Impact evaluation in a multi-input multi-output setting: Evidence on the effect of additional resources for schools

Giovanna D'Inverno[a], Mike Smet[a,b], Kristof De Witte[a,c]

[a]Leuven Economics of Education Research, Faculty of Business and Economics, KU Leuven, Belgium
[b]Work and Organisation Studies, Faculty of Business and Economics, KU Leuven, Belgium
[c] TIER, Maastricht University, Maastricht, The Netherlands

January 30, 2019

Abstract:
This paper proposes an innovative approach to evaluate the causal impact of a policy change in a multi-input multi-output setting. It combines varied insights from the econometric impact evaluation techniques and the efficiency analysis. In particular, the current paper accounts for endogeneity issues by introducing a quasi-experimental setting within a conditional multi-input multi-output efficiency framework and decompose the overall efficiency between 'group-specific' efficiency (i.e., reflecting internal managerial inefficiency) and 'program' efficiency (i.e., explaining the impact of the policy intervention on performance). This framework allows the researcher to interpret the efficiency scores in terms of causality. The practical usefulness of the methodology is demonstrated through an application to secondary schools in Flanders, Belgium. By exploiting an exogenous threshold, the paper examines whether additional resources for disadvantaged students impact the efficiency of schools. The empirical results indicate that additional resources do not causally influence efficiency around the threshold.

JEL-Classification: H52, I22, I24, I28
Keywords: Impact evaluation, Efficiency, Causal inference, Secondary education

Contact: Giovanna.Dinverno@kuleuven.be, https://orcid.org/0000-0002-9738-4121; Mike.Smet@kuleuven.be, https://orcid.org/0000-0002-1304-4809; Kristof.DeWitte@kuleuven.be, https://orcid.org/0000-0003-0505-8642.
Corresponding author: Giovanna D'Inverno.

1. Introduction

There has been an increasing pressure for evidence-based interventions to channel the budgetary resources in the most appropriate way towards well-defined priorities (OECD 2017b, c). This puts forth the intricate nature of either 'effectiveness' or 'efficiency' of interventions. Effectiveness assesses whether the policy has reached its pursued goal, whereas efficiency examines whether it has been done by using the minimum amount of resources or producing the maximum amount of outputs. However, the occurrence of endogeneity might stall the attempts of the researcher in the domain of policy evaluation to go beyond correlational evidence. Endogeneity might arise from 'omitted variables' that influence the outcomes under consideration and are correlated with other independent variables, from 'self-selection' into the treatment, from non-random measurement errors, or from 'reverse causality', which refers to a two-way relationship capable of generating a self-reinforcing mechanism in the allocation of the resources and/or in the outcome that can be observed. The econometric impact (or program) evaluation literature has proposed consolidated policy evaluation techniques to address endogeneity issues, such as Regression Discontinuity Design (RDD), Difference-in-Differences (DiD) or Instrumental Variables (IV) (Abadie and Cattaneo 2018; Angrist and Pischke 2009). By contrast, the efficiency literature has just recently started addressing the endogeneity problem in the frontier estimation. The use of state-of-the-art techniques, such as the robust and the conditional analysis in the nonparametric formulation (Simar et al. 2016) or advanced tools in the parametric formulation (Amsler et al. 2016), might mitigate measurement errors in the frontier estimation, however, they still do not address the other endogeneity issues. Due to this, there is an emerging literature that caters its attention towards endogeneity in efficiency, from both a theoretical perspective and empirical application by using tools proposed by the impact evaluation literature (for a comprehensive review, Santín and Sicilia 2017b). Thus, this paper contributes to this emerging literature by providing a framework to overcome these endogeneity issues and evaluate the causal impact of a policy change on efficiency.

In this study, we propose an innovative procedure to capture the causal impact of a policy intervention on efficiency, whenever the treatment status depends on an exogenously set threshold. We combine insights from a regression discontinuity approach with insights from metafrontier and conditional efficiency measurement, integrating two streams of literature. For the efficiency literature, the suggested approach builds on the seminal paper conducted by Charnes et al. (1981) that distinguished management practices from program effects; however, we move beyond correlational evidence to a causal interpretation of the findings. For the impact evaluation literature, the followed approach is innovative as it allows impact evaluation in a multi-input and multi-output setting, and successfully grasps synergies in the input/output mix, rather than considering one output at the time. Moreover, we do not only investigate whether a policy has an impact on the outcome, but we can also explore the mechanisms leading to the observed outcome. For example, we can analyze how the resources allocated for the policy intervention have been used, regardless of whether it is effective or, if not, even explaining why.

The suggested approach can be implemented to evaluate the impact of a policy from a performance perspective and can also be adapted to different frontier model specification and field of applications.[1] Additionally, it can be seen as a complementary tool to the effectiveness analysis. In this regard, it might be a procedure to detect why a policy might be

[1] To stimulate further applications, the code is available upon request.

or not effective: for example, a policy might not lead to the expected outcomes and thus ineffective, because of the mismanagement of the resources and thus inefficient.

To show the practical usefulness of the proposed procedure, we examine the efficiency effects of a large-scale (both in number of students and in funds) 'Equal Educational Opportunity (EEO) program' in Flanders, Belgium. Particularly, we evaluate the impact of additional funding provided to schools which pass an exogenously determined percentage of disadvantaged students. Similar programs are popular in many countries as socio-economic status has been widely recognized as one of the most important aspects that impact educational outcomes (Agasisti et al. 2018; Dahl and Lochner 2012; Haveman and Wolfe 1995) and labor market outcomes (Grenet 2013; Oosterbeek and Webbink 2007; Pischke and von Wachter 2008; Stephens and Yang 2014). Moreover, governmental authorities have encouraged various programs and policies to inhibit the impact of socio-economic factors onto the pedagogical achievements (Gibbons et al. 2018), such as voucher programs (Muralidharan and Sundararaman 2015), class size reduction (Duflo et al. 2015) and additional funding (Leuven et al. 2007).

This paper is the first to provide causal evidence on the efficiency implications of providing additional funding to schools. There might be an impact on efficiency as the additional funding might result in a different educational production function for the schools (Levin 1974; Hanushek 1979, 2002). Thus, schools with additional funding can generate more outputs with the provided resources. With reference to the debate about the efficiency and effectiveness of school resources on educational outcomes, unsolved endogeneity problems might lead to biased results and explain the ambiguous findings of the literature (Hanushek 2006; Jackson et al. 2016). First, endogeneity might arise from the various sources mentioned above while estimating the educational production function (Cazals et al. 2016; Cordero et al. 2015; Mayston 2003; Santín and Sicilia 2017a, c; Simar et al. 2016). Second, this might also occur when extending the focus of the efficiency in education studies from the overall production frontier estimation to the program efficiency evaluation. Since the seminal paper by Charnes et al. (1981), various researchers and scholars intended to disentangle program efficiency from the managerial one, in the attempt to disentangle a component attributable to the context or the program under which a school operates from a component related to its internal managerial characteristics. Such decomposition aids in differentiating evidence of good school managerial practices from a bad one or evidence of good programs from a bad school management. However, the endogeneity might arise in this framework as well, leading to biased program/managerial efficiency estimates and preventing from causal interpretation of the findings. In the empirical application of the current study, we tackle endogeneity issues both for the education production function estimation and in the decomposition between managerial and program efficiency by using the procedure proposed in this paper.

This paper contributes to four main strands of literature. First, it contributes to the emerging operational research literature dealing with endogeneity issues in non-parametric frontier estimation (Cazals, Fève, Florens, and Simar 2016; Cordero, Santín, and Sicilia 2015; Simar, Vanhems, and Van Keilegom 2016). Second, it adds onto the literature pertaining to the impact evaluation in efficiency by providing causal interpretation of the findings. Third, it contributes to the literature bridging the gap between effectiveness and efficiency, by combing regression discontinuity together with conditional metafrontier approach in the efficiency framework. Fourth, from an empirical perspective, the current study contributes to the economics of education literature by providing new impact evaluation evidence on an 'Equal Educational Opportunity (EEO) program'. As many countries are struggling with similar equal educational opportunities challenges (UN General Assembly 2015), the empirical findings will be relevant beyond the specific Flemish context.

The remainder of this paper is organized as follows. Section 2 explains the suggested approach to handle endogeneity issues in efficiency impact evaluation. Section 3 shows the empirical application to an education context. Section 4 presents the steps and their relative implementation together with the empirical findings for secondary education. To conclude, Section 5 presents a critical discussion of the main methodological aspects and outlines the ways to move forward along the path traced by this paper.

2. Methodology

To assimilate the causal impact of a policy intervention on efficiency, we proceed in three steps. First, to tackle endogeneity in the production frontier, we focus on the treated and control group around an exogenous cutoff. Second, we disentangle the overall efficiency into a *managerial* and a *program* component. Because of the quasi-experimental setting defined in the first step, we can give causal interpretation to the estimates obtained in this second step. Third, using a conditional efficiency analysis we explore potential mechanisms.

Step 1. Tackling the endogeneity issue in frontier estimation

The literature pertaining to the econometric impact evaluation has developed and consolidated a range of techniques that address endogeneity issues, such as Regression Discontinuity Design (RDD), Difference-in-Differences (DiD) and Instrumental Variables (IV) (Abadie and Cattaneo 2018; Angrist and Pischke 2009). These techniques are capable of estimating the casual effect of the policy intervention by comparing a group of treated observations with those of the untreated ones, which have similar characteristics. The latter group is meant to represent what would have happened if the treated units had not received the treatment, namely the counterfactual, isolating in this way the impact of the intervention (Schlotter et al. 2011).

The proposed approach deals with a policy intervention where the treatment is assigned to observations based on whether a specific covariate c, the "assignment variable", falls below or above a certain cutoff value c_0: this is the quasi-experimental setting handled in the regression discontinuity design (Cattaneo et al. 2015; Lee and Lemieux 2010). Following the RDD standard notation:

$$D_i = \begin{cases} 1 & if\ c_i \geq c_0 \\ 0 & if\ c_i < c_0 \end{cases} \qquad (2.1)$$

where D_i denotes the treatment status of unit i and it is a deterministic and discontinuous function of c_i (Angrist and Pischke 2009): when $D_i = 1$, the unit is subject to the policy intervention and hence it is assigned to the treated group, otherwise to the control group.[2]

If the units have no precise control over the assignment variable, "there is a striking consequence: the variation in the treatment in a neighborhood of the threshold is 'as good as randomized'" (Lee and Lemieux 2010, p.293). Therefore, the treated and the untreated units are comparable, thus, the observations right below the cutoff can be perceived as a valid counterfactual for those that are right above the cutoff. Due to this reason, we might want to

[2] Specifically, the proposed approach follows the idea behind the sharp RDD (presence of perfect compliance) and accordingly the estimates measure average treatment effects. However, further research should extend the approach to a fuzzy RDD framework (presence of imperfect compliance, i.e. units might not receive the treatment even if they are eligible for it) and interpret accordingly the estimates as local average treatment effects. Moreover, it is straightforward to see that the treatment status as introduced in formula (2.1) might work also in the other way around, that is $D_i = 1$ if $c_i \leq c_0$ and $D_i = 0$ otherwise.

exclude the influence of observations far from the threshold and thus focus on more similar units. Following the insights of the nonparametric regression discontinuity design, the attention is restricted over a narrow window of observations. The choice of the width of the window is a crucial step and in the RDD literature it is mentioned as the problem of bandwidth selection (Calonico et al. 2014b; Imbens and Kalyanaraman 2012). The bandwidth should be neither too small nor too big. If the bandwidth were too small, there would be handful of observations to require meaningful estimates; whereas, if the bandwidth were too big, there would be too many observations, bringing into the analysis heterogeneity and confounding factors. For the choice of the optimal bandwidth h, we follow the idea behind the nonparametric local linear regression method and specifically adopt the robust data-driven bandwidth selection procedure proposed by Calonico et al. (2014b). Consequently, we restrict the full sample by considering only observations with $c_i \in [c_0 - h, c_0 + h]$, that is within h distance from the cutoff and hence the name *h% discontinuity sample* (Angrist and Lavy 1999; Leuven et al. 2007). The units with $c_i \in [c_0 - h, c_0)$ constitute to the control group, while the units with $c_i \in [c_0, c_0 + h]$ the treated group. In the practical implementation, the selection procedure requires the output variable and the assignment variable (also referred to as "running" variable or "forcing" variable in the RDD literature). Given the multi-input multi-output framework of the production frontier estimation and to handle the variability on the output side, for the current study, the researchers obtain as many ideal bandwidths as the number of outputs that can be considered for the efficiency analysis, varying between a lower and upper bound. In the spirit of local linear regression methods, having a range of optimal bandwidths (differently from the RDD applications where one outcome at the time is considered) is not a matter of concern, but rather a tool to check the robustness of the causal estimates (Lee and Lemieux 2010).

To support the internal validity of the RDD setting, there are several conditions that must be focused upon (Lee and Lemieux 2010). First and foremost, it is fundamental to check the hypothesis of no precise control over the assignment variable, as units might have incentive in manipulating it to benefit of the policy intervention. In the RDD literature the way to rule out sorting around the threshold is mainly twofold. First, baseline covariates should be similar in treated and control groups and have the same distribution so to support randomization around the cutoff. Second, a more formal test is suggested to check the continuity of the assignment variable density function (McCrary 2008). In addition to no manipulation, it is necessary to have a clear discontinuous jump in the probability of treatment at the cutoff point. If these conditions are met and the *h%* discontinuity sample with treated and control units is constructed, it is possible to proceed further with the second proposed step in the study.

Step 2. Decomposing the overall efficiency

Once the endogeneity issue has been solved by focusing on observations just right below and above the cutoff, we can proceed to the second step. In the second step, the performance evaluation of the units under analysis in a multi-input multi-output framework and its decomposition into a *managerial* and a *program* component are emphasized upon.

