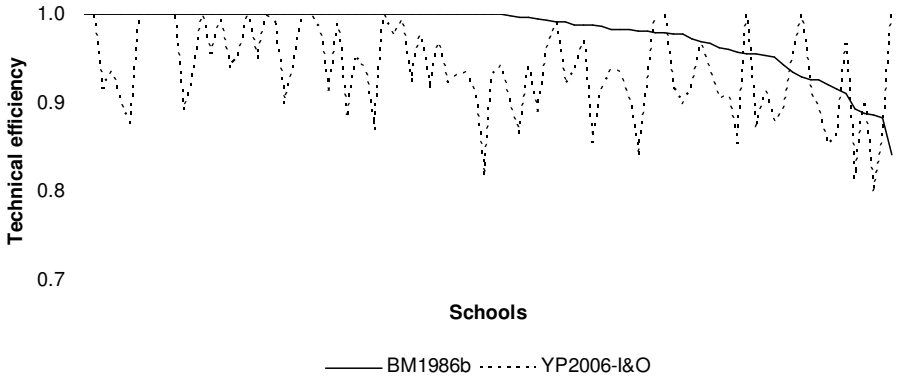


Figure 21
Efficiency scores provided by BM1986b and YP2006-I for each school – first graph

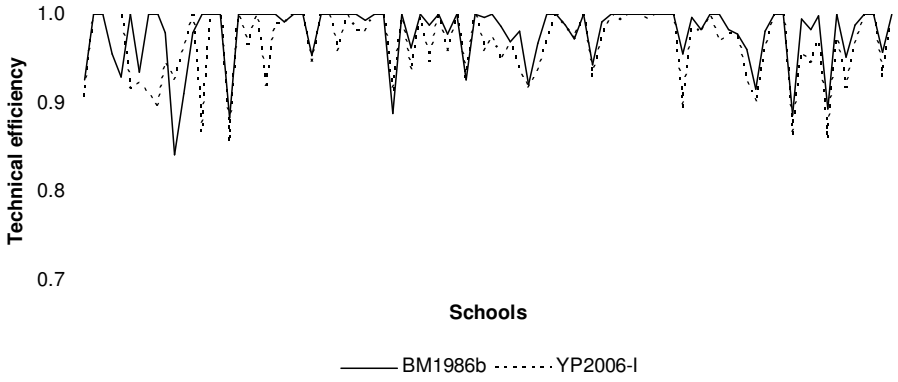


Figure 22
Efficiency scores provided by BM1986b and YP2006-I for each school – second graph

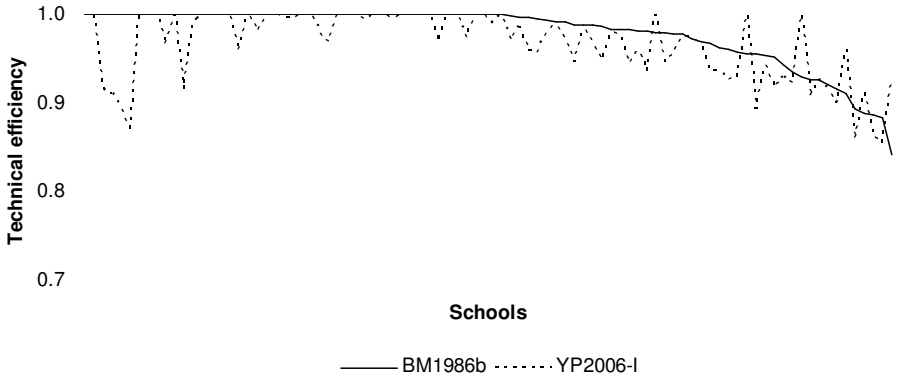


Figure 23
Efficiency scores provided by BM1986b and H2014 for each school – first graph

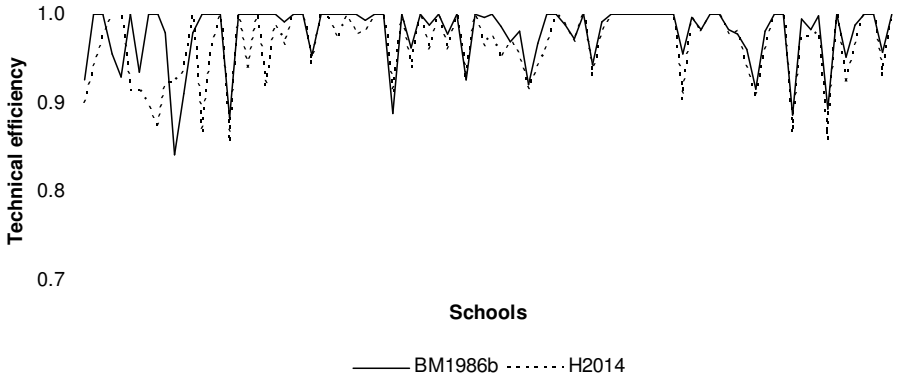


Figure 24
Efficiency scores provided by BM1986b and H2014 for each school – second graph

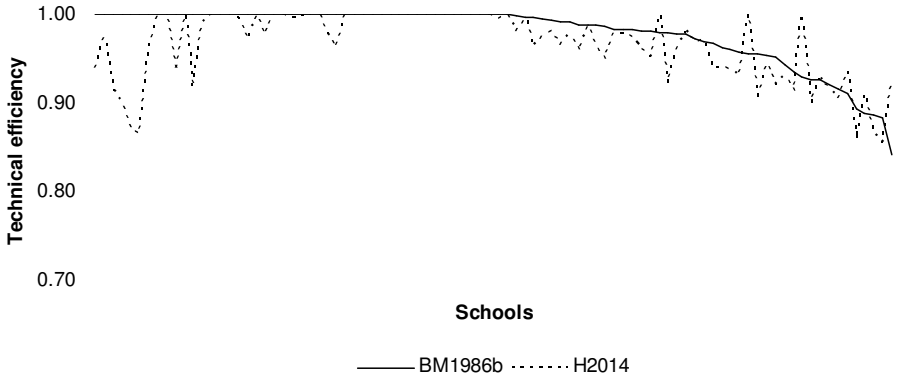


Figure 25
Efficiency scores provided by BM1986b and VRS for each school – first graph

