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] *KU Leuven, Faculty of Business and Economics (FEB), Leuven, Belgium*
[b] *KU Leuven, Work and Organisation Studies, Faculty of Business and Economics (FEB), Leuven, Belgium*
[c] *Maastricht University, UNU–MERIT, Maastricht, The Netherlands*

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 insights from econometric impact evaluation techniques and 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 by decomposing 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.

Keywords: Data envelopment analysis, Impact evaluation, Efficiency, Causal inference, Equal Educational Opportunities

☆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 4^{th} 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, EWEPA 2019 in London, EURO 2019 in London, XXVIII Meeting of AEDE in Las Palmas, Ana Camanho, Chris O'Donnell, Jonas Månsson, Maria C. A. Silva, Tommaso Agasisti, Jill Johnes, Geraint Johnes, Dániel Horn, Kristiaan Kerstens, Antonio Peyrache, Jaap Bos, Daniel Santín, Gabriela Sicilia, Fritz Schiltz, Vítezslav Titl, Steven Groenez, Melissa Tuytens, Ides Nicaise, Thomas Wouters, Jolien De Norre, Nele Havermans, the 'SONO Opvolgingsgroep' and three anonymous referees 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).

*Corresponding author.
Email addresses: giovanna.dinverno@kuleuven.be (Giovanna D'Inverno), mike.smet@kuleuven.be (Mike Smet), kristof.dewitte@kuleuven.be (Kristof De Witte)

1. Introduction

There has been increasing pressure for evidence-based interventions to channel budgetary resources in the most appropriate way towards well-defined priorities (OECD, 2017b). 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). This paper contributes to this emerging literature by providing a framework to overcome these endogeneity issues and by evaluating 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 can 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 specifications 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 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 (Dahl and Lochner, 2012) and labor market outcomes (Grenet, 2013; 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). 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 (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, 2018; 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

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

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 et al., 2016; Cordero et al., 2015; Simar et al., 2016). Second, it adds to the literature pertaining to the impact evaluation in efficiency by providing a causal interpretation of the findings. Third, it contributes to the literature bridging the gap between effectiveness and efficiency, by combining regression discontinuity together with conditional metafrontier approach in the efficiency framework. The proposed approach brings closer the idea of policy impact evaluation to the concept of efficiency, being a complementary tool for policy evaluation. For example, if it is true that an effective policy can be inefficient, an inefficient policy can be the reason why a policy might be not effective. 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, the empirical findings are 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, we explore the role of the environmental variables to unravel potential mechanisms.

2.1. Step 1. Tackling the endogeneity issue in frontier estimation: a Regression Discontinuity Design approach

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 causal 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:

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

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 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

[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.

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 (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.

2.2. Step 2. Decomposing the overall efficiency: a conditional metafrontier approach

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 $x = (x_1, \ldots, x_k) \in \mathbb{R}^{K+}$into a vector of outputs $y = (y_1, \ldots, y_l) \in \mathbb{R}^{L+}$ and that can be presented in the following standard formulation (Afriat, 1972):

$$y = f(x) \tag{2.2}$$

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

[3] For a more formal discussion on the axiomatic framework, we refer for example to Shepard (1970) and Kerstens et al. (2019), among others.

general production function implicitly neglects potential inefficiencies in the production process (Santín and Sicilia, 2017b). Therefore, we can add an efficiency component u:

$$y = f(x) \cdot u \tag{2.3}$$

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 the 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., 2019; 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 papers 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 Santín, 2018; 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 efficiency of each unit i belonging to one of the two groups by estimating a group-specific local production frontier

[4] For a comprehensive overview, we refer the interested reader to Kerstens et al. (2019).

(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 conditional version of the robust Free Disposal Hull (FDH) model also known as conditional *order-m* (Deprins et al., 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 specifications. 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). Finally, the conditional approach is well suited to mimic the RDD approach by including the assignment variable as an environmental variable in the model estimation, as well as the other covariates as we will discuss in the next step. [5]

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

[5] "Covariates", "environmental variables", "contextual variables" are used interchangeably throughout this paper. The first term is mostly used in the impact evaluation literature, while the other two in the efficiency one.

$$\widehat{\theta}^{s}_{m,n}(x, y \mid c) = \int_{0}^{\infty} (1 - \hat{F}_{X|Y,c,n}(ux \mid y, c))^{m} du \tag{2.5}$$

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, y the outputs and c the assignment variable. 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, as the evaluated observation is not necessarily part of the reference set. (Daraio and Simar, 2007a). A nonparametric kernel function and a bandwidth parameter b have to be selected using smoothing techniques to handle the assignment variable in the estimation.

2.3. Step 3. Including the environmental variables

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). [6] If the presence of the environmental factors is significant, the decomposition of the overall efficiency scores on step 2 loses its relevance. From this perspective, controlling for environmental variables becomes not only interesting but essential in the estimation of the production frontier.