For explanatory purposes, let's start by considering a general production function that converts a vector of inputs $x = (x_1, \dots, x_k) \in \mathbb{R}^{K+}$ into a vector of outputs $y = (y_1, \dots, y_l) \in \mathbb{R}^{L+}$ and that can be presented in the following standard formulation (Afriat 1972):

$$y = f(x) \qquad (2.2)$$

where *f(.)* is the technology that determines the output production together with the inputs. Following O'Donnell (2016), a technology can be defined as "a technique, method or system for transforming inputs into outputs […] it is convenient to think of a technology as a book of instructions, or recipe". The set containing all the feasible input-output combinations for a given technology is labelled "production possibility set". In line with the axiomatic approach to production theory, it is common to assume certain axioms or properties concerning the technology, including no free lunch, free disposability of inputs and outputs and closedness. [3] However, this general production function implicitly neglects potential inefficiencies in the production process (Santín and Sicilia 2017b). Therefore, we can add an efficiency component u:

$$y = f(x) \cdot u \qquad (2.3)$$

Specifically, $u = 1$ suggests that the inputs are efficiently managed producing the maximum achievable output given the existing technology. If $u \in (0,1)$, the decision making unit (DMU) is not fully exploiting its capacity and, therefore, the observed level of outputs is determined not only by the used inputs and the available technology, but also by the level of mismanagement *u*. In the production frontier approach, the basic idea is to represent the relationship between inputs and outputs by encompassing all the observations under analysis. Referring to the production possibility set introduced above, its boundary represents the frontier. The "best practice" DMUs constitute the efficiency frontier and envelope all the other DMUs under analysis. Accordingly, the farther from the efficiency frontier, the more inefficient is the unit in the process of transforming inputs into outputs.

Looking at equation (2.3), an increase in the outputs can be obtained by a change in inputs (x), technology $(f(.))$ or efficiency (u). However, there might be spillover effects from one component to another one, which makes the idea of isolating one effect at a time a little puzzling. Furthermore, we do not know *a priori* the direction of the treatment impact on the production activity of the treated units. For example, on one hand, an increase in the inputs might result in scale economies and let the units achieve some targets otherwise not feasible (therefore producing spillover effects on the production technology or on the internal management efficiency). On other hand, additional resources might lead to a 'wealth effect', i.e. a significant amount of resources would be liable to be misused which can be observed in the general public spending framework (Cherchye et al. 2018; D'Inverno et al. 2018). In a multidimensional framework, more inputs might have an impact on one output, but not on others.

The efficiency literature dealing with impact evaluation proposes different approaches to evaluate group performance. Since the seminal paper by Charnes et al. (1981), Grosskopf and Valdmanis (1987), Månsson (1996), researchers have tried to disentangle program efficiency from the managerial one, in the attempt to distinguish a component attributable to the context or the program under which the DMU operates from a component related to its internal managerial characteristics (Aparicio et al. 2017; Aparicio and Santin 2017; Camanho and Dyson 2006; Johnson and Ruggiero 2014). In the procedure we propose, we adapt the concept of the non-parametric metafrontier approach developed by Battese and Rao (2002), Battese et al. (2004), and formalized by O'Donnell et al. (2008).[4]

Specifically, we consider the treated and the control group determined in step 1 by restricting the focus on units right above and below the exogenous cutoff. We measure the

[3] For a more formal discussion on the axiomatic framework, we refer for example to Shepard (1970) and Kerstens et al. (2018), among others.

[4] For a comprehensive overview, we refer the interested reader to Kerstens et al. (2018).

efficiency of each unit i belonging to one of the two groups by estimating a group-specific local production frontier (TE_i^D), where $D \in \{0, 1\} = \{Control, Treated\}$. Additionally, we measure the efficiency of each unit i belonging to the *h*% discontinuity sample (i.e., where both treated and control units are present) by estimating an overall production frontier (TE_i^*). The *program* efficiency is computed for each unit i as follows:

$$Program\ efficiency_i^D = \frac{TE_i^*}{TE_i^D} = \frac{Overall\ efficiency_i}{Managerial\ efficiency_i^D} \tag{2.4}$$

where $D \in \{0, 1\} = \{ Control, Treated\}$. The distance of a DMU from its (group-specific) local frontier measures the '*managerial* efficiency', which signified the level of efficiency in terms of internal management. The distance between the local and the overall frontier captures the '*program* efficiency', which emphasizes the level of efficiency linked to the fact that the units belongs or not to the treated group. Accordingly, it can be interpreted as the causal effect of the policy intervention on efficiency. In this way, we are successful in distinguishing the extent to which the overall performance of a DMU is due to its own internal managerial efficiency and to the policy impact.

As for the frontier estimation of the production process, we rely on a nonparametric formulation. Specifically, the current study considers the robust Free Disposal Hull (FDH) model also known as *order-m* (Deprins and Simar 1984; Cazals et al. 2002; Daraio and Simar 2005) for a number of reasons. First of all, being fully nonparametric, it avoids imposing any specific parametric assumption, which is preferable, as we do not a priori observe the exact relationship between inputs and outputs. This avoids specification biases and remains consistent with the nonparametric approach proposed in the previous step for the Regression Discontinuity Design. Second, it reduces the impact of atypical observations (outliers or measurement errors). Instead of the full frontier obtained enveloping all the observations, we construct a partial frontier focusing on a subsample of *m* DMUs randomly drawn from the full sample of *n* observations. In this way, the influence of outlying or extreme observations can be mitigated and the estimates are more robust compared to those obtained with the standard FDH methodology. Third, it allows for multiple inputs and outputs simultaneously: there is no need for restrictive choice in inputs and outputs as required in other model specification. Fourth, it does not assume any convexity, which otherwise might lead to unfeasible input-output combinations. Fifth, it has interesting asymptotical properties and tests (Kneip et al. 2015, 2016).

More formally, following Daraio and Simar (2007a), the input-oriented order-m efficiency estimator ($\hat{\theta}_{m,n}^s$) for an observation i is defined in its probability formulation as follows:

$$\hat{\theta}_{m,n}^s(x, y) = \int_0^\infty (1 - \hat{F}_{X|Y,n}(ux|y))^m\, du \tag{2.5}$$

where $s = \{Control, Treated, Overall\ h\%\ discontinuity\ sample\}$, n is the size of the sample from which $m < n$ units are drawn, x the inputs and y the outputs. The obtained efficiency score per unit reflects the extent to which the unit succeeds in converting its multiple inputs into multiple outputs. Due to the subsampling, there might arise 'super-efficient' observations: these units are more efficient than the average of *m* units producing at least their level of output and randomly drawn from the full sample of *n* units (Daraio and Simar 2007a).

Step 3. Detecting the environmental variable influence: a Conditional approach

Environmental variables, beyond the control of the observations' management, affect not only the distribution of the efficiency scores, but also their attainable sets (Cazals et al. 2002; Daraio and Simar 2005, 2007b; De Witte and Kortelainen 2013).[5] Thus, in the third step, heterogeneity in the estimation of the production frontier of step 2 is included. Using a conditional efficiency framework, the efficiency estimates are not only determined by the inputs (*x*) and the outputs (*y*), but also by the environmental variables (*z*) under a non-separable production context (Cazals et al. 2016). Following Daraio and Simar (2007a), the input-oriented conditional order-m efficiency estimator ($\hat{\theta}^{s}_{m,n}$) is defined in its probability formulation as follows:

$$\hat{\theta}^{s}_{m,n}(x,y|z) = \int_{0}^{\infty} (1 - \hat{F}_{X|Y,Z,n}(ux|y,z))^{m}\, du \qquad (2.6)$$

where $s = \{Control, Treated, Overall\ h\%\ discontinuity\ sample\}$, n is the size of the sample from which $m < n$ units are drawn, x the inputs, y the outputs and z the contextual variables. For this estimation, a nonparametric kernel function and a bandwidth parameter *b* have to be selected using smoothing techniques, properly handling discrete and continuous environmental variables. Due to the subsampling, there might arise 'super-efficient' observations, as the evaluated observation is not necessarily part of the reference set.

It should be noticed that this further step is not redundant with respect to the regression discontinuity design approach, but rather complementary as it addresses different aspects. First, as in the spirit of the RDD, the environmental characteristics that are not pre-determinants of the treatment status should not be statistically different across the treated and the control groups, but nonetheless are included in the regression to provide more accurate estimates (Calonico et al. 2016; Lee and Lemieux 2010). Second, the direct inclusion of the environmental variables handles left heterogeneity across the treated and the control samples (especially for the upper bound of the optimal bandwidth range because it covers units farther from the threshold). Third, an additional source of information can be obtained while performing the conditional analysis. By comparing the conditional and the unconditional efficiency estimates

$$Q^{s,z}_{m} = \hat{\theta}^{s}_{m,n}(x,y|z)/\hat{\theta}^{s}_{m,n}(x,y) \qquad (2.7)$$

we can causally evaluate the direction of influence of environmental variables on the production process by performing a nonparametric statistical inference (Bădin et al. 2012; Daraio and Simar 2007a p. 115). By definition, the environmental variables are non-discretionary; therefore in principle the DMUs cannot directly change them as they would. However, knowing the influence of these variables can help the policy makers to enact more targeted interventions and provide further help.

[5] In the efficiency literature alternative interpretations of the "environmental variables" can be found. For example, Doms and Bartelsman (2000) define "factors behind the patterns" the forces that can influence the production processes. O'Donnell et al. (2017) distinguish between the characteristics of the production environment defined as variables that are physically involved in the production process and the characteristics of the market or institutional environment. More examples are in Daraio and Simar (2007a). In the current approach, we consider the environmental variables in their broadest sense, namely variables which are not under the control of the managers and that affect both the attainable set and the distribution of the efficiency scores, without making any *a priori* distinction of the variables at hand.

3. Empirical application to secondary schools

This section applies the procedure described in Section 2 to evaluate the causal impact of additional funding for schools with disadvantaged students on school performance. As a starting point, we use the educational production function (Levin 1974; Hanushek 1979, 2002), which models the conversion of multidimensional inputs (e.g., school resources, peers, innate ability, motivation) into educational outcomes (e.g., student achievement, attendance rate, job market success). The educational production is deemed to be efficient when the observed outputs are generated using the lowest amount of resources (or alternatively if the observed inputs are transformed into the highest amount of outputs).[6] However, endogeneity issues might arise from various sources when estimating the educational production function (Cazals et al. 2016; Cordero et al. 2015; Santín and Sicilia 2017a, c; Simar et al. 2016) and this occurs quite often in the education sector (Cordero et al. 2015; Mayston 2003). For example, there could be a potential impact of unobservable factors that correlate with the measured variables, such as the innate ability of the student, motivations or other family information that might not be retrieved. There might be problems of self-selection wherein the parents decide the schools for their children's' enrollment or teachers subjective choice of selecting a school, confounding the real underlying production process. There also might be reinforcing mechanisms in the allocation of school resources as, for example, in the allocation of additional funding or good teachers, leading to reverse causality (De Witte and López-Torres 2017). In addition, endogeneity issues might arise in the attempt to disentangle a component attributable to the context or the program under which a school operates from a component related to its internal managerial characteristics, leading to biased program/managerial efficiency estimates and preventing from causal interpretation of the findings. Accordingly, unsolved endogeneity problems might explain contradicting findings about the efficiency and effectiveness of school resources on educational outcomes (Hanushek 2006; Jackson et al. 2016).

3.1 The 'Equal Educational Opportunities' program

The Flemish Community of Belgium strives to ensure the presence of equal educational opportunities over the last decades (Nusche et al. 2015) for various reasons. According to the OECD PISA surveys, Flanders experiences a high disparity in basic skills and achievement, largely explained by the student socio-economic background (OECD 2013, 2017a). The performance gap for students with a migrant background is the highest in the OECD; this gap is furthermore enhanced due to uneven distribution of experienced teachers (Nusche et al. 2015). Moreover, in the Flemish Community of Belgium, there is large segregation in schools determined by secondary school track choice. Though in theory, the choice between tracks adds up to the abilities and ambitions of the students, general education is still considered as the most prestigious choice rather than one entail with vocational education. In the absence of standardized exams, this creates segregation in schools (De Witte and Hindriks 2017). Also, the school population in the Flemish Community is increasingly heterogeneous in terms of poverty, language, culture and family structure. Projections suggest that the population growth will be concentrated in disadvantaged groups, mainly consisting of first and second-generation migrants. Therefore, the equity challenge is noteworthy and could even worsen in the next years (European Commission 2017).

---

[6] For a comprehensive overview of the different levels of analysis, the main inputs/outputs/contextual variables and the methodological approaches considered in the efficiency in education literature, we refer to the recent reviews by Johnes 2015, De Witte and López-Torres 2017, G. Johnes et al. 2017, J. Johnes et al. 2017.

The ‘Equal Educational Opportunities (“*gelijkeonderwijskansenbeleid*, GOK”) program’ promoted by the Flemish Ministry of Education was initiated in 2002. According to the policies of the program, additional funding is provided to secondary schools with a significant number of disadvantaged students. Though there is considerable freedom for the use of funding, these extra resources can only be used for hiring additional teachers and teacher support (hence, equivalently expressed in teaching hours). The criteria for being considered a “disadvantaged” student slightly changed over the years. Before 2008, the focus was more educational outcome oriented, however, since then, the definition of a disadvantaged student has shifted its focus to the background characteristics of the students in order to support those who hail from a low-economic background. Specifically, five indicators are considered: (i) the student receives an educational grant (proxy for the family income); (ii) the student’s mother does not have a secondary education degree (proxy for parental educational background); (iii) the student lives outside of family; (iv) the parent is part of the travelling population; (v) the student does not speak Dutch (i.e., the native language) at home. Thus, a school is liable for additional teaching hours if a weighted share of students meets at least one of these indicators and it exceeds an exogenously set threshold. For the first stage of secondary education (first two years), the cutoff is set at a minimum share of 10% disadvantaged students. For the second and third stage of secondary education (last four or five years), the cutoff level is at 25%. The difference in the threshold for the first and the second/third stage is due to historical reasons (Nusche et al. 2015). The total amount of additional funding assigned to a school is decided every three years, on the basis of the amounts and the type of disadvantaged students per school in the year before the start of the three-year cycle. Moreover, to avoid fragmentation of resources, eligible schools receive the extra funding only if they generate at least six teaching hours. Further details on Flemish education system and the program are provided in Appendix A.