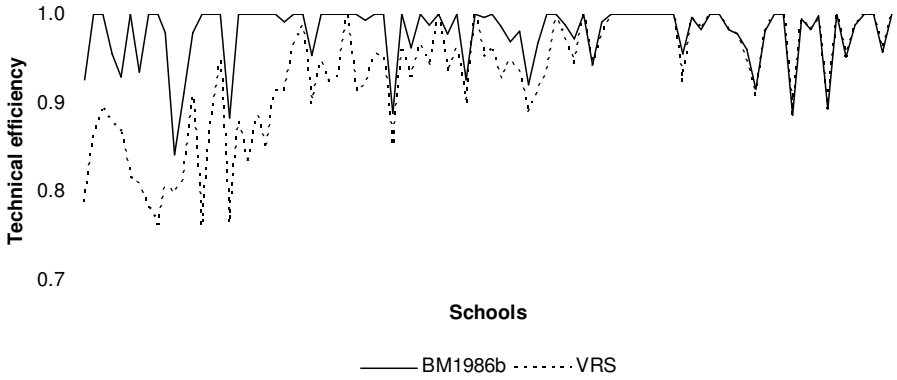


Figure 26
Efficiency scores provided by BM1986b and VRS for each school – second graph

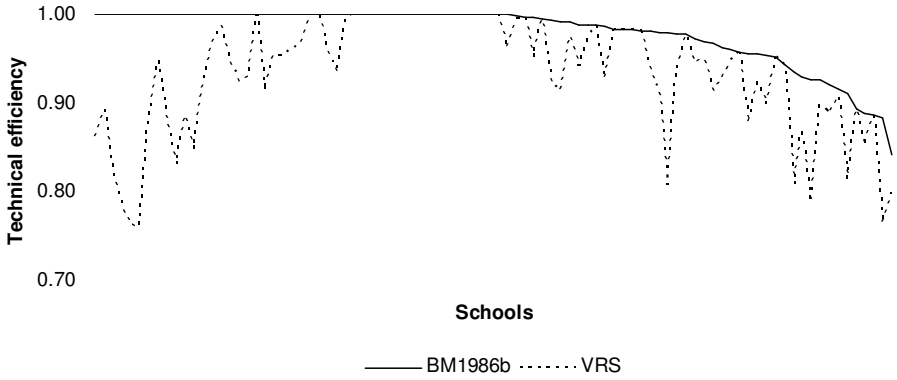


Figure 27
Efficiency scores provided by BM1986b and C1981 for each school – first graph

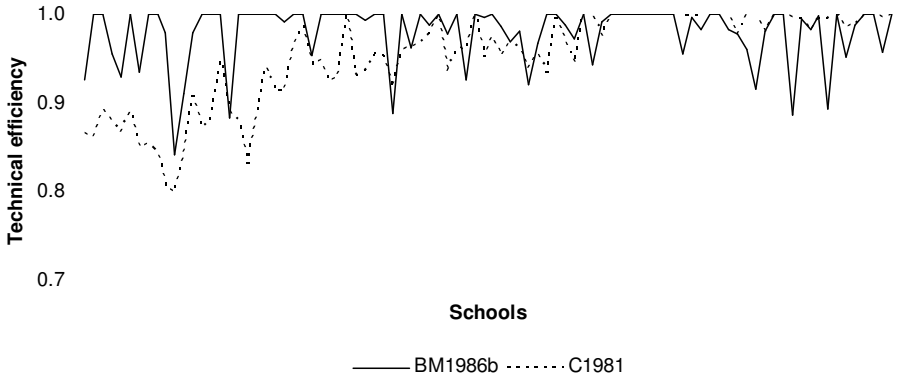


Figure 28
Efficiency scores provided by BM1986b and C1981 for each school – second graph

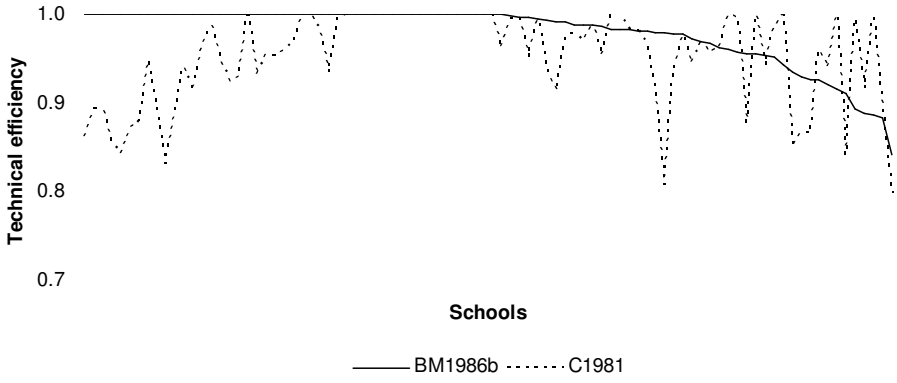


Figure 29
Efficiency scores provided by R1991 and YP2006-I&O for each school – first graph