With respect to the inclusion of covariates, the RDD literature is quite varied. 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 improve the precision and provide more accurate estimates (Lee and Lemieux, 2010; Calonico et al., 2019). Others suggest mainly the inclusion of imbalanced variables, when it is plausible to assume that all the relevant characteristics are observed in the data (Frölich and Huber, 2019). The direct inclusion of the environmental variables handles left heterogeneity across the treated and the control samples and leads to consistent estimation.

Especially in small-sample empirical applications, it is not advised to include all the observed covariates not to lose statistical power in the conditional estimation. For this reason, step 3 can be seen as the necessary further step to undertake in presence of imbalanced variables, if found any after checking their statistical difference between treated and control group as suggested in step 1. We consider the complete model as a robustness check when enough data are available for the estimation procedure.

[6] In the efficiency literature alternative interpretations of the "environmental variables" can be found. For example, Bartelsman and Doms (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.

Using a conditional efficiency framework, the efficiency estimates are not only determined by the inputs (x), the outputs (y) and the assignment variable (c), but also by the other environmental variables (z) under a non-separable production context (Cazals et al., 2016). Adapting the notation, the input-oriented conditional order-m efficiency estimator ($\widehat{\theta}^{s}_{m,n}$) is defined as follows:

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

For this estimation, a nonparametric kernel function and a bandwidth parameter b have to be selected using smoothing techniques, properly handling discrete and/or continuous environmental variables.

To conclude, an additional source of information can be obtained while performing the conditional analysis. By comparing the conditional and the unconditional (namely without environmental variables) efficiency estimates

$$Q^{s,c,z}_{m} = \widehat{\theta}^{s}_{m,n}(x, y \mid c, z) / \widehat{\theta}^{s}_{m,n}(x, y) \tag{2.7}$$

we can causally evaluate the direction of the environmental variables influence together with the assignment variable role 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.

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).[7] 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, 2018; 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

[7] 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); Johnes et al. (2017a,b).

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.

3.1. The 'Equal Educational Opportunities' program

The Flemish education system is organised into three educational networks, i.e. official education organised by the Flemish community, government-aided public education run by municipal or provincial authorities, and government-aided private education organised by a private person or organisation, consisting primarily of catholic schools. The majority of Flemish schools are government-aided private schools. Despite all networks receive similar government funding and are free of tuition, private schools attract, on average, students with a higher socioeconomic status. Further, Flemish secondary education is organized in a tracking system. Students can choose between programmes in an academic, technical, artistic, or vocational education track.

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, 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).

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 support secondary schools with a significant number of disadvantaged students. A positive impact of the program would consist in the re-

duction of the gap on the academic results (that is, the outputs of the educational process) between the schools that do and do not receive the additional inputs. The program could be perfectly effective without requiring these schools to improve their level of efficiency (that is, in the input-output relationship), but this is not what the legislator expected when designing it. For example, Article VI.5 of the GOK Decree (2002) states that "for the use of resources, schools must develop a school-specific equal opportunities policy. Schools are therefore no longer approached as executors of a policy that is set out for them, but are expected to autonomously develop their own GOK policy within the flexible frameworks and instruments that the government provides" (Poesen-Vandeputte and Nicaise, 2012). 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.[8]

3.2. Data and variables

We observe 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

[8] 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.

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.

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); in the Flemish law teaching hours are linearly determined by the number of students and depend on the study field (De Witte et al., 2019). 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; in the Flemish law also operating grants are linearly determined by the number of students and depend on the study field (De Witte et al., 2019). 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

Since the initial conceptualization of the educational production function, there has been perceived the need to measure the school performance beyond student achievements (e.g. test scores), accounting for the school's ability to provide students with tools to succeed in their later-stage challenges (Levin, 1974; Hanushek, 1979). Following this rationale, the mission of secondary schools covers different objectives and involves different temporal horizons, namely to succeed in promoting students' short-term educational outcomes and long-term lifelong learning opportunities (Silva et al., 2019). Accordingly, both dimensions need to be considered so to account for these complementary objectives and to assess the efficiency of the conversion of resources in these educational results.

For the purpose of analysis, a comprehensive definition of output has been considered to represent all these aspects, including intermediate outputs (throughputs), short-term and long-term outputs (outcomes), looking at the most suitable ones for the Flemish context and in line with the standard literature on efficiency in education and education economics.

The first output is *share 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 final school exams. 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 either of 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. This variable can be seen in the same way as student test scores, commonly used as output in the literature (see for all De Witte and López-Torres, 2017). The second output mentions the *share of students without grade retention* , defined as the complement to the proportion of students experiencing grade retention in secondary school. [9] Grade retention has been considered an important dimension to be looked at in the education economics literature (see for example Rosenfeld, 2010) as well as passing rates in the efficiency in education literature as a measure of educational quality (see for example Grosskopf et al., 2014). It should be noted that 24% of the 15-years old in Flanders experienced grade retention, which is double from the OECD average. The third output variable consists of the *share 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.[10] This variable is not that common as such in the efficiency in education literature, but it is rather assimilated to the use of attendance rate in previous studies (see for example Bradley et al., 2001; Daneshvary and Clauretie, 2001; Grosskopf and Moutray, 2001). In these studies this variable has been considered an output on a par with student test scores. However, it could be interpreted in a different way and whether to be considered as a throughput or output depends also on the other included variables and their timing. To this extent, the *share of students without problems of absenteeism* can be seen as throughput with respect to grade retention and success in the final school exams.[11] Finally, based on the arguments made above, the *share of students enrolled in*

[9] Following Jones and Waguespack (2011), grade retention is "the practice in which children are required to repeat a grade level in school because they failed to meet required benchmarks or grade level standards".