The empirical analysis of the current study is focused on the second and third cycle of secondary education whose cutoff is set at 25%. Also, to avoid redundancy, following this juncture, the second and third cycle of secondary education is referred as to secondary education.[7]

3.2 Data and variables

We use an unique dataset of 642 secondary schools covering the school year 2011/2012, starting year of a new three-year cycle, and representing more than 90% of all the secondary schools in Flanders. The Flemish Ministry of Education provided us with rich data at pupil and school level. At the student level, data contain information on the disadvantaged student indicators, student characteristics (e.g., gender, nationality) and field of study. Furthermore, we have information on educational outcomes that involve the short term (problematic absenteeism, grade retention and certificate obtained at the end of the school year) and the long term (enrolment in higher education). At school level, the collected data include information on the percentage of disadvantaged students, school location, educational track (general, technical, vocational or artistic education), school size, whether the school received additional funding in the previous years, amount of operational grants, teacher information (e.g., gender, degree, seniority) and number of teaching hours.

[7] At a threshold of 10% it is more likely to have non-compliers (eligible but not treated) due to the second eligibility criteria: even if the observed share of disadvantaged students might be above the set threshold for determining treatment eligibility, it might not be enough to generate a minimum of 6 hours.

3.2.1 Inputs

School funding resources are essentially provided across three categories: staffing hours, operating grants and capital (Nusche et al. 2015). However, for the current study, capital expenditure has not been considered for the cross-sectional focus of the analysis; therefore, we use two input variables obtained from the administrative data. The first variable is *teaching hours per student*, which measures the number of total teaching hours, keeping in consideration both the standard teaching hours and the extra conducted for disadvantaged students (if any). As discussed earlier, the change in inputs due to the policy might result in spillover effects on the production technology or on the internal management efficiency; thus, the additional teaching hours cannot be ignored, but rather accounted for (see also Section 2 – Step 2). As a second variable, we use the *operating grants per student*, which measures the total budget distributed among schools to cover their expenses. To reduce the variability across the units under analysis, we consider the amount of teaching hours and operating grants per student. The two inputs are expressed in ratios, which are not a matter of concern given the FDH model adopted for the frontier estimation (Olesen et al. 2015, 2017).

3.2.2 Outputs

In the efficiency of education literature, educational outcomes have been measured as student achievement or more generally student engagement, focusing both on the short-term and long-term benefits (De Witte and López-Torres 2017). For the purpose of analysis, four varied outputs have been considered to represent all these aspects. The first output is the s*hare of students that can progress to the next school year without any restrictions,* which measures the proportion of students that obtain 'A certificate'. In the absence of standardized test scores, 'A certificate' serves as a good proxy for student performance. At the end of the school year, each student receives three types of certificates, namely, "A", "B" or "C", on the basis of their respective final school exam session. A student obtaining an "A certificate" is allowed to progress to the following year level without any restrictions in the program. In the latter two scenarios, the student can progress but only in specific programs or has to repeat the year. The second output variable consists of the s*hare of students without problems of absenteeism.* This output quantifies the proportion of students that are not problematically absent, that is students who have not missed school for more than 30 half school days. This variable signifies the engagement of students in school in educational activities, promoting better learning in the short term and lifetime opportunities in the long term.[8] The third output mentions the s*hare of students without grade retention* which can be considered as the complement of grade retention (Rosenfeld 2010). Accordingly, this variable measures the proportion of students that progress through school without experiencing grade retention in secondary education. It should be noted that 24% of the 15-years old in Flanders experienced grade retention, which is double from the OECD average. Finally, the s*hare of students enrolled in higher education* measures the proportion of students that started either an academic or professional bachelor. This output considers the role of school in providing enough encouragement for students to focus their attention on higher education and pursuing lifelong opportunities.[9]

---

[8] https://www.brookings.edu/research/going-to-school-is-optional-schools-need-to-engage-students-to-increase-their-lifetime-opportunities/

[9] It should be noted that although the share of students that can progress to the next school year without any restrictions captures how the school promotes the student attainment and the share of students without problems of absenteeism captures the student engagement, the share of students without grade retention

3.2.3 Contextual variables

The educational production function is influenced by characteristics that are not under the direct control of the school management but must be controlled in the analysis (Haelermans and Ruggiero 2013, 2017; Ruggiero 2000). Three groups of contextual variables have been identified – school, teacher and student characteristics.

School characteristics

First, consider *school track*. Students can choose among four tracks: general, artistic, technical and vocational secondary education. General education is perceived as the most prestigious track while vocational is considered as the least one. This apparent division generates segregation in student allocation across the schools, which are mostly observed in differences in the average socio-economic levels. To understand and capture this phenomenon, we consider a dummy variable equal to one if the school offers general secondary education (*School track – General education*).

Second*,* among the literature catering to education economics, the importance of *school size* has been stated with considerable relevance. There has been a noticeable relationship between the school size effects and the possible existence of scale economies in the literature. Interestingly, the evidence can be mixed if looking at the student socio-economic characteristics (Leithwood and Jantzi 2009). School principals cannot refuse student enrolments by law (unless the school faces capacity restrictions); consequently, school size is an exogenous variable that is not under the control of the school management. However, this still affects the manner in which schools alter resources into educational outcomes and, therefore, it is worth controlling for it.

Third, the *share of students changing school* measures the share of students that change their school and enroll themselves in a different school in the next year. This variable captures how many students leave the school or are pushed away from the school they are currently enrolled in, and, as such, it may serve as a proxy for selection in and of schools.

Fourth, *previously treated school* is a dummy equal to one if the school received additional teaching hours in the previous three-year cycle (started in the school year 2008/2009). In this manner, we can handle the influence on the school management of being already a recipient of extra resources. This influence might work in two different directions – the school understands that they can employ their resources in a better manner in the new cycle which is the "learning effect", or the provision of additional resources hamper the management and create a "wealth effect".

Fifth, *private education* refers to the educational networks that act as "umbrella organization" for the school governing bodies: public education organized by the central government, public education organized by municipalities or provinces, and private education. These networks differ mainly in the competent government authority and the manner in which they are managed, that is, either publicly or privately. However, despite the mentioned educational networks, schools have to attain the same general goals.

Finally, *school with special need students* is a dummy variable equal to one if the school is eligible for additional funding to support integration of special need students.

Teacher characteristics

The role of teacher quality and school principals in the pedagogical domain has been increasingly acknowledged (Hanushek and Woessmann 2015; OECD 2017b; De Witte and Van

---

embeds partly both the aspects in a complementary fashion. The rather low correlation coefficients (0.6359, 0.3932, 0.3784) further prove this statement.

Klaveren 2014; De Witte and Rogge 2011) and, thus, has to be taken into account for the analysis of the current study.
The variable of *teacher seniority* measures the experience of teachers in a respective school; it ranges from 1 to 7, wherein 1 refers to the least experienced teachers (0-5 years) and 7 to the most experienced ones (>30 years). The second variable *teacher diploma* quantifies the share of teachers that have the precise diploma to teach the subject they are assigned to ("vereiste bekwaamheidsbewijzen") or one at a similar level ("voldoend geachte bekwaamheidsbewijzen"), as opposed to another type of diploma representing the minimum level required for teaching. The third variable mentions s*chool principal seniority* that measures the seniority of school principals and is measured in a similar manner as to the experience of teachers; it ranges from 1 to 7, where 1 refers to the least experienced and 7 to the most experienced school principal. The fourth variable is the *teacher age*, which ranges from 1 to 8, where 1 refers to the youngest teachers (<30 year old) and 8 to the oldest ones (60+). The fifth variable, which is, t*eacher full-time* represents the share of teachers that have a full-time contract, as opposed to a part-time contract. Finally, *female teachers* is the share of female teachers working in a school.

Student characteristics
The student population of the school has been proxied with the help of the following three variables. The s*hare of students with grade retention in primary school* measures the share of students that experienced grade retention in primary school, and can be perceived as a proxy for the cognitive skill of the pupil. The s*hare of special need students in primary school* posits as a representative for pupil's cognitive skill the school has to deal with. Third, the s*hare of male students* measures the proportion of male students in a school. Earlier evidence highlights the difference between the performance of male and female students and accordingly, this study includes this characteristic (Cipollone and Rosolia 2007).

4. Results

4.1 Step 1: a Regression Discontinuity Design approach

To evaluate the causal impact on efficiency of additional funding provided to schools, we exploit the cutoff exogenously set at 25% share of disadvantaged students in the second and third cycle of secondary education. Observations right above and below the 25% cutoff are selected by the CCT optimal bandwidth (Calonico et al. 2014a). Since four outputs have been considered for the main analysis, there are four selected bandwidths ranging between 6% and 8% (for more details see Appendix B.1). Without loss of generality, the researchers can focus the analysis on the extreme optimal bandwidth values, 6% and 8%. Thus, the 6% discontinuity sample, as the smallest focus on observations, is obtained along with the 8% discontinuity sample, as the largest one. To focus the discussion, we provide critical discussion for the 6% discontinuity sample in the main text, while the results are provided for the 8% discontinuity sample in Appendix D. For completeness, in Appendix F we provide the 'naive' estimates for the full sample, with the caveat that here endogeneity issues have not been dealt with and accordingly we cannot give causal interpretation of the findings, differently from what we intend to do in the following by applying the proposed approach.

To provide a sound causal interpretation, it is crucial to validate the established RDD setting; given that schools above the threshold receive additional resources, there might be manipulation around the threshold. Although this is unlikely due to the use of administrative data to crosscheck multiple indicators used in determining the percentage of disadvantaged students, however we check whether there is sorting around the threshold. As a first indication for manipulation, we test if the baseline characteristics around the threshold are similar. Close to the cutoff, the schools in the control and treatment group should be similar, except for the treatment.[10] Table 1 suggests that the two groups are not statistically different in means for all the control variables considered, but for few exceptions. These exceptions are mostly related to student characteristics such as the share of disadvantaged students and the share of special needs students in primary school, which will serve as contextual variables in the current analysis. It has also been observed that the schools below the threshold tend to focus more on general education schools and not treated before. Moreover, Table 2 signifies that the treated group has, on average, a higher level of inputs, but a lower level of outputs. On the one hand, the difference in inputs and outputs may be a consequence of the different share of pupils in school tracks between the control and treated group. In a similar way, there are differences in the operating grants and the outputs between general and the other school tracks.[11] This is indicative of the occurrence of inefficiency in the treated group. However, the analysis proposed by this paper helps in measuring the efficiency from an input/output mix perspective, disentangling the source of this inefficiency and detecting the possible mechanisms behind the observed picture.

Table 1. Sample means for control/treated group and population. Control variables.

	Below threshold		*Above threshold*		Full sample		*p*-value
School track – General education	0.794	(0.407)	0.493	(0.504)	0.640	(0.482)	0.0002
School size (log)	6.176	(0.449)	6.186	(0.476)	6.181	(0.461)	0.8916
Share of students changing school	0.0978	(0.0364)	0.0929	(0.0363)	0.0953	(0.0363)	0.4281
Previously treated school	0.221	(0.418)	0.704	(0.460)	0.468	(0.501)	0.0000
Education provider							0.561
Public education	0.191		0.197		0.194		
Public municipal education	0.074		0.123		0.101		
Private education	0.735		0.676		0.705		
School with special need students	0.441	(0.500)	0.507	(0.504)	0.475	(0.501)	0.4406
Teacher seniority	3.922	(0.348)	3.867	(0.356)	3.894	(0.352)	0.3627
Teacher diploma	0.973	(0.0308)	0.963	(0.0360)	0.968	(0.0338)	0.0879
School principal seniority	5.334	(1.119)	5.432	(1.031)	5.384	(1.072)	0.5905
Teacher age	4.188	(0.316)	4.161	(0.316)	4.174	(0.315)	0.6163
Teacher full-time	0.299	(0.109)	0.312	(0.0983)	0.306	(0.104)	0.4601
Female teachers	0.595	(0.118)	0.571	(0.123)	0.583	(0.121)	0.2318
Share of students with grade retention in primary school	0.0952	(0.0566)	0.148	(0.0654)	0.122	(0.0665)	0.0000
Share of special need students in primary school	0.0141	(0.0238)	0.0318	(0.0334)	0.0232	(0.0303)	0.0005
Share of male students	0.474	(0.161)	0.533	(0.211)	0.504	(0.190)	0.0670

[10] Again, for brevity, in this section we report the means for the 6% discontinuity sample. In Appendix D.1, there is the table listing the means for the 8% discontinuity sample.

[11] To account for similar observed differences between schools, we perform the analysis by limiting the sample to only vocational schools or general education schools. The analysis suggests robust findings to the main outcomes. Results are available upon request from the authors.

Share of disadvantaged students	0.220	(0.0188)	0.281	(0.0187)	0.251	(0.0357)	0.0000
Observations (schools)	68		71		139		

Note: Results for 6%-discontinuity sample (8%-discontinuity sample in Appendix B.2). Standard deviation in parentheses. *p*-values obtained from t-test to examine whether the control and the treated group variables are statistically different in means.

Table 2. Sample means for control/treated group and population. Input and output variables.

	Below threshold		*Above threshold*		Full sample		*p*-value
Inputs							
Teaching hours per student	2.120	(0.408)	2.389	(0.431)	2.257	(0.440)	0.0002
Operating grants per student	915.5	(82.54)	985.8	(138.2)	951.4	(119.3)	0.0004
Outputs							
Share of students progressing to next school year without restrictions	65.96	(5.261)	61.88	(6.417)	63.88	(6.206)	0.0001
Share of students without problems of absenteeism	99.68	(0.550)	99.35	(0.584)	99.51	(0.589)	0.0009
Share of students without grade retention	94.53	(2.757)	93.53	(3.431)	94.02	(3.149)	0.0594
Share of students enrolled in higher education	75.46	(15.38)	62.34	(17.37)	68.76	(17.64)	0.0000
Observations (schools)	68		71		139		

Note: Results for 6%-discontinuity sample (8%-discontinuity sample in Appendix B.2). Standard deviation in parentheses. *p*-values obtained from t-test to examine whether the control and the treated group variables are statistically different in means.

To formally test for the presence of manipulation, a McCrary manipulation test (McCrary 2008) using a Local-Polynomial Density Estimation as proposed by Cattaneo et al. (2018) has been conducted. Also, in this case, the results in Table 3 do not point to any manipulation around the threshold. In addition, we graphically check in Figure 1 the frequency distributions of the schools with respect to the assignment variable (the share of disadvantaged students) for different ranges and there is no evidence of any sorting around the threshold.

Table 3. McCrary manipulation test.