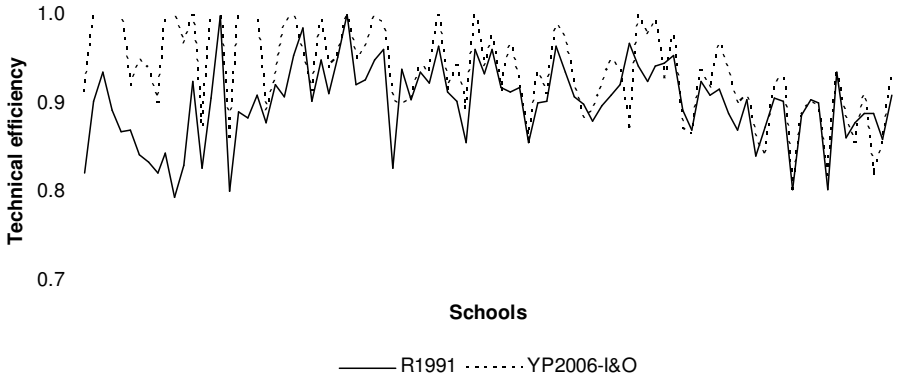


Figure 30
Efficiency scores provided by R1991 and YP2006-I&O for each school – second graph

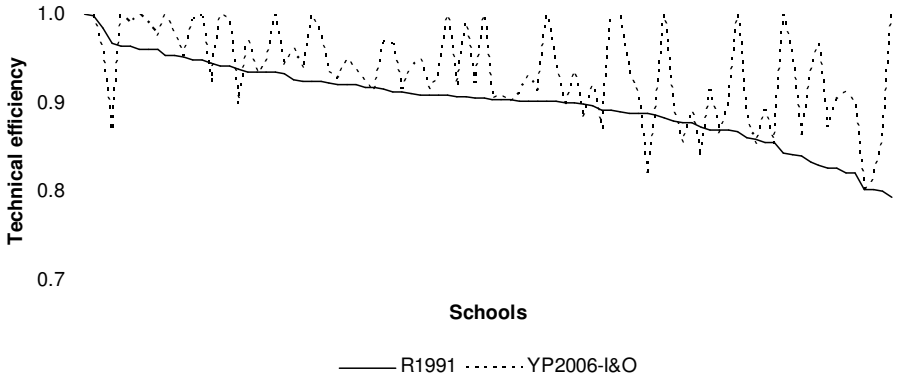


Figure 31
Efficiency scores provided by R1991 and YP2006-I for each school – first graph

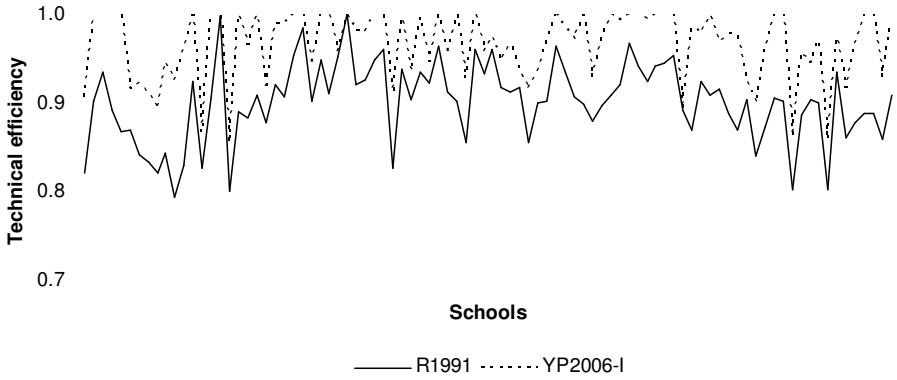


Figure 32
Efficiency scores provided by R1991 and YP2006-I for each school – second graph

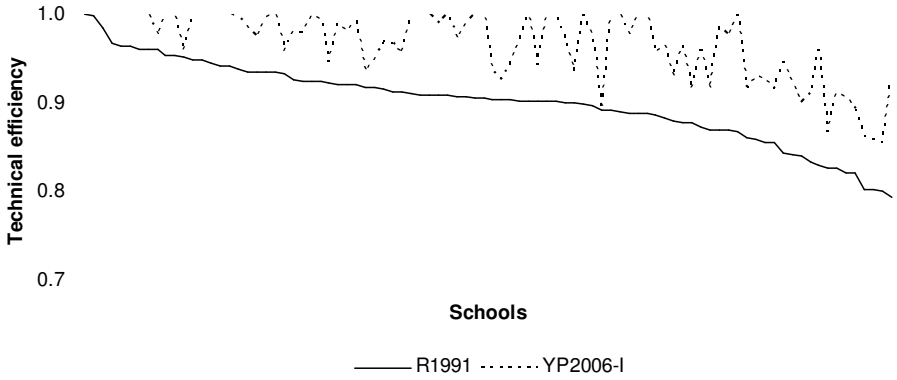


Figure 33
Efficiency scores provided by R1991 and H2014 for each school – first graph

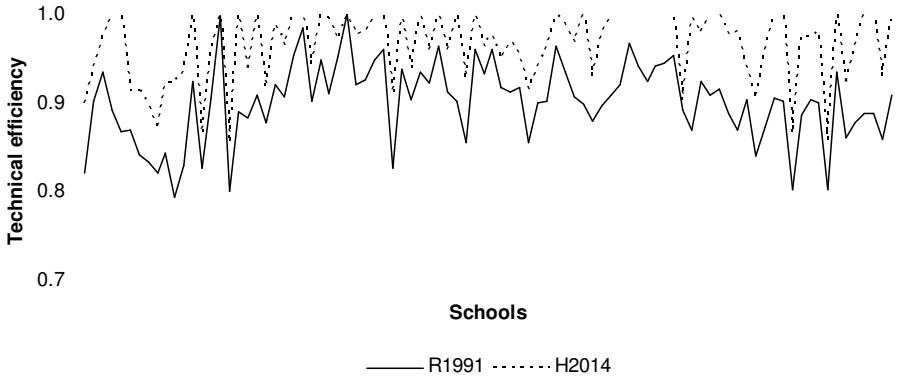


Figure 34
Efficiency scores provided by R1991 and H2014 for each school – second graph