	Bandwidths		Number of schools		Test	
	Below	*Above*	# Below	# Above	T	*p*-value
$h_{-} = h_{+}$	0.06	0.06	68	71	0.3252	0.7450
Observations in the full sample			236	406		

Note: Results for 6%-discontinuity sample (8%-discontinuity sample in Appendix B.3).

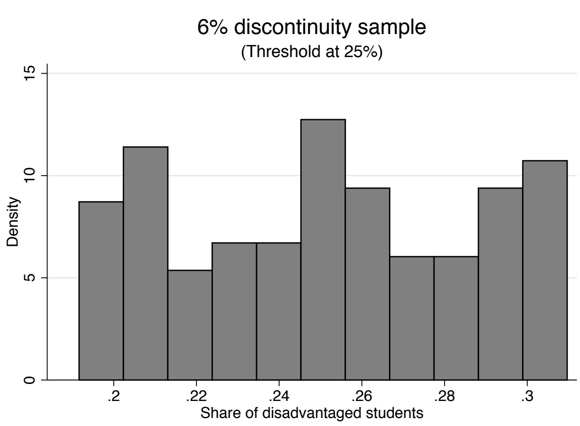

Figure 1. Frequency distribution of the schools with respect to the share of disadvantaged students

In addition, the presence of discontinuity in the probability of treatment has to be examined. Figure 2 shows the probability of treatment when the cutoff is exogenously set at 25% of disadvantaged students in a school and displays a discontinuous jump at the cutoff. The jump in the probability of treatment at the cutoff is not sharp from 0 to 1 as it would be expected in a sharp RDD setting (Lee and Lemieux 2010). We are aware of the limits that this might bring into our empirical application, but we believe also that this is not a matter of concern for two main reasons. First of all, the imperfect compliance observed is due to the additional requirement of generating a minimum of 6 hours, which can be easily excluded as the case of imperfect take-up. Moreover, we performed as a robustness check the analysis with and without the units that are eligible but not receiving the treatment. These results are consistent (see Section 4.5). Therefore, we are confident that the quasi-experimental data at hand are able to show the potential of the tool proposed in this paper and to provide sound policy recommendations. In terms of interpretation, the imperfect compliance results in local average treatment effects. More in general, in case of perfect compliance the average program efficiency scores can be interpreted as average treatment effects, consistently with the sharp Regression Discontinuity Designs (Lee and Lemieux 2010). We consider dealing with imperfect compliance as scope for future research.

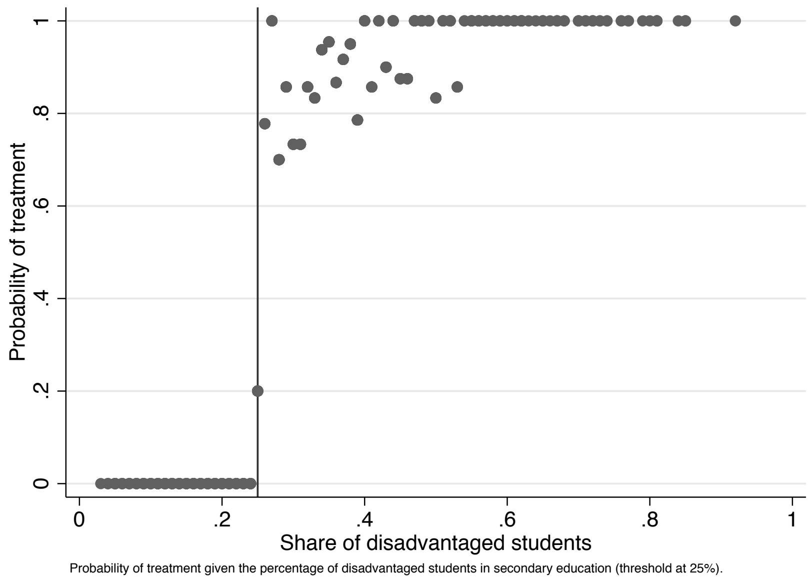

Figure 2. Discontinuity in the probability of treatment

4.2 Step 2: a Metafrontier approach

In step 2, for the groups of schools distinguished in step 1, we estimate the educational production frontier using an input-oriented robust FDH model. We compute the efficiency scores for each school under analysis following equation (2.5). As for the choice of *m*, a sensitivity analysis shows that *m=40* is warranted, even across different discontinuity samples (see plots in Appendix C). We recall that, from an economic perspective, the value *m* can be interpreted as the number of (randomly drawn) potential competing schools producing at least the same level of output as the unit under observation (Daraio and Simar 2007a). First, we estimate the pooled frontier for the whole discontinuity sample. The efficiency score indicates the overall level of efficiency of the school under analysis. Then, we estimate group-specific frontiers, separately for the treated and the control group so to disentangle the overall efficiency into a component related to managerial efficiency and another to program efficiency. The obtained efficiency scores for the group-specific frontiers measure the internal managerial efficiency level of the schools. Residually, we compute the level of program efficiency, as explained in Section 2 - Step 2.

Table 4 shows the average scores of the overall, managerial and program efficiency for the 6% discontinuity sample (results for 8% discontinuity sample are similar and presented in Appendix D.1), without controlling the operational environment (operational environment has been controlled in the next subsection). We interpret the complement to 1 of the average overall efficiency and managerial efficiency as the detected level of inefficiency. The average overall efficiency is 5 percentage points higher for control schools, but the average school-specific efficiency is about 2 percentage points higher for treated schools. This suggests that treated schools have a more homogenous production technology (i.e., their efficiency scores are closer to each other). However, the overall efficiency level is lower among the treated schools pointing at the presence of a higher waste of resources, that is almost 20% (obtained as 1-0.803) versus 14% (obtained as 1-0.855), and this can be explained by the program efficiency component.[12] A program efficiency score for the treated schools lower than 1 denotes that the treated-specific frontier is further from the overall frontier compared to the control-specific frontier. This puts forth the notion that treated schools do not successfully convert more resources into more outputs around the threshold; the schools could have achieved similar output with less amount of resources as observed for similar but untreated schools.

Thanks to the regression discontinuity setting, we can go beyond the correlation interpretation of the findings and provide instead causal inference: around the threshold the extra resources allocated because of the policy intervention do not promote a better overall school performance. The program efficiency of the untreated schools amounts to 1.002, suggesting that the untreated schools are mainly constituting the metafrontier.[13] As discussed before, this estimate can be considered as a local average treatment effect.

To check if the differences in performance between the control and the treated group are statistically different, we complement the analysis with a non-parametric statistical test (Charnes et al. 1981, Vaz and Camanho 2012). The non-parametric Wilcoxon–Mann–Whitney has been performed to examine whether the control and the treated groups are from populations with the same distribution: *p*-values are reported in Table 4. Alternative tests are available, but they are not appropriate for decomposed efficiency scores (Kneip et al. 2016).

Table 4. Descriptive statistics of the efficiency scores.

	Below threshold				*Above threshold*				*p*-value
	mean	sd	min	max	mean	sd	min	max	
Unconditional									
Overall efficiency	0.8554	0.0837	0.6504	1.0000	0.8026	0.0996	0.4945	1.0000	0.0051
School efficiency	0.8538	0.0848	0.6434	1.0000	0.8789	0.1151	0.5192	1.0007	0.0248
Program efficiency	1.0021	0.0029	1.0000	1.0108	0.9160	0.0560	0.7272	1.0000	0.0000

[12] It should be noted that the results for focusing on general and vocational schools only suggest similar findings.

[13] Recall that efficiency scores > 1 point to 'super-efficient' observations, which is due to the resampling technique discussed in Section 2. A score of 1.002 can be interpreted as the schools are performing 0.2% better than expected.

Observations	68	71

Note: Results for 6%-discontinuity sample (8%-discontinuity sample in Appendix D.1). *p*-values obtained from the non-parametric Wilcoxon–Mann–Whitney test to examine whether the control and the treated groups are from populations with the same distribution.

4.3 Step 3: a Conditional approach

Environmental variables play a prominent role in the educational production process estimation (Brennan et al. 2014; Cherchye et al. 2010; Cordero et al. 2017; Johnes 2015). These variables have been often included in a two-stage procedure that implicitly assumes a "separability condition" (Daraio and Simar 2007b), which seems an unrealistic assumption in educational applications. For instance, if schools with more low SES students receive more resources, the separability condition is violated as SES directly affects the educational production process. In coherence with the mentioned reason, a non-separable production context has been opted in the current study; hence, a robust conditional model inclusive of the contextual variables in the frontier specification has been estimated.

Table 5 shows that, in line with the insights provided by the regression discontinuity design, addition of the contextual variables in the frontier estimation does not alter the findings incurred in step 2 (even if the conditional estimates are higher than the unconditional ones): this holds for the discontinuity samples obtained considering the range of optimal bandwidths computed in step 1. As in the unconditional efficiency estimates, program efficiency scores are systematically lower for treated schools rather than for the control ones.

Nevertheless, by systematically adding control variables to the analysis, the results in Table 5 suggest that including school characteristics influences the obtained efficiency scores most. For example, in model specification 2 (teacher characteristics) the average difference in program efficiency between the control and treated groups almost vanishes to as little as 3.9 percentage points, although the variation in the program efficiency scores remains larger in the treated schools. In the most elaborated model specification 4 (School, Teacher and Student characteristics), although the variation in the program efficiency is larger for the treated schools, the average difference in program efficiency between the control and treated group amounts to 2.5 percentage points. This suggests that the policy did not improve the efficiency of the treated schools, but did not harm them as well.

Again, to check if the differences in performance between the control and the treated group are statistically different, the non-parametric Wilcoxon–Mann–Whitney test was performed to examine whether the control and the treated groups are from populations with the same distribution: *p*-values are reported in Table 5 and suggest significant differences.

Table 5. Descriptive statistics of the efficiency scores.

	Below threshold				*Above threshold*				*p*-value
	mean	sd	min	max	mean	sd	min	max	
Conditional 1 - School characteristics									
Overall efficiency	0.9093	0.0770	0.6962	1.0000	0.8570	0.0990	0.6154	1.0000	0.0024
School efficiency	0.9079	0.0736	0.6941	1.0000	0.9296	0.0829	0.6811	1.0000	0.0182
Program efficiency	1.0014	0.0201	0.8942	1.0720	0.9218	0.0649	0.7719	1.0022	0.0000
Conditional 2 - Teacher characteristics									
Overall efficiency	0.9639	0.0607	0.7336	1.0000	0.9287	0.0935	0.5337	1.0000	0.0071
School efficiency	0.9463	0.0693	0.7190	1.0000	0.9477	0.0812	0.5391	1.0000	0.3976
Program efficiency	1.0201	0.0418	0.9373	1.1810	0.9808	0.0672	0.7630	1.2296	0.0000

Conditional 3 - Student characteristics									
Overall efficiency	0.9180	0.0787	0.7330	1.0000	0.9138	0.0924	0.4911	1.0000	0.9562
School efficiency	0.9278	0.0773	0.7450	1.0000	0.9508	0.0787	0.5241	1.0000	0.0112
Program efficiency	0.9901	0.0393	0.8262	1.0904	0.9615	0.0611	0.7861	1.1011	0.0007
Conditional 4 - School & Teacher & Student characteristics									
Overall efficiency	0.9664	0.0466	0.8559	1.0000	0.9603	0.0607	0.6767	1.0000	0.6123
School efficiency	0.9682	0.0469	0.8532	1.0000	0.9860	0.0314	0.8234	1.0000	0.0018
Program efficiency	0.9984	0.0203	0.9163	1.0574	0.9736	0.0493	0.7908	1.0552	0.0003
Observations	68				71				

Note: Results for 6%-discontinuity sample (8%-discontinuity sample in Appendix D.1). *p*-values obtained from the non-parametric Wilcoxon–Mann–Whitney test to examine whether the control and the treated groups are from populations with the same distribution.
The conditional models include the following variables:
Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students
Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers
Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students
Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority, Teacher diploma, % students with problems in primary school, % students with special needs in primary school

In summary, according to the evidence incurred by the analysis pursued so far, treated schools do not successfully convert the additional resources to perform better around the threshold, unless school and pupil characteristics are accounted for. Stated differently, resources allocated where there is a relatively small share of disadvantaged students (25% cutoff) and/or a little amount of resources seem to miss to the desired policy outcome.

4.4 Statistical inference

Next, we analyze by a conditional efficiency model the statistical inference by comparing conditional and unconditional estimates along the contextual variables of interest, by means of a nonparametric regression and considering 2000 bootstrap samples. This can be utilized to explore the direction of the influence of these variables with respect to the efficiency assessment. To reduce the course of dimensionality, only few variables per model specification are included. Table 6 summarizes the main findings obtained for the different conditional models considered above, listing the median influence of the contextual variables and the p-values for the significance tests (Li and Racine 2007). Graphically, the smoothed regression line can be interpreted as the marginal effect of the contextual variable under focus on the attainable set. For a more intuitive interpretation of the findings, we consider the ratio of unconditional over conditional estimates: if the smoothed nonparametric regression is increasing, then the variable is favourable to the efficiency, otherwise the opposite holds (De Witte and Schiltz 2018).

The model specifications that include school characteristics reveal that secondary schools providing general education have a favorable influence on the efficiency. This is not surprising as more disadvantaged students will be concentrated in vocational schools, creating a more problematic context where to promote school engagement compared to the other schools, and as vocational schools receive more inputs. Corresponding with the evidence generated, the share of students that change their school in the next year plays an unfavorable influence on the education production, as they are the most problematic ones and for this reason somehow pushed away. As revealed from the nonparametric regression plot, the private education schools have a favorable influence and the opposite holds for the schools which had received additional resources in the previous three-year cycle, signifying the aspects of a lack of learning effect in management of these extra resources. Moreover, the favorable influence that emerges for the school size points at the presence of scale economies in the educational production, or, alternatively, it might capture the decreasing input coefficients of the financing mechanism. In regards to the teacher characteristics, teachers holding a significant amount of experience plays a significant role on efficiency both from the teacher and the school principal side. The same applies when teachers have a diploma specifically related with the topic they teach; this favors the education delivery. Having a full time contract and the teacher age instead play an unfavorable role. All student characteristics in the analysis play an unfavorable influence; it is more likely that schools where students experience grade retention in primary education or students in special need schools face more problematic students and, therefore, face an unfavorable environment for the education production. As for the variables measuring the share of males in class, the evidence is consistent with the literature, as there is evidence that females outperform male student quite often (Cipollone and Rosolia 2007).

Table 6. Direction of the influence of the contextual variables.

	Conditional 1			Conditional 2			Conditional 3			Conditional 4		
	Influence	p-value		Influence	p-value		Influence	p-value		Influence	p-value	
School characteristics												
General education	Favorable	0.2605	***							Favorable	0.0035	***
School size	Unfavorable	0.134								Favorable	0.364	
% Change school	Unfavorable	0.0035	***							Unfavorable	0	***
Previously treated	Favorable	0.2605								Unfavorable	0.0505	*
Private education	Favorable	0.093	*									
Special needs school	Favorable	0.1585										
Teacher characteristics												
Teacher seniority				Favorable	0.193					Favorable	0.096	
Teacher diploma				Favorable	0.202					Favorable	0.758	
Teacher age				Unfavorable	0.2635							
School principal seniority				Favorable	0.04	**						
Teacher contract				Unfavorable	0.029	**						
% female teachers				Favorable	0.0055							
Student characteristics												
Primary retention							Unfavorable	0.9975		Unfavorable	0.135	
Special students in primary							Unfavorable	0.1825		Unfavorable	0.134	
% Man							Unfavorable	0.5795				

* p < 0.10, ** p < 0.05, *** p < 0.01

Note: Results for 6%-discontinuity sample (8% discontinuity sample in Appendix D.2).

The conditional models include the following variables:

Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students

Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers

Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students

Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority & diploma, % students with problems in primary school, % students with special needs in primary school

4.5 Robustness checks

To test the robustness of the results, we perform several analyses on subsamples. By using the subsamples, we explicitly compare 'like with likes'. First, to account for the presence of imperfect compliance, the main analysis is performed excluding the eligible but not treated schools. The results of this analysis are listed in Appendix E. A second series of robustness tests examines the sensitivity of the results with respect to the underlying (un)observed heterogeneity. As schools at both sides of the exogenously set threshold might have different characteristics which remain unobserved to the researcher, or as the treatment might have heterogeneous effects in different types of schools, the sample is limited to only vocational or only general education schools.

On average, program efficiency scores are lower for treated schools and there is a higher average overall efficiency among control schools. This analysis signifies that schools fail to convert resources into more outputs, even when the eligible but not treated schools are excluded. Controlling for the school and pupil characteristics significantly reduces the gap in the program efficiency scores, making the scores reach a point where the difference is no longer significant. This suggests that the policy did not improve the efficiency of the treated schools, but did not harm them as well. Overall, results seem to be very robust. This gives us confidence that schools receiving additional resources and located just above the threshold do not successfully convert them into more output. Nevertheless, accounting for the school and pupil characteristics, the difference in program efficiency largely disappears.

5. Discussion and policy implications

This paper proposed an innovative approach to evaluate the causal impact of a policy intervention on efficiency, by combining insights from impact evaluation techniques and the standard efficiency analysis. Specifically, we designed a three-step procedure that can be utilized whenever the treatment status depends on an exogenously set threshold. In the first step, we focus on the observations around the threshold to handle potential endogeneity issues and, accordingly, we define a discontinuity sample in the spirit of a regression discontinuity design (RDD). In such a manner, we distinguish two groups of units very similar in their baseline characteristics but different in the treatment (treated *versus* untreated). In the second step, we adapt the concept of the nonparametric metafrontier approach to decompose the overall efficiency into a `managerial' and a `program' efficiency component. To do so, we estimate both a group-specific local production frontier for each group and a pooled production frontier for the discontinuity sample: the program efficiency is obtained residually by comparing the latter with the former. In the third step, we include heterogeneity in the estimation of the production frontier of step 2 by proposing a conditional analysis. Furthermore, the comparison between conditional and unconditional estimates leads to insightful statistical inference, detecting the direction of the influence of the contextual variables under a non-separable production context. Due to the quasi-experimental setting introduced in step 1, casual interpretation to the estimates can be granted.

We showcase the practical usefulness of the devised methodology evaluating the causal impact on school performance of the 'Equal Educational Opportunities' program, promoted by the Flemish Ministry of Education in Belgium since 2002 to support schools with (a large share of) disadvantaged students in secondary education. Specifically, the program assigns additional resources to the schools that exceed the 25% exogenously set threshold of disadvantaged students. To validate the regression discontinuity setting, a number of checks that indicated the absence of manipulation around the threshold were performed. For the educational production frontier estimation, we considered two inputs (namely the *total teaching hours per student*, including the additional hours, and the *operating grants per student*) and four outputs (namely *Share of students progressing through school without any restrictions*, *Share of students without problems of absenteeism*, *Share of students without* *grade retention*, *Share of students enrolled in higher education*). Whereas, to go forth with conditional analysis, three sets of contextual variables were chosen, such as characteristics of schools, teachers and students.

Examining schools close to the exogenously determined cutoff level, the results indicate that additional resources do not causally influence efficiency around the threshold. In particular, the schools close to the threshold and receiving the additional resources have lower program efficiency. These results seem to be very robust to several sub-analysis (e.g. by education track and different bandwidth). Nevertheless, despite the assumption that schools close to the threshold are very similar, some observed characteristics might still be different. Using a conditional efficiency model, we account for the school, teacher and pupil characteristics. The results of the conditional efficiency analyses indicate that the difference in program efficiency largely disappears.

The proposed approach follows the idea behind the sharp regression discontinuity design, namely in presence of perfect compliance: units are eligible for the treatment and they receive it. However, further research should extend the approach to a fuzzy regression discontinuity design framework, namely in presence of imperfect compliance: this occurs whenever there are units that do not receive the treatment, even if they are eligible for it, for instance due to additional requirements that these units miss to meet or in case of imperfect take-up.

Acknowledgments

We would like to thank Johan Vermeiren, senior expert at the Flemish ministry for education, for providing us with the data and helpful information. We also owe gratitude to participants of the 4th LEER conference on Education Economics, DEA40 in Birmingham, NAPW X in Miami, Efficiency in Education Conference in Huddersfield and Budapest, AIRO 2018 in Taormina, AMASES XLII in Naples, Ana Camanho, Chris O'Donnell, Jonas Månsson, Maria Silva Portela, Tommaso Agasisti, Jill Johnes, Geraint Johnes, Dániel Horn, Daniel Santin, Gabriela Sicilia, Sergio Perelman, Fritz Schiltz, Vítezslav Titl, Steven Groenez, Melissa Tuytens, Ides Nicaise, Thomas Wouters, Jolien De Norre, Nele Havermans and the 'SONO Opvolgingsgroep' for their useful comments and insights. This research was funded by 'Steunpunt SONO' by the Flemish government. Giovanna D'Inverno also gratefully acknowledges financial support from Research Foundation – Flanders, FWO (Postdoctoral Fellowship 12U0219N).

References

Abadie A, Cattaneo MD (2018) Econometric Methods for Program Evaluation. *Annu. Rev. Econom.* 10(1):465–503.

Afriat SN (1972) Efficiency Estimation of Production Functions. *Int. Econ. Rev. (Philadelphia).* 13(3):568–598.

Agasisti T, Avvisati F, Borgonovi F, Longobardi S (2018) *Academic resilience* (Paris).

Amsler C, Prokhorov A, Schmidt P (2016) Endogeneity in stochastic frontier models. *J. Econom.* 190(2):280–288.

Angrist JD, Lavy V (1999) Using Maimodines' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Q. J. Econ.* 114(2):533–575.

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press, Princeton).

Aparicio J, Crespo-Cebada E, Pedraja-Chaparro F, Santín D (2017) Comparing school ownership performance using a pseudo-panel database: A Malmquist-type index approach. *Eur. J. Oper. Res.* 256(2):533–542.

Aparicio J, Santin D (2017) A note on measuring group performance over time with pseudo-panels. *Eur. J. Oper. Res.* 0:1–9.

Bădin L, Daraio C, Simar L (2012) How to measure the impact of environmental factors in a nonparametric production model. *Eur. J. Oper. Res.* 223(3):818–833.

Bartelsman EJ, Doms ME (2000) Understanding Productivity: Lessons from Longitudinal Microdata. *J. Econ. Lit.* XXXVIII(September):569–594.

Battese GE, Rao DSP (2002) Technology Gap , Efficiency , and a Stochastic Metafrontier Function. *Int. J. Bus. Econ.* 1(2):87–93.

Battese GE, Rao DSP, O'Donnell CJ (2004) A Metafrontier Production Function for Estimation of Technical Efficiencies and Technology Gaps for Firms Operating Under Different Technologies. *J. Product. Anal.* 21:91–103.

Brennan S, Haelermans C, Ruggiero J (2014) Nonparametric estimation of education productivity incorporating nondiscretionary inputs with an application to Dutch schools. *Eur. J. Oper. Res.* 234(3):809–818.

Calonico S, Cattaneo MD, Farrell MH, Titiunik R (2016) Regression discontinuity designs using covariates. *URL http//www-personal. umich. edu/~ cattaneo/papers/Calonico-Cattaneo-/Farrell-Titiunik_2016_wp. pdf*.

Calonico S, Cattaneo MD, Titiunik R (2014a) Robust data-driven inference in the regression-discontinuity design. *Stata J.* 14(4):909–946.

Calonico S, Cattaneo MD, Titiunik R (2014b) Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica* 82(6):2295–2326.

Camanho AS, Dyson RG (2006) Data envelopment analysis and Malmquist indices for measuring group performance. *J. Product. Anal.* 26(1):35–49.

Cattaneo MD, Frandsen BR, Titiunik R (2015) Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *J. Causal Inference* 3(1):1–24.

Cattaneo MD, Jansson M, Xinwei M (2018) *Manipulation testing based on density discontinuity*

Cazals C, Fève F, Florens JP, Simar L (2016) Nonparametric instrumental variables estimation for efficiency frontier. *J. Econom.* 190(2):349–359.

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

Charnes A, Cooper WW, Rhodes E (1981) Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. *Manage. Sci.* 27(6):668–697.

Cherchye L, De Witte K, Ooghe E, Nicaise I (2010) Efficiency and equity in private and public education: A nonparametric comparison. *Eur. J. Oper. Res.* 202(2):563–573.

Cherchye L, De Witte K, Perelman S (2018) A Unified Productivity-Performance Approach Applied to Secondary Schools. *J. Oper. Res. Soc.* In Press.

Cipollone P, Rosolia A (2007) Social Interactions in High School: Lessons from an Earthquake. *Am. Econ. Rev.* 97(3):948–965.

Cordero JM, Cristóbal V, Santín D (2017) Causal Inference on Education Policies: a Survey of Empirical Studies Using Pisa, Timss and Pirls. *J. Econ. Surv.* 00(0):1–38.

Cordero JM, Santín D, Sicilia G (2015) Testing the accuracy of DEA estimates under endogeneity through a Monte Carlo simulation. *Eur. J. Oper. Res.* 244(2):511–518.

D'Inverno G, Carosi L, Ravagli L (2018) Global public spending efficiency in Tuscan

municipalities. *Socioecon. Plann. Sci.* 61:102–113.
Dahl GB, Lochner L (2012) The Impact of Family Income on Child Achievement : Evidence from the Earned Income Tax Credit. *Am. Econ. Rev.* 102(5):1927–1956.
Daraio C, Simar L (2005) Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *J. Product. Anal.* 24(1):93–121.
Daraio C, Simar L (2007a) *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications* (Springer Science & Business Media).
Daraio C, Simar L (2007b) Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. *J. Product. Anal.* 28(1–2):13–32.
De Witte K, Hindriks J (2017) De geslaagde school.
De Witte K, Van Klaveren C (2014) How are teachers teaching? A nonparametric approach. *Educ. Econ.* 22(1):3–23.
De Witte K, Kortelainen M (2013) What explains the performance of students in a heterogeneous environment? Conditional efficiency estimation with continuous and discrete environmental variables. *Appl. Econ.* 45(17):2401–2412.
De Witte K, López-Torres L (2017) Efficiency in education: A review of literature and a way forward. *J. Oper. Res. Soc.* 68(4):339–363.
De Witte K, Rogge N (2011) Accounting for exogenous influences in performance evaluations of teachers. *Econ. Educ. Rev.* 30(4):641–653.
De Witte K, Schiltz F (2018) Measuring and explaining organizational effectiveness of school districts: Evidence from a robust and conditional Benefit-of-the-Doubt approach. *Eur. J. Oper. Res.* 267(3):1172–1181.
De Witte K, Titl V, Holz O, Smet M (2017) Funding formulas in compulsory education. Report for the German Community of Belgium.
Deprins D, Simar L (1984) Measuring labor inefficiency in post offices. *Perform. Public Enterp. Concepts Meas. M. Marchand, P. Pestieau H. Tulkens (eds.), Amsterdam, North-holl.*:243–267.
Duflo E, Dupas P, Kremer M (2015) School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. *J. Public Econ.* 123:92–110.
European Commission (2017) Education and Training Monitor 2017 - Belgium.
Gibbons S, McNally S, Viarengo M (2018) Does Additional Spending Help Urban Schools? An Evaluation Using Boundary Discontinuities. *J. Eur. Econ. Assoc.* 16(5):1618–1668.
Grenet J (2013) Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws. *Scand. J. Econ.* 115(1):176–210.
Grosskopf S, Valdmanis V (1987) Measuring hospital performance: A Non-parametric Approach. *J. Health Econ.* 6:89–107.
Haelermans C, Ruggiero J (2013) Estimating technical and allocative efficiency in the public sector: A nonparametric analysis of Dutch schools. *Eur. J. Oper. Res.* 227(1):174–181.
Haelermans C, Ruggiero J (2017) Non-parametric estimation of the cost of adequacy in education: The case of Dutch schools. *J. Oper. Res. Soc.* 68(4):390–398.
Hanushek E, Woessmann L (2015) *Universal Basic Skills, What Countries Stand to Gain*
Hanushek EA (1979) Conceptual and empirical issues in the estimation of educational production functions. *J. Hum. Resour.*:351–388.
Hanushek EA (2002) Publicly provided education. *Handb. public Econ.* 4:2045–2141.
Hanushek EA (2006) Chapter 14 School Resources. *Handb. Econ. Educ.* 2(06):865–908.
Haveman R, Wolfe B (1995) The Determinants of Children ' s Attainments : A Review of Methods and Findings. *J. Econ. Lit.* 33(4):1829–1878.
Imbens G, Kalyanaraman K (2012) Optimal Bandwidth Choice for the Regression Discontinuity