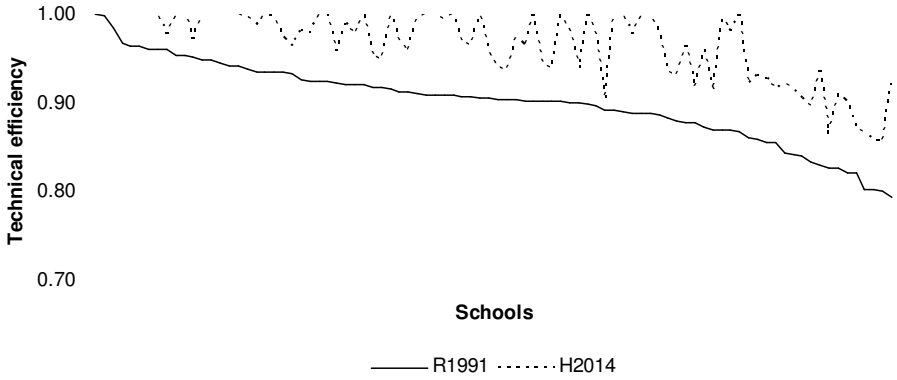


Figure 35
Efficiency scores provided by R1991 and VRS for each school – first graph

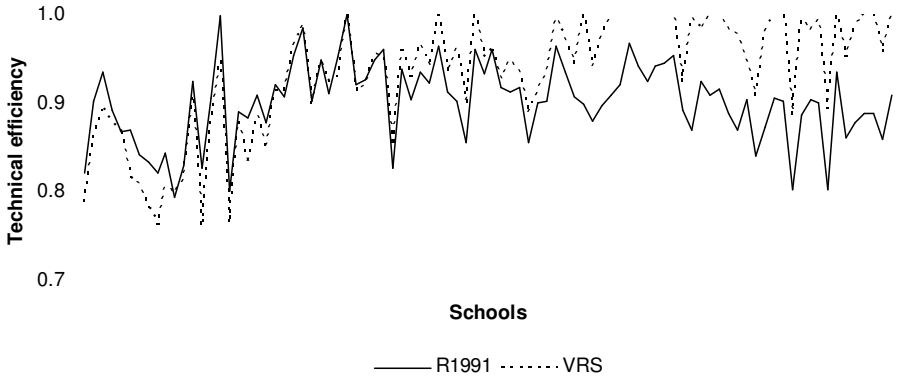


Figure 36
Efficiency scores provided by R1991 and VRS for each school – second graph

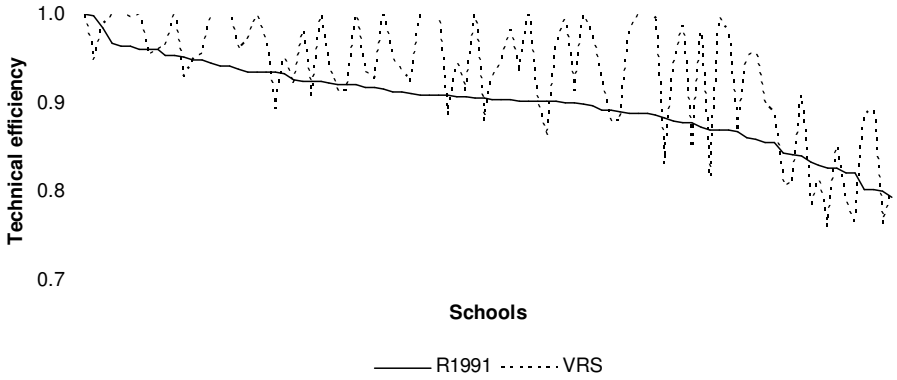


Figure 37
Efficiency scores provided by R1991 and C1981 for each school – first graph

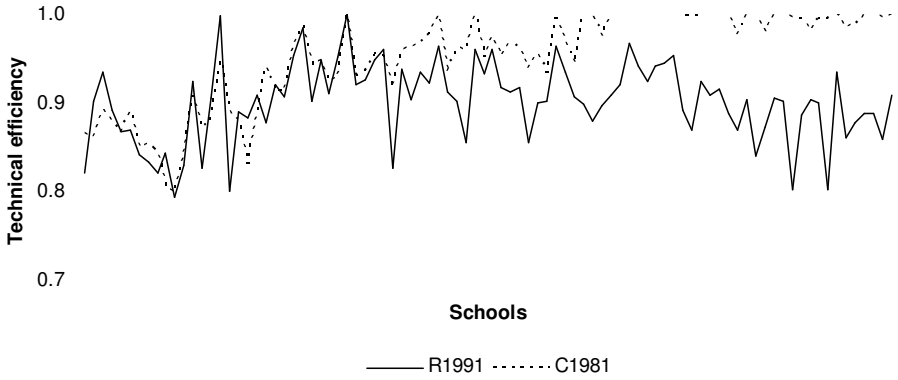


Figure 38
Efficiency scores provided by R1991 and C1981 for each school – second graph

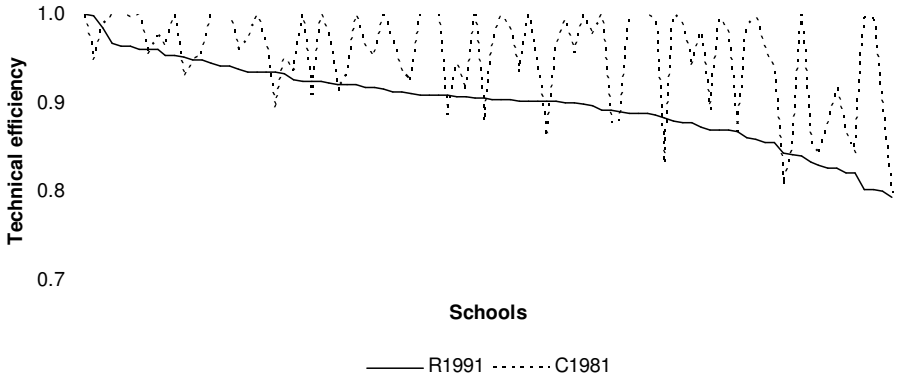


Figure 39
Efficiency scores provided by YP2006-I&O and YP2006-I for each school – first graph

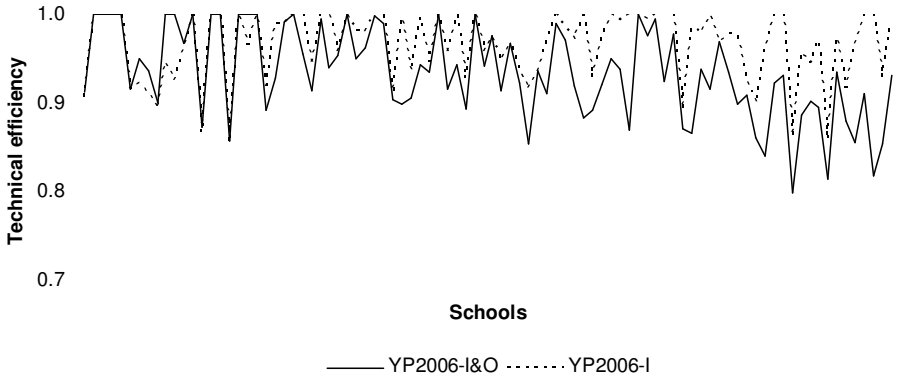


Figure 40
Efficiency scores provided by YP2006-I&O and YP2006-I for each school – second graph

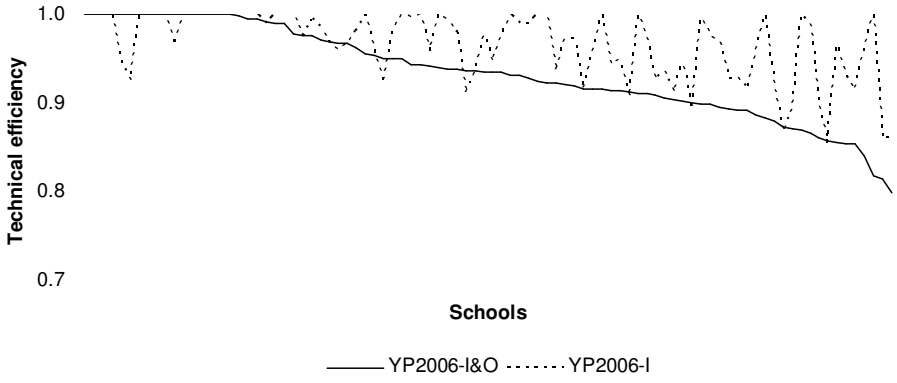


Figure 41
Efficiency scores provided by YP2006-I&O and H2014 for each school – first graph

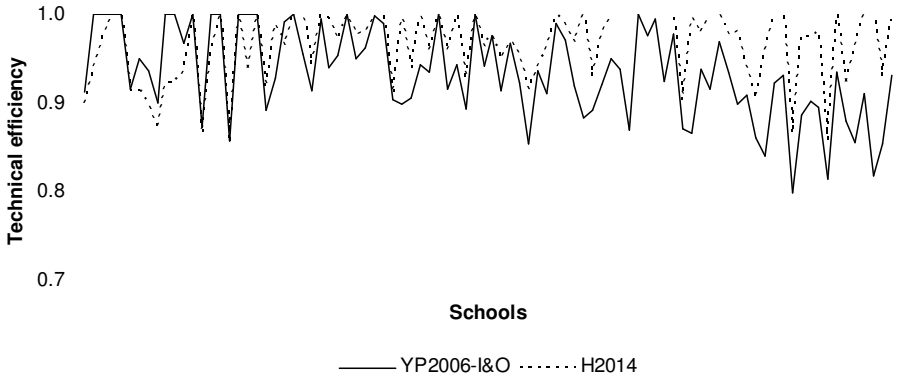


Figure 42
Efficiency scores provided by YP2006-I&O and H2014 for each school – second graph

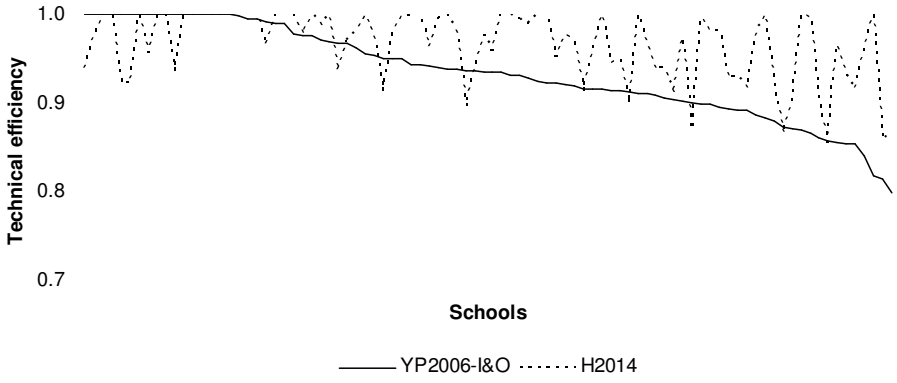


Figure 43
Efficiency scores provided by YP2006-I&O and VRS for each school – first graph

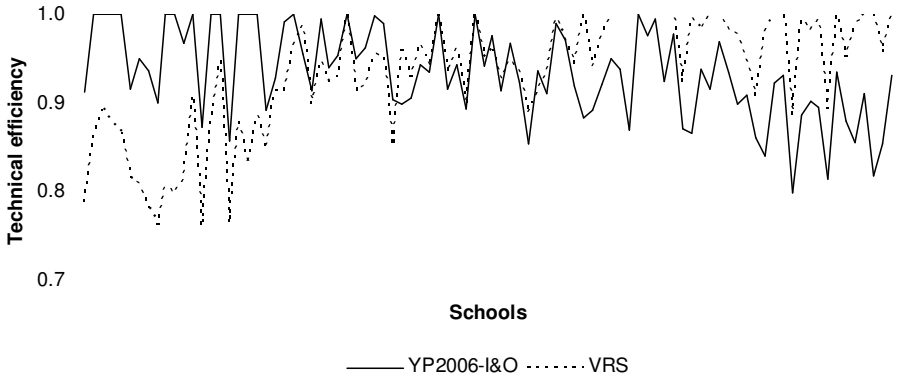


Figure 44
Efficiency scores provided by YP2006-I&O and VRS for each school – second graph