Estimator. *Rev. Econ. Stud.* 79:933–959.
Jackson CK, Johnson RC, Persico C (2016) The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. *Q. J. Econ.* 131(1):157–218.
Johnes G, Johnes J, Agasisti T, López-Torres L (2017) *Handbook of Contemporary Education Economics* Edward Elgar Publishing, ed.
Johnes J (2015) Operational research in education. *Eur. J. Oper. Res.* 243(3):683–696.
Johnes J, Portela M, Thanassoulis E (2017) Efficiency in education. *J. Oper. Res. Soc.* 68(4):331–338.
Johnson AL, Ruggiero J (2014) Nonparametric measurement of productivity and efficiency in education. *Ann. Oper. Res.* 221(1):197–210.
Kerstens K, O'Donnell C, Woestyne I Van de (2018) Metatechnology Frontier and Convexity: A Restatement. *Eur. J. Oper. Res.* (xxxx):1–13.
Kneip A, Simar L, Wilson PW (2015) When bias kills the variance: Central Limit Theorems for DEA and FDH efficiency scores. *Econom. Theory* 31(2):394–422.
Kneip A, Simar L, Wilson PW (2016) Testing Hypotheses in Nonparametric Models of Production. *J. Bus. Econ. Stat.* 34(3):435–456.
Lee DS, Lemieux T (2010) Regression Discontinuity Design in Economics. *J. Econ. Lit.* 20(1):281–355.
Leithwood K, Jantzi D (2009) A Review of Empirical Evidence About School Size Effects: A Policy Perspective. *Rev. Educ. Res.* 79(1):464–490.
Leuven E, Lindahl M, Oosterbeek H, Webbink D (2007) The Effect of Extra Funding for Disadvantaged Pupils on Achievement. *Rev. Econ. Stat.* 89 (4)(November):721–736.
Levin HM (1974) Measuring Efficiency in Educational Production. *Public Finan. Q.* 2(1):3–24.
Li Q, Racine JS (2007) *Nonparametric econometrics: theory and practice* (Princeton University Press).
Månsson J (1996) Market Technical Efficiency and Ownership The Case of Booking Centres in the Swedish Taxi Market. *J. Transp. Econ. Policy* 30(1):83–93.
Mayston DJ (2003) Measuring and managing educational performance. *J. Oper. Res. Soc.* 54(7):679–691.
McCrary J (2008) Manipulation of the running variable in the regression discontinuity design : A density test. *J. Econom.* 142:698–714.
Muralidharan K, Sundararaman V (2015) The Aggregate Effect of School Choice: Evidence from a two-stage experiment in India Karthik. *Q. J. Econ.* 130(3):1011–1066.
Nusche D, Miron G, Santiago P, Teese R (2015) OECD Reviews of School Resources: Flemish Community of Belgium 2015. *OECD Rev. Sch. Resour.* (OECD Publishing, Paris).
O'Donnell CJ (2016) Using information about technologies, markets and firm behaviour to decompose a proper productivity index. *J. Econom.* 190(2):328–340.
O'Donnell CJ, Fallah-Fini S, Triantis K (2017) Measuring and analysing productivity change in a metafrontier framework. *J. Product. Anal.* 47(2):117–128.
OECD (2013) *Education at a Glance 2013* (OECD Publishing).
OECD (2017a) *Educational Opportunity for All: Overcoming Inequality throughout the Life Course*
OECD (2017b) *Government at a Glance 2017*
OECD (2017c) *The Funding of School Education: Connecting Resources and Learning*
Olesen OB, Petersen NC, Podinovski V V. (2015) Efficiency analysis with ratio measures. *Eur. J. Oper. Res.* 245(2):446–462.
Olesen OB, Petersen NC, Podinovski V V. (2017) Efficiency measures and computational approaches for data envelopment analysis models with ratio inputs and outputs. *Eur. J.*

Oper. Res. 261(2):640–655.
Oosterbeek H, Webbink D (2007) Wage effects of an extra year of basic vocational education. *Econ. Educ. Rev.* 26(4):408–419.
Pischke JS, von Wachter T (2008) Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation. *Rev. Econ. Stat.* 90(3):592–598.
Rosenfeld MJ (2010) Nontraditional Families and Childhood Progress Through. *Demography* 47(3):755–775.
Ruggiero J (2000) Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. *J. Oper. Res. Soc.* 51:906–912.
Santín D, Sicilia G (2017a) Dealing with endogeneity in data envelopment analysis applications. *Expert Syst. Appl.* 68:173–184.
Santín D, Sicilia G (2017b) Impact evaluation and frontier methods in education: a step forward. Johnes G, Johnes J, Agasisti T, López-Torres L, eds. *Handb. Contemp. Educ. Econ.* (Edward Elgar Publishing, Inc.), 211–245.
Santín D, Sicilia G (2017c) Using DEA for measuring teachers' performance and the impact on students' outcomes: evidence for Spain. *J. Product. Anal.*:1–15.
Schlotter M, Schwerdt G, Woessmann L (2011) Econometric methods for causal evaluation of education policies and practices: A non-technical guide. *Educ. Econ.* 19(2):109–137.
Shepard RW (1970) *Theory of Cost and Production Function* (NJ: Princeton University Press, Princeton).
Simar L, Vanhems A, Van Keilegom I (2016) Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *J. Econom.* 190(2):360–373.
Stephens MJ, Yang DY (2014) Compulsory Education and the Benefits of Schooling. *Am. Econ. Rev.* 104(6):1777–1792.
UN General Assembly (2015) Transforming our world: the 2030 Agenda for Sustainable Development. *New York United Nations*.
Vaz CB, Camanho AS (2012) Performance comparison of retailing stores using a Malmquist-type index. *J. Oper. Res. Soc.* 63(5):631–645.

Appendix Online - Supplementary material

Appendix A: The Flemish education system and its equal educational opportunities program

In the Flemish Community of Belgium, education is compulsory from age 6 to 18. Compulsory education includes two levels: primary (6-12 years old) and secondary (12-18 years old) education. Parents have full freedom to select any primary or secondary school for their children. In the domain of secondary education there are four ability tracks. General secondary education prepares students for higher education; artistic secondary education provides general education but with concentrated emphasis on arts; while technical secondary education takes a more technical approach, intended to provide students with the necessary skills to start a professional career, it also provides them with sufficient knowledge to enroll in higher education. This is in contrast with the vocational secondary education track that explicitly trains students for a specific occupation. In theory, the choice between the mentioned four tracks is up to the abilities and ambitions of the students, however, general education is perceived to be the most prestigious choice, whereas vocational is considered the least. In the absence of standardized exams, this creates segregation in schools (De Witte and Hindriks 2017). The segregation can be observed in the significant differences in the average SES levels among schools.

Schools are funded through their school boards. The funding of teaching hours for schools in Flanders is done on the basis of the number of students enrolled and point envelopes (De Witte et al. 2017). A key aspect in the Flemish school funding mechanism concerns the way additional funding is obtained for staff in supporting low socio-economic status students. Specifically, in secondary education, there is no adjustment in the formula, but schools might receive additional teaching hours based on the "Equal Educational Opportunities (*gelijkeonderwijskansenbeleid*, *GOK*) program", enacted in the Flemish Community of Belgium since 2002 (Nusche et al. 2015). The additional resources are assigned to school solely on the basis of an exogenously defined cutoff; thus, schools are eligible for funding only if they have more than 25% disadvantaged students in the second and third stage of secondary education (and 10% in the first stage of secondary education). Despite attaining the minimum share of disadvantaged students, schools need to apply for the funding. The eligibility criteria for defining 'disadvantaged students' shifted slightly throughout the years. Before 2008, the focus was mainly on the educational outcomes of students: a disadvantaged student was defined as a student who satisfies at least one of the following indicators. (i) The student has two or more years of grade retention; (ii) The student was part of a program for non-Dutch speaking newcomers; (iii) Students in vocational or technical education who received a school advise to repeat the year or to change their field of study. After 2008, the focus shifted to the socio-economic status of students. In particular, there are 5 equal opportunities indicators specified in the decree. To each of these indicators a weight, expressed in points is assigned. Below, find the 5 indicators with their respective point-values. The indicator school grant has 2-point values, one for students that only indicate this indicator and one (potentially together with non-Dutch home speakers) for those that indicate at least one other as well.

1. Parents belong to the travelling population (Roma, circus etc.) This indicator has a weight coefficient of 0.8 points.
2. The mother does not own a degree of secondary education. This indicator has a weight coefficient of 0.6 points.
3. The student is temporarily or permanently admitted outside of the family. This indicator has a point value of 0.8 points.

4. The family receives one or more school grants. If this is the only indicator checked, the point value is 0.4. This weight is corrected as the number of students that meet this requirement is then multiplied by 0.4417. This brings the real point value to 0.17668. When the student also checks another indicator the weight is set at 0.18 points.

5. The language the student speaks at home is not Dutch. This indicator has a weight coefficient of 0.2 points. For students that meet multiple indicators the weights are cumulative up to a maximum of 1.2 points per student.

The weight coefficient of 0.4417 for school grants also counts towards the count of weighted disadvantaged students. All other indicators are weighted as one in this regard. This calculation happens at the school level. Afterwards the points generated in the first cycle are summed and multiplied by 1.5 when the school is domiciled in the Brussels Capital Region or if the school has more than 55% disadvantaged students. If the school meets both criteria the multiplication happens twice. The total amount of points is multiplied with 0.2916 teacher hours.

The point values of students in second and third grade are also summarized. This value is than multiplied by 1.5 when the school is domiciled in the Brussels Capital Region or if the school has more than 55% disadvantaged students. If the school meets both criteria the multiplication happens twice. The total amount of points is multiplied with 0.1225 teacher hours.

A school receives the sum of these teacher hour students only if the result over all cycles yields 6 extra teacher hours or more. The calculation happens every 3 years (GOK-period) and during this period the additional hours remain the same. The extra teacher hours can be used across cycles as long as they aim to improve equal educational outcomes.

The total amount of additional funding for a school is decided upon every three years and is based on the amount and type of the disadvantaged students per school in the year before the start of the 3-year cycle. The latest cycles started in the school years 2008/2009, 2011/2012 and 2014/2015. Interestingly, schools are flexible in the use of these additional inputs. This might increase the differences in inefficiency between schools.

Appendix B: Bandwidth and manipulation tests

B.1 Optimal bandwidths

The following table lists the optimal bandwidths computed for each output under analysis using the 'rdrobust package' in Stata (Calonico et al. 2014a). Without loss of generality, we can focus on a range of optimal bandwidths between 6% and 8% and accordingly we obtain the 6% discontinuity sample, as the smallest focus on observations, as well as the 8% discontinuity sample, as the largest one.

Table 7. Optimal bandwidths. Threshold at 25% of disadvantaged students.

		Number of schools	
Outputs	Bandwidths	# Below	# Above
Share of students progressing to next year without restrictions	0.084	99	115
Share of students without problems of absenteeism	0.084	99	115
Share of students without grade retention	0.053	56	62
Share of students enrolled in higher education	0.082	97	113
Observations in the full sample		236	406

Note: Bandwidths computed using the 'rdrobust package' in Stata (Calonico, Cattaneo and Rocío Titiunik 2014).

B.2 Comparison control and treated group for different samples

The following tables list the variable sample means for the control and the treated group of schools, respectively below and above the exogenously set threshold, together with the full sample mean. The table shows the means for the 8% discontinuity sample of schools under analysis. The last column of each table reports the *p*-values obtained from t-test conducted to examine whether the control and the treated group variables are statistically different in means. Specifically, this test provides valuable information on the discontinuity samples under analysis. First, it gives a preliminary overview of the relation among the inputs and the outputs across treated and control group and gives the basis for a more in-depth analysis as suggested by this paper. Second, it checks whether control and treated groups have similar environmental characteristics and to which extent the regression discontinuity design mimics a randomized experiment.

Table 8 shows that the treated group has, on average, a higher level of inputs, but a lower level of outputs. This might suggest the presence of inefficiency in the treated group: the analysis proposed by this paper helps in disentangling the source of this inefficiency and in detecting the possible mechanisms behind the observed picture. As for the control variables, Table 9 displays that the two groups are not statistically different in means for all the variables we consider, but for few exceptions, mostly related to student characteristics: the conditional analysis is able to capture this left heterogeneity.

Table 8. Sample means for control/treated group and population. Input and output variables. 8% discontinuity sample.

	Below threshold		*Above threshold*				
	Control		Treated		Total		p-value
Inputs							
Teaching hours per student	2.127	(0.503)	2.479	(0.451)	2.316	(0.506)	0.0000
Operating grants per student	912.7	(83.53)	1005.4	(157.9)	962.5	(136.7)	0.0000
Outputs							
Share of students progressing to next year without restrictions	66.07	(5.762)	61.01	(7.864)	63.35	(7.400)	0.0000
Share of students without problems of absenteeism	99.68	(0.722)	99.26	(0.706)	99.45	(0.741)	0.0001
Share of students without grade	94.62	(3.034)	93.31	(3.329)	93.92	(3.254)	0.0045

retention							
Share of students enrolled in higher education	77.19	(14.34)	58.60	(17.11)	67.19	(18.37)	0.0000
Observations (school level)	92		107		199		

Note: Results for 8%-discontinuity sample. Standard deviation in parentheses. *p*-values obtained from t-test to examine whether the control and the treated group variables are statistically different in means.

Table 9. Sample means for control/treated group and population. Control variables. 8% discontinuity sample.