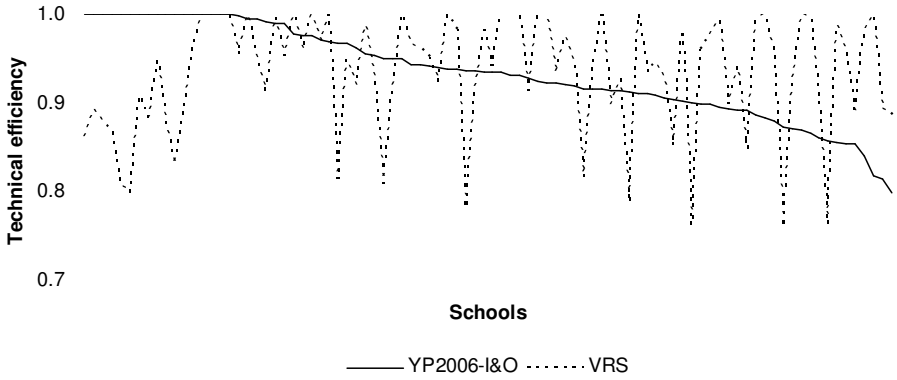


Figure 45
Efficiency scores provided by YP2006-I&O and C1981 for each school – first graph

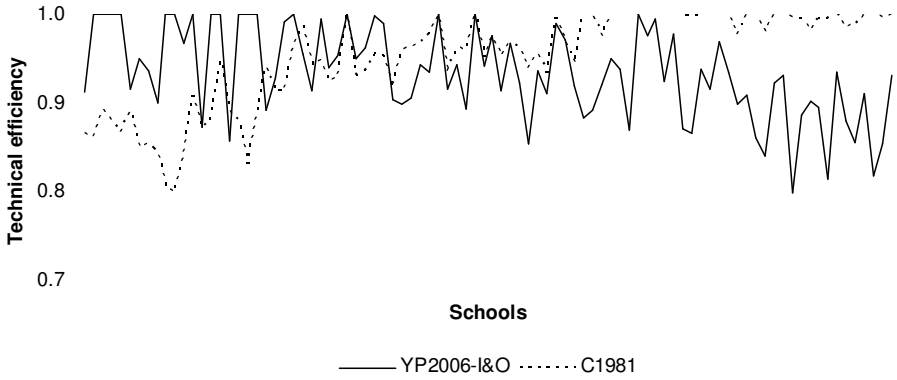


Figure 46
Efficiency scores provided by YP2006-I&O and C1981 for each school – second graph

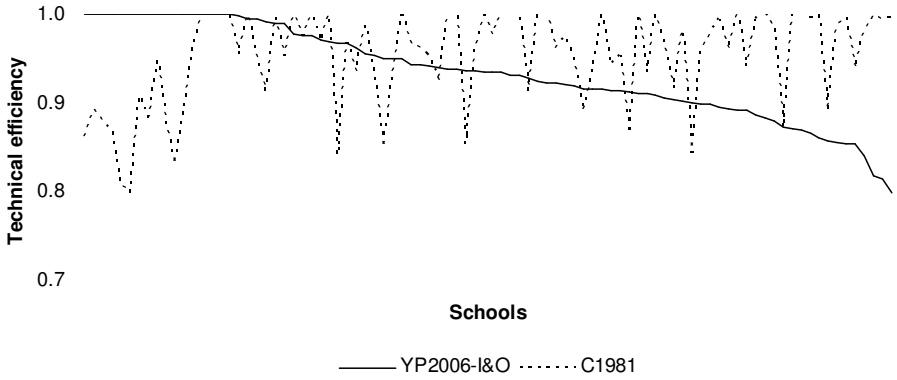


Figure 47
Efficiency scores provided by YP2006-I and H2014 for each school – first graph

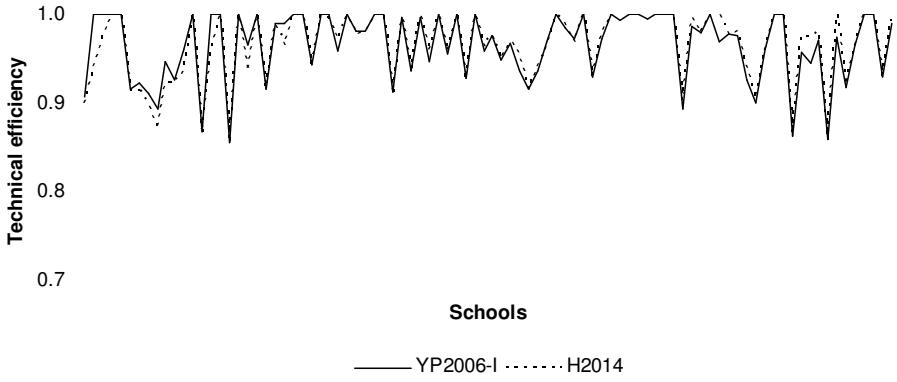


Figure 48
Efficiency scores provided by YP2006-I and H2014 for each school – second graph