	Below Control		*Above* Treated		Total		*p*-test
School track - General	0.804	(0.399)	0.393	(0.491)	0.583	(0.494)	**0.0000**
School size (log)	6.150	(0.472)	6.193	(0.460)	6.173	(0.465)	0.5203
Share of students changing school	0.0984	(0.0472)	0.0967	(0.0380)	0.0975	(0.0424)	0.7749
Previously treated school	0.185	(0.390)	0.720	(0.451)	0.472	(0.500)	**0.0000**
Education provider							0.124
Public education	0.195		0.159				
Municipal education	0.054		0.140				
Private education	0.750		0.701				
School with special need students	0.424	(0.497)	0.551	(0.500)	0.492	(0.501)	0.0735
Teacher seniority	3.869	(0.366)	3.854	(0.361)	3.861	(0.362)	0.7772
Teacher diploma	0.965	(0.0404)	0.961	(0.0349)	0.963	(0.0375)	0.3797
School principal seniority	5.295	(1.175)	5.451	(1.002)	5.379	(1.085)	0.3150
Teacher age	4.162	(0.331)	4.170	(0.302)	4.166	(0.315)	0.8599
Teacher full-time	0.290	(0.114)	0.305	(0.0972)	0.298	(0.105)	0.3171
Female teachers	0.597	(0.111)	0.576	(0.133)	0.586	(0.123)	0.2345
Share of students with grade retention in primary school	0.0908	(0.0578)	0.162	(0.0655)	0.129	(0.0713)	**0.0000**
Share of special need students in primary school	0.0118	(0.0211)	0.0349	(0.0323)	0.0242	(0.0299)	**0.0000**
Share of male students	0.463	(0.142)	0.536	(0.240)	0.502	(0.204)	**0.0106**
Share of disadvantaged students	0.210	(0.0236)	0.294	(0.0245)	0.255	(0.0483)	**0.0000**
Observations (school level)	92		107		199		

Note: Results for 8%-discontinuity sample. Standard deviation in parentheses. *p*-values obtained from t-test to examine whether the control and the treated group variables are statistically different in means.

B.3 Manipulation tests

The following table shows the results of the manipulation test implemented using the ‘rddensity package’ in Stata (Cattaneo et al. 2018). There is no evidence of sorting around the cutoff, independently on whether we specify the bandwidth at both sides of the cutoff.

Table 10. Manipulation tests for secondary education. Threshold at 25% share of disadvantaged students

	Bandwidths		Number of schools		Test	
	Below	*Above*	# Below	# Above	T	p-value
$h_- = h_+$						
	0.06	0.06	68	71	0.3252	0.7450
	0.08	0.08	92	107	0.2151	0.8297
$h_- \neq h_+$						
	0.116	0.096	149	128	0.4433	0.6576
Observations in the full sample			236	406		

Note: Results obtained using the ‘rddensity package’ in Stata (Cattaneo et al. 2018). The first two tests have been obtained by specifying the bandwidth at both sides of the cutoff (6% is the lower bound and 8% is the upper bound of the computed optimal bandwidth range) to construct the density estimators on the two sides of the cutoff. The third one has been obtained without specifying the bandwidth.

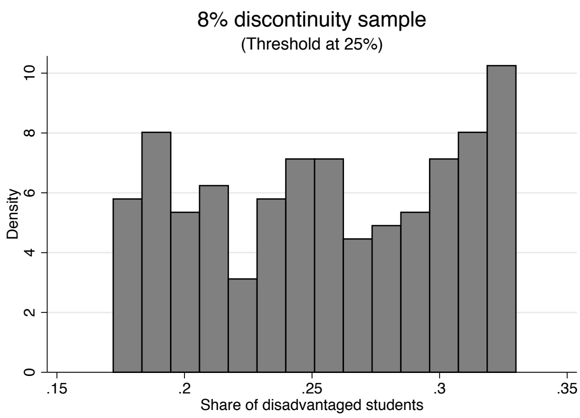

Figure 3. Frequency distribution of the schools with respect to the share of disadvantaged students for the 6% discontinuity sample.

Appendix C: Choice of *m* - Figures

Depending on the choice of the partial sample size, *m*, the share of super-efficient observations varies: the size of the drawn sample (*m*) with respect to the total sample size *n* influences the probability of the observation under analysis not to belong to the production frontier. The value of *m* is set to attain a sufficiently small decrease in the share of super-efficient schools for different control/treated/overall groups and for different bandwidths (here, *m*=40).

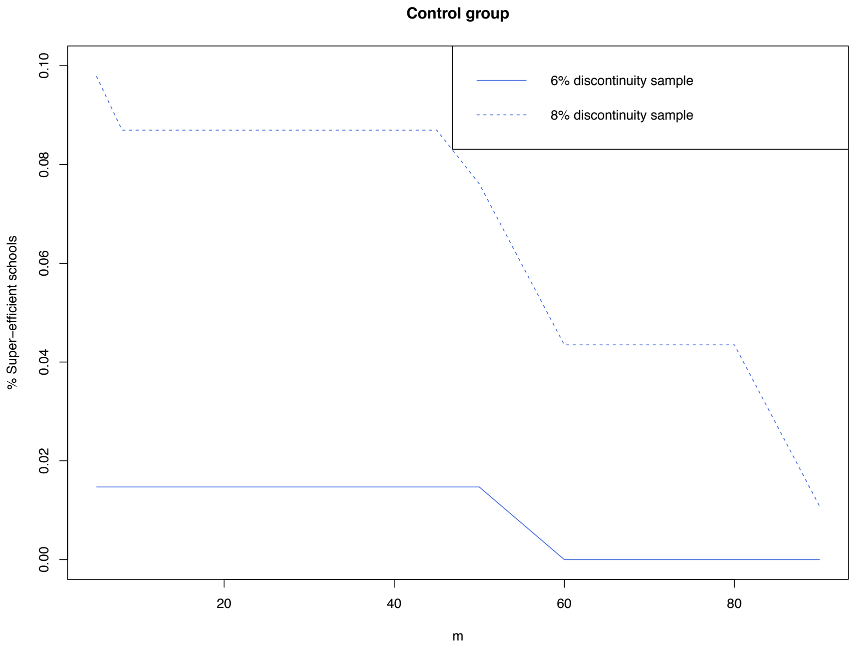

Figure 4. Marginal decrease in percentage of super-efficient schools. Control group.

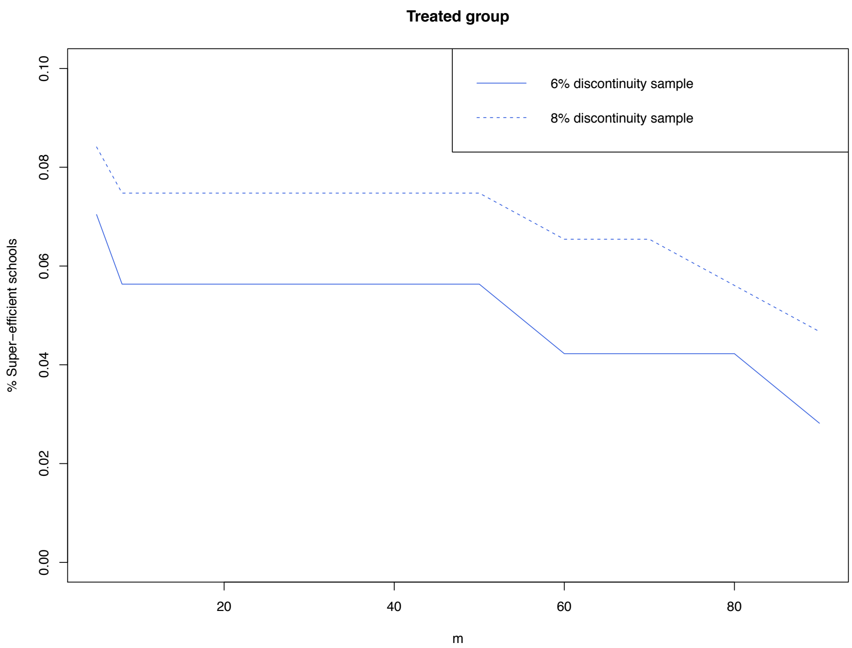

Figure 5. Marginal decrease in percentage of super-efficient schools. Treated group.

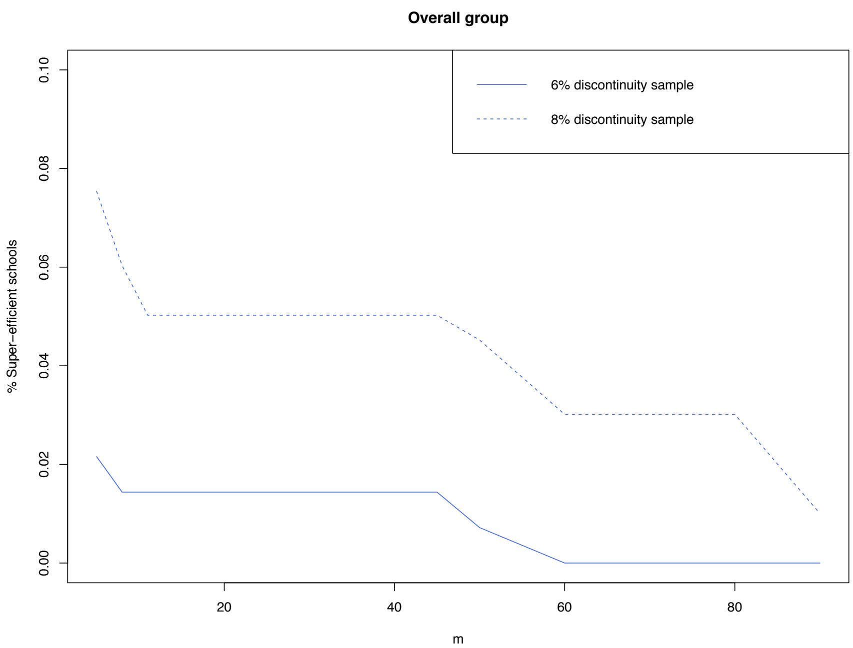

Figure 6. Marginal decrease in percentage of super-efficient schools. Overall group.

Appendix D: Complete descriptive statistics of the efficiency estimates

In this section, we present the results of the main efficiency analysis. We consider two inputs (*Teaching hours per student*, *Operating grants per student*), four outputs (*Share of students with that can* progress to the next school year without any restrictions, *Share of students without problems of absenteeism*, *Share of students progressing though school*, *Share of students enrolled in higher education*), three groups of contextual variables (School, Teacher and Student characteristics) and *m* is set to 40.

In the 6% discontinuity sample there are 68 schools below the threshold and 71 above.
In the 8% discontinuity sample there are 92 schools below the threshold and 107 above.

D.1 Descriptive statistics of the efficiency scores for 8% discontinuity sample.

Table 10. Descriptive statistics of the efficiency scores. 8% discontinuity sample

	Below threshold				*Above threshold*				*p-value*
	mean	sd	min	max	mean	sd	min	max	
Unconditional									
Overall efficiency	0.8611	0.0887	0.6431	1.0000	0.7826	0.1061	0.4966	1.1201	0.0000
School efficiency	0.8598	0.0899	0.6345	1.0000	0.8553	0.1219	0.5194	1.0423	0.7068
Program efficiency	1.0017	0.0027	1.0000	1.0136	0.9179	0.0537	0.7169	1.0746	0.0000
Conditional 1 - School characteristics									
Overall efficiency	0.9138	0.0726	0.7245	1.0000	0.8596	0.0941	0.6245	1.0000	0.0001
School efficiency	0.9133	0.0735	0.6928	1.0000	0.9304	0.0820	0.6655	1.0000	0.0393
Program efficiency	1.0008	0.0172	0.9540	1.0540	0.9246	0.0666	0.6838	1.0033	0.0000
Conditional 2 - Teacher characteristics									
Overall efficiency	0.9654	0.0611	0.7382	1.0000	0.9092	0.0898	0.5773	1.0000	0.0000
School efficiency	0.9599	0.0662	0.6956	1.0000	0.9537	0.0742	0.5711	1.0000	0.0886
Program efficiency	1.0077	0.0561	0.8004	1.2067	0.9543	0.0688	0.7024	1.1554	0.0000
Conditional 3 – Student characteristics									
Overall efficiency	0.9177	0.0806	0.7241	1.0000	0.9238	0.0865	0.4919	1.0000	0.5527
School efficiency	0.9075	0.0823	0.7248	1.0000	0.9538	0.0721	0.5348	1.0000	0.0000
Program efficiency	1.0116	0.0233	0.9761	1.1212	0.9687	0.0600	0.7889	1.2155	0.0000
Conditional 4 – School & Teacher & Student characteristics									
Overall efficiency	0.9578	0.0564	0.7936	1.0000	0.9597	0.0619	0.5536	1.0000	0.9132
School efficiency	0.9537	0.0599	0.7811	1.0000	0.9767	0.0438	0.7400	1.0000	0.0048
Program efficiency	1.0046	0.0131	0.9736	1.0546	0.9825	0.0481	0.7481	1.1248	0.0000
Observations	92				107				

p-values obtained from the non-parametric Wilcoxon–Mann–Whitney test to examine whether the control and the treated groups are from populations with the same distribution.
The conditional models include the following variables:
Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students
Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers
Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students
Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority, Teacher diploma, % students with problems in primary school, % students with special needs in primary school

D.2 Results on statistical inference for 8% bandwidth.

Table D.1. Direction of the influence of the contextual variables. 8% discontinuity sample.

	Conditional 1			Conditional 2			Conditional 3		Conditional 4		
	Influence	p-value		Influence	p-value		Influence	p-value	Influence	p-value	
School characteristics											
General education	Favorable	0.000	***						Favorable	0.005	***
School size	Favorable	0.174							Favorable	0.0105	**
% Change school	Favorable	0.003	***						Favorable	0	***
Previously treated	Unfavorable	0.0015	***						Unfavorable	0.0155	**
Private education	Favorable	0.165									
Special needs school	Favorable	0.2535									
Teacher characteristics											
Teacher seniority				Favorable	0.017	**			Favorable	0.1135	
Teacher diploma				Favorable	0.6905				Unfavorable	0.2345	
Teacher age				Unfavorable	0.010	**					
School principal seniority				Favorable	0.263						
Teacher contract				Unfavorable	0.000	***					
% female teachers				Favorable	0.000	***					
Student characteristics											
Primary retention							Unfavorable	0.9715	Unfavorable	0.139	
Special students in primary							Unfavorable	0.578	Unfavorable	0.1275	
% Man							Unfavorable	0.1175			

* p < 0.10, ** p < 0.05, *** p < 0.01

The conditional models include the following variables:

Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students

Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers

Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students

Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority & diploma, % students with problems in primary school, % students with special needs in primary school

Appendix E: Robustness check results

In this section we present the analysis performed to test the robustness of the main results excluding the eligible but not treated schools, so to address some left potential criticisms and we make sure that these elements do not drive the findings.