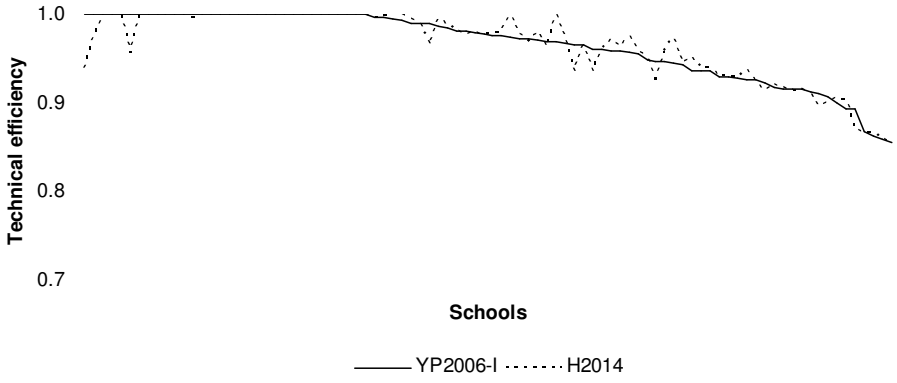


Figure 49
Efficiency scores provided by YP2006-I and VRS for each school – first graph

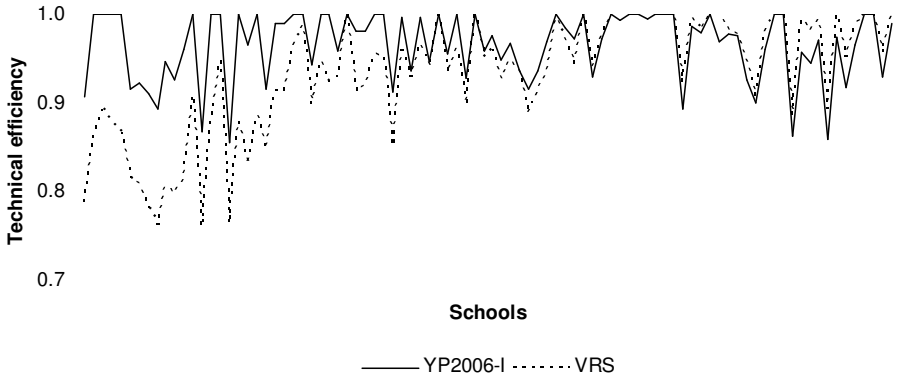


Figure 50
Efficiency scores provided by YP2006-I and VRS for each school – second graph

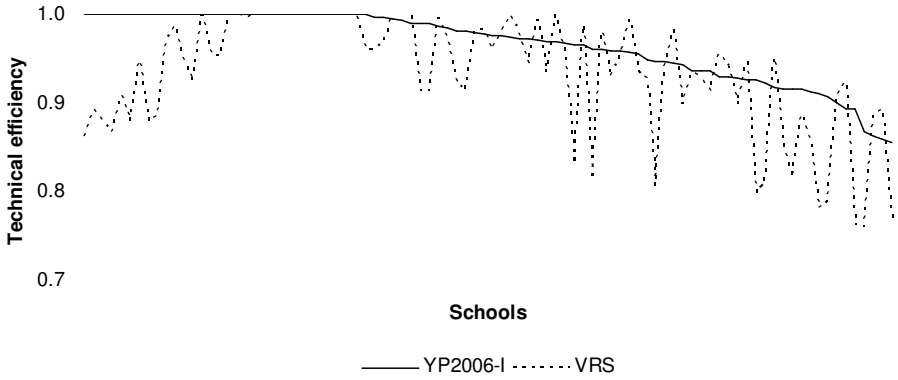


Figure 51
Efficiency scores provided by YP2006-I and C1981 for each school – first graph

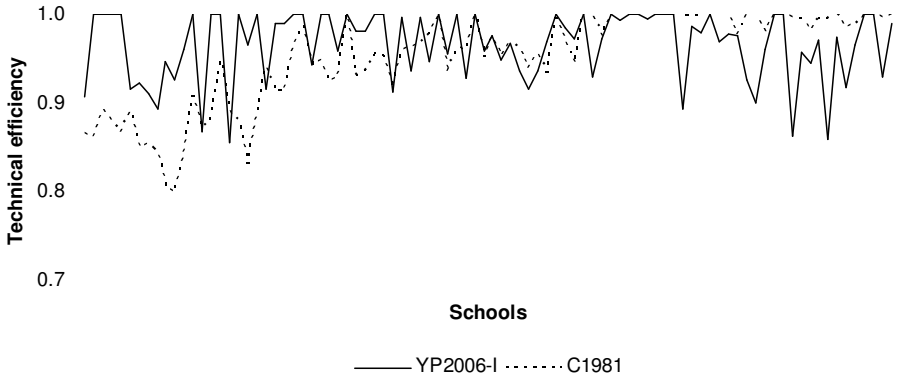


Figure 52
Efficiency scores provided by YP2006-I and C1981 for each school – second graph

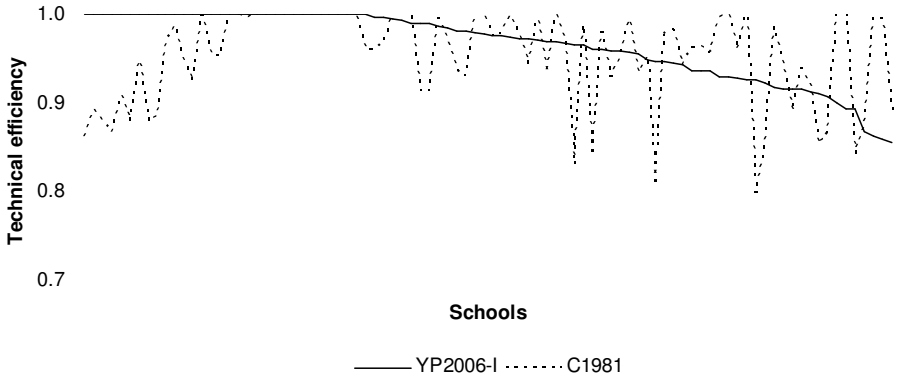


Figure 53
Efficiency scores provided by H2014 and VRS for each school – first graph