Specifically, we present the results of the efficiency analysis where we exclude from the sample the schools that are eligible but not treated because unable to generate a minimum of six teaching hours (for further explanation, see also Appendix A). Nevertheless, we consider the same optimal bandwidth range, between 6% and 8%.

For the following estimation, we consider two inputs (*Teaching hours per student*, *Operating grants per student*), four outputs (*Share of students progressing through school without any restrictions*, *Share of students without problems of absenteeism*, *Share of students without grade retention*, *Share of students enrolled in higher education*), three groups of contextual variables (School, Teacher and Student characteristics) and *m* is set to 40.

In the 6% discontinuity sample there are 68 schools below the threshold and 48 above.

In the 8% discontinuity sample there are 92 schools below the threshold and 89 above.

E.1 Descriptive statistics of the efficiency scores for 6% discontinuity sample.

Table 11. Descriptive statistics of the efficiency scores. 6% discontinuity sample

	Below threshold				*Above threshold*				*p-value*
	mean	sd	min	max	mean	sd	min	max	
Unconditional									
Overall efficiency	0.8549	0.0840	0.6491	1.0000	0.7877	0.0906	0.6115	1.0000	0.0004
School efficiency	0.8538	0.0848	0.6434	1.0000	0.8740	0.1089	0.6415	1.0002	0.1266
Program efficiency	1.0015	0.0021	1.0000	1.0088	0.9045	0.0576	0.7272	1.0000	0.0000
Conditional 1 - School characteristics									
Overall efficiency	0.9183	0.0724	0.7106	1.0000	0.8697	0.0874	0.6456	1.0000	0.0027
School efficiency	0.9079	0.0736	0.6941	1.0000	0.9640	0.0550	0.7707	1.0000	0.0000
Program efficiency	1.0118	0.0197	0.9838	1.0989	0.9025	0.0772	0.6716	1.0000	0.0000
Conditional 2 - Teacher characteristics									
Overall efficiency	0.9536	0.0647	0.7373	1.0000	0.8946	0.0932	0.6667	1.0000	0.0001
School efficiency	0.9463	0.0693	0.7190	1.0000	0.9703	0.0478	0.7677	1.0000	0.0576
Program efficiency	1.0082	0.0168	0.9704	1.0740	0.9216	0.0811	0.7212	1.0146	0.0000
Conditional 3 – Student characteristics									
Overall efficiency	0.9230	0.0784	0.7353	1.0000	0.9312	0.0778	0.7196	1.0000	0.5698
School efficiency	0.9278	0.0773	0.7450	1.0000	0.9649	0.0564	0.7734	1.0000	0.0038
Program efficiency	0.9953	0.0353	0.8501	1.1223	0.9648	0.0521	0.8335	1.0410	0.0025
Conditional 4 – School & Teacher & Student characteristics									
Overall efficiency	0.9657	0.0481	0.8550	1.0000	0.9643	0.0455	0.8184	1.0000	0.5413
School efficiency	0.9682	0.0469	0.8532	1.0000	0.9838	0.0354	0.8381	1.0000	0.0250
Program efficiency	0.9975	0.0167	0.9081	1.0349	0.9809	0.0451	0.8279	1.0955	0.0054
Observations	68				48				

p-values obtained from the non-parametric Wilcoxon–Mann–Whitney test to examine whether the control and the treated groups are from populations with the same distribution.
The conditional models include the following variables:
Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students
Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers
Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students
Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority, Teacher diploma, % students with problems in primary school, % students with special needs in primary school

E.2 Descriptive statistics of the efficiency scores for 8% discontinuity sample.

Table 12. Descriptive statistics of the efficiency scores. 8% discontinuity sample

	Below threshold				*Above threshold*				*p-value*
	mean	sd	min	max	mean	sd	min	max	
Unconditional									
Overall efficiency	0.8611	0.0887	0.6431	1.0000	0.7826	0.1061	0.4966	1.1201	0.0000
School efficiency	0.8598	0.0899	0.6345	1.0000	0.8553	0.1219	0.5194	1.0423	0.7747
Program efficiency	1.0017	0.0027	1.0000	1.0136	0.9179	0.0537	0.7169	1.0746	0.0000
Conditional 1 - School characteristics									
Overall efficiency	0.9138	0.0726	0.7245	1.0000	0.8596	0.0941	0.6245	1.0000	0.0000
School efficiency	0.9133	0.0735	0.6928	1.0000	0.9304	0.0820	0.6655	1.0000	0.0083
Program efficiency	1.0008	0.0172	0.9540	1.0540	0.9246	0.0666	0.6838	1.0033	0.0000
Conditional 2 - Teacher characteristics									
Overall efficiency	0.9654	0.0611	0.7382	1.0000	0.9092	0.0898	0.5773	1.0000	0.0000
School efficiency	0.9599	0.0662	0.6956	1.0000	0.9537	0.0742	0.5711	1.0000	0.1348
Program efficiency	1.0077	0.0561	0.8004	1.2067	0.9543	0.0688	0.7024	1.1554	0.0000
Conditional 3 – Student characteristics									
Overall efficiency	0.9177	0.0806	0.7241	1.0000	0.9238	0.0865	0.4919	1.0000	0.5963
School efficiency	0.9075	0.0823	0.7248	1.0000	0.9538	0.0721	0.5348	1.0000	0.0000
Program efficiency	1.0116	0.0233	0.9761	1.1212	0.9687	0.0600	0.7889	1.2155	0.0000
Conditional 4 – School & Teacher & Student characteristics									
Overall efficiency	0.9578	0.0564	0.7936	1.0000	0.9597	0.0619	0.5536	1.0000	0.9075
School efficiency	0.9537	0.0599	0.7811	1.0000	0.9767	0.0438	0.7400	1.0000	0.0077
Program efficiency	1.0046	0.0131	0.9736	1.0546	0.9825	0.0481	0.7481	1.1248	0.0000
Observations	92				89				

p-values obtained from the non-parametric Wilcoxon–Mann–Whitney test to examine whether the control and the treated groups are from populations with the same distribution.
The conditional models include the following variables:
Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students
Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers
Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students
Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority, Teacher diploma, % students with problems in primary school, % students with special needs in primary school

Appendix F: Results for the full sample.

For completeness, in this section we present the analysis performed for the full sample, with the caveat that here endogeneity issues have not been dealt with and accordingly we cannot give causal interpretation of the findings, differently from what we intend to do in the present paper by applying the proposed approach.

First, in Tables 13 and 14 we show the baseline characteristics when all the observations of the sample are under analysis, so including also the ones far from the threshold. We can observe that if we consider the full sample, then the two groups (below and above the threshold) turn out to be systematically different for almost every variable. Specifically, on average treated schools have a bigger size, a higher share of students that change school in the next year, a higher share of students with special needs, less teachers with a proper diploma for the subject they are responsible for and there are less privately managed schools. If we overlook these remarkable differences and we do not deal with this heterogeneity in the performance assessment, we might estimate a confounded educational production process and end up with biased efficiency estimates, preventing from going beyond the correlation of the findings.

Then, in Table 15 we provide the efficiency results for the full sample. We recall that we consider two inputs (*Teaching hours per student*, *Operating grants per student*), four outputs (*Share of students progressing through school without any restrictions*, *Share of students without problems of absenteeism*, *Share of students without grade retention*, *Share of students enrolled in higher education*), three groups of contextual variables (School, Teacher and Student characteristics) and m is set to 40. There are 236 schools below the threshold and 406 above. Interestingly, results are similar with those obtained from the discontinuity samples, with few exceptions though. Program efficiency scores of the treated schools are close to 1 whenever student-related characteristics are included in the frontier estimation. Exploring the distribution of the efficiency scores, we observe that the further the schools from the cutoff, the higher the program efficiency scores of the treated schools whenever students' characteristics are taken into account. However, we cannot give causal interpretation to this evidence because of the explained confounding factors. Accordingly, we can conclude that including in the analysis the schools far from the cutoff set at 25%, these additional resources might play a role, either because the treatment intensity is higher or because resources are allocated in more problematic contexts. Nevertheless, the fact remains that we should handle these considerations with caution.

Table 13. Sample means for control/treated group and population. Input and output variables. Full sample.

	Below threshold Control		*Above threshold* Treated		Total		p-value
Inputs							
Teaching hours per student	1.897	(0.447)	2.768	(0.511)	2.448	(0.644)	0.0000
Operating grants per student	869.4	(79.21)	1080.6	(170.2)	1003.0	(176.1)	0.0000
Outputs							
Share of students progressing to next year without restrictions	67.85	(5.315)	58.20	(8.466)	61.75	(8.794)	0.0000
Share of students without problems of absenteeism	99.84	(0.485)	97.67	(3.696)	98.47	(3.133)	0.0000
Share of students without grade retention	95.83	(2.733)	92.63	(3.399)	93.81	(3.523)	0.0000
Share of students enrolled in higher education	84.13	(13.23)	45.21	(18.75)	59.52	(25.28)	0.0000
Observations (school level)	236		406		642		

Note: Results for full sample. Standard deviation in parentheses. p-values obtained from t-test to examine whether the control and the treated group variables are statistically different in means.

Table 14. Sample means for control/treated group and population. Control variables. Full sample.

	Below Control		*Above* Treated		Total		*p*-test
School track - General	0.911	(0.285)	0.342	(0.475)	0.551	(0.498)	**0.0000**
School size (log)	6.275	(0.475)	6.092	(0.503)	6.159	(0.501)	**0.0000**
Share of students changing school	0.0887	(0.0451)	0.0988	(0.0509)	0.0951	(0.0491)	**0.0124**
Previously treated school	0.0763	(0.266)	0.877	(0.329)	0.583	(0.494)	**0.0000**
Education provider							**0.0000**
Public education	0.097		0.298				
Municipal education	0.03		0.121				
Private education	0.873		0.581				
School with special need students	0.364	(0.482)	0.527	(0.500)	0.467	(0.499)	**0.0001**
Teacher seniority	3.869	(0.360)	3.823	(0.400)	3.840	(0.386)	0.1487
Teacher diploma	0.973	(0.0330)	0.946	(0.0440)	0.955	(0.0423)	**0.0000**
School principal seniority	5.422	(1.182)	5.509	(1.065)	5.477	(1.109)	0.3355
Teacher age	4.134	(0.331)	4.178	(0.312)	4.162	(0.320)	0.0951
Teacher full-time	0.288	(0.123)	0.304	(0.0981)	0.298	(0.108)	0.0716
Female teachers	0.596	(0.102)	0.559	(0.144)	0.573	(0.131)	**0.0006**
Share of students with grade retention in primary school	0.0555	(0.0534)	0.241	(0.0959)	0.173	(0.122)	**0.0000**
Share of special need students in primary school	0.00524	(0.0145)	0.0531	(0.0392)	0.0355	(0.0398)	**0.0000**
Share of male students	0.458	(0.126)	0.535	(0.265)	0.506	(0.227)	**0.0000**
Share of disadvantaged students	0.154	(0.0546)	0.431	(0.139)	0.329	(0.176)	**0.0000**
Observations (school level)	236		406		642		

Note: Results for full sample. Standard deviation in parentheses. *p*-values obtained from t-test to examine whether the control and the treated group variables are statistically different in means.

Table 15. Descriptive statistics of the efficiency scores. Full sample

	Below threshold				*Above threshold*				*p-value*
	mean	sd	min	max	mean	sd	min	max	
Unconditional									
Overall efficiency	0.8696	0.0842	0.6259	1.0014	0.7062	0.0953	0.4880	1.1425	0.0000
School efficiency	0.8692	0.0848	0.6186	1.0014	0.7902	0.1122	0.5195	1.1210	0.0000
Program efficiency	1.0006	0.0015	1.0000	1.0118	0.8960	0.0472	0.6565	1.0584	0.0000
Conditional 1 - School characteristics									
Overall efficiency	0.9159	0.0672	0.6773	1.0000	0.8551	0.0904	0.6342	1.0000	0.0000
School efficiency	0.9068	0.0692	0.6634	1.0000	0.8953	0.0911	0.6655	1.0000	0.3613
Program efficiency	1.0105	0.0227	0.9225	1.1780	0.9569	0.0608	0.6525	1.0837	0.0000
Conditional 2 - Teacher characteristics									
Overall efficiency	0.9520	0.0670	0.6495	1.0000	0.8438	0.1271	0.5075	1.0000	0.0000
School efficiency	0.9501	0.0680	0.6463	1.0000	0.9183	0.0957	0.5621	1.0000	0.0000
Program efficiency	1.0022	0.0149	0.9137	1.1065	0.9206	0.1159	0.6265	1.3487	0.0000
Conditional 3 – Student characteristics									
Overall efficiency	0.8992	0.0757	0.6708	1.0000	0.9258	0.0847	0.4911	1.0000	0.0000
School efficiency	0.9039	0.0747	0.7113	1.0001	0.9332	0.0787	0.5432	1.0000	0.0000
Program efficiency	**0.9949**	0.0228	0.8632	1.0514	**0.9940**	0.0701	0.7260	1.2332	0.0124
Conditional 4 – School & Teacher & Student characteristics									
Overall efficiency	0.9230	0.0646	0.7746	1.0000	0.9663	0.0563	0.5166	1.0000	0.0000
School efficiency	0.9333	0.0634	0.7789	1.0000	0.9601	0.0584	0.7102	1.0000	0.0000
Program efficiency	**0.9890**	0.0223	0.8622	1.0057	**1.0075**	0.0433	0.5585	1.1630	0.0000

Observations	236	406

p-values obtained from the non-parametric Wilcoxon–Mann–Whitney test to examine whether the control and the treated groups are from populations with the same distribution.

The conditional models include the following variables:

Conditional 1: School track (General education), School size, % of students changing school, Previously treated school, Private education, School with special need students

Conditional 2: Teacher seniority, Teacher diploma, School principal seniority, Teacher age, Teacher type of contract, % female teachers

Conditional 3: % students with problems in primary school, % students with special needs in primary school, % male students

Conditional 4: School track (General education), School size, % of students changing school, Previously treated school, Teacher seniority, Teacher diploma, % students with problems in primary school, % students with special needs in primary school