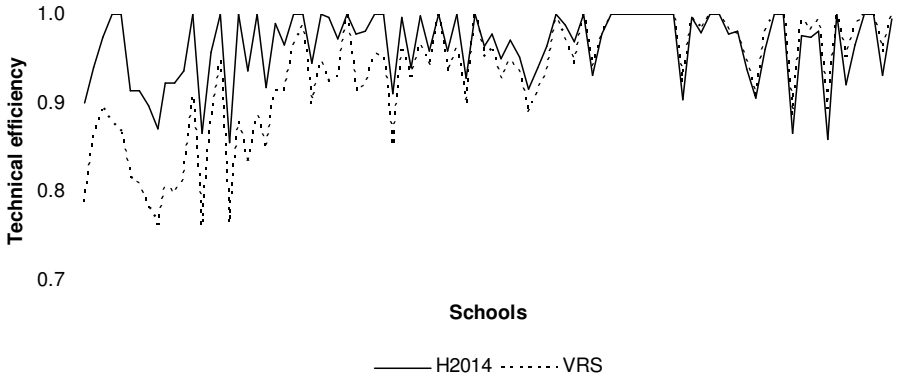


Figure 54
Efficiency scores provided by H2014 and VRS for each school – second graph

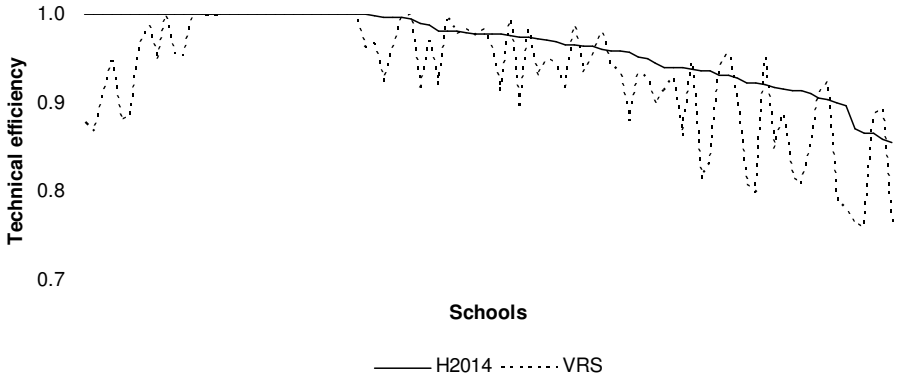


Figure 55
Efficiency scores provided by H2014 and C1981 for each school – first graph

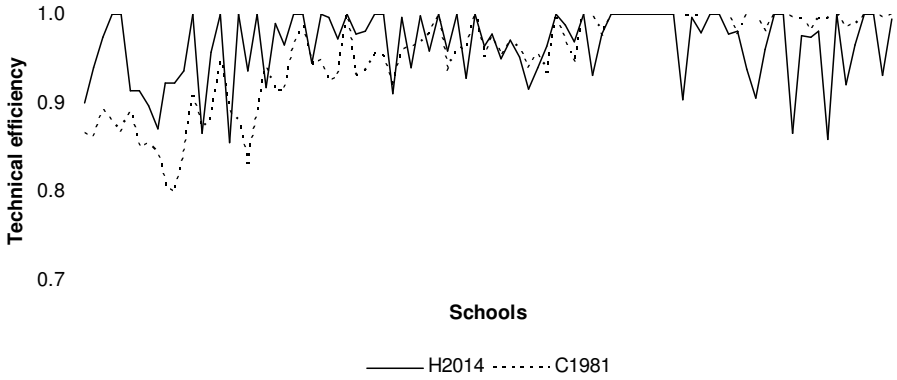


Figure 56
Efficiency scores provided by H2014 and C1981 for each school – second graph

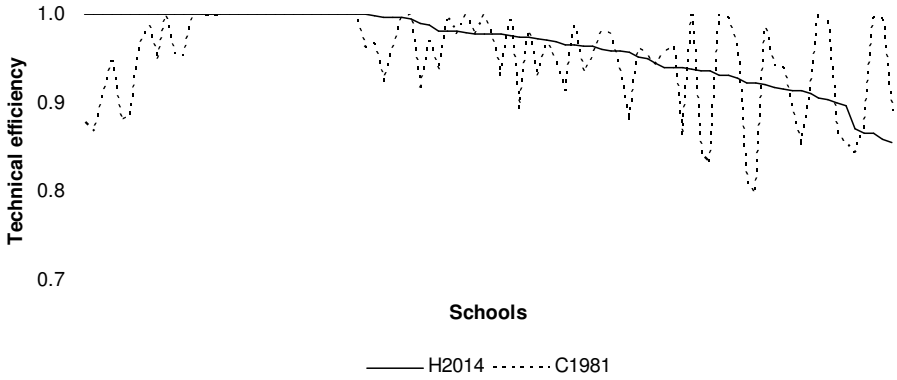


Figure 57
Efficiency scores provided by VRS and C1981 for each school – first graph

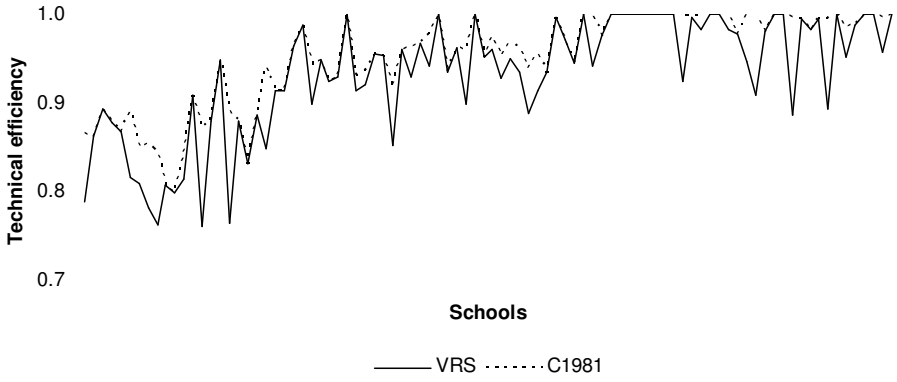


Figure 58
Efficiency scores provided by VRS and C1981 for each school – second graph