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

[11] 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 embeds partly both the aspects in a complementary fashion. The rather low correlation coefficients (0.6359, 0.3932, 0.3784) and

higher education is considered to account for a longer-term result (see for example Silva et al., 2019). This variable 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.

As partially different timing in these outputs might rise some doubts, we provide a series of robustness checks where we test for different combinations of output specifications.

3.2.3 Contextual variables

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".

some robustness checks including separately the outputs further prove this statement.

Fifth, *education provider* refers to the educational networks that act as "umbrella organization" for the school governing bodies: *public education* organized by the central government, *public municipal 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; De Witte and Rogge, 2011) and, thus, has to be taken into account when checking the characteristics across schools.

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 *share 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 *share of special need students in primary school* posits as a representative for pupil's cognitive skill the school has to deal with. Third, the *share 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).

To conclude, we recall the assignment variable "share of disadvantaged students" used to determine the treatment status.

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). [12] 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 the bandwidth equal to 6%, we restrict the full sample by considering only the schools whose share of disadvantaged students is between $(25\% - 6\%)$ and $(25\% + 6\%)$. Specifically, the schools between $(25\% - 6\%)$ and 25% constitutes the control group, while the schools between 25% and $(25\% + 6\%)$ the treated group. In each group respectively 68 and 71 schools are identified.

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. This is unlikely due to the use of administrative data to crosscheck multiple indicators used in determining the percentage of disadvantaged students. Moreover, the program is fully exogenous for parents when they do their school choice, because the elements parents should take decisions upon are neither observed nor publicly disclosed (Palmaccio et al., 2020). Parents do not (and cannot) observe either the funding level or the provided teaching hours, as this information is contained in the administrative data. Likewise, parents do not observe whether a school has been previously treated or not, as this is not a publicly disclosed information, but only traceable in the administrative data. To complement these arguments, we check in the data 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.[13] Table 1 suggests that the two groups are not statistically different in means for most of the control variables considered. However, a small number of control variables is statistically different in means. We include these dissimilar baseline variables as environmental variables in step 3. These environmental variables are mostly related to student characteristics such as the share of students with grade retention in primary school and the share of special needs students in primary school, which will serve as contextual variables in the current analysis. It has also

[12] Recently alternative bandwidth selection procedures have been proposed (e.g. Calonico et al., 2017). Our results are very similar across different procedures.

[13] Again, for brevity, in this section we report the means for the 6% discontinuity sample. In Appendix B.2, there is the table listing the means for the 8% discontinuity sample.

been observed that the schools below the threshold tend to focus more on general education schools and they have not received additional funding in the previous cycle. 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.[14] On the other hand, this may be 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. [15]

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 (null hypothesis of no manipulation). 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.

Furthermore, 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 (local) average treatment effects, consistently with the sharp Regression Discontinuity Designs and with the idea of "local compliers", being the units close to the threshold (Lee and Lemieux, 2010;

[14] 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.

[15] When using a multiplier model specification for the efficiency analysis, the weights might offer interesting and complementary insights about the different weighting on inputs and outputs across the treated and the control group.

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
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.

Frölich and Huber, 2019). We consider dealing with imperfect compliance as scope for future research.

Table 3: 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).

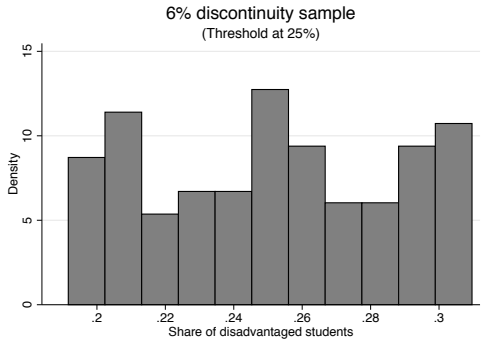

Figure 1:

Distribution of the schools with respect to the share of disadvantaged students

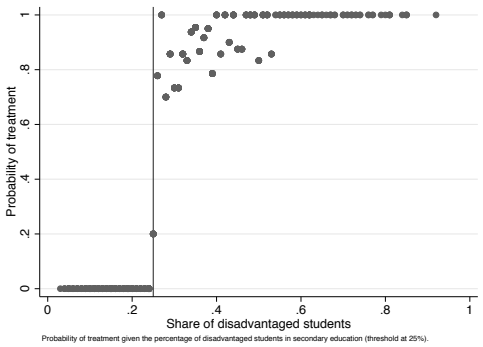

Figure 2:

Discontinuity in the probability of treatment

4.2. Step 2: a conditional metafrontier approach

In step 2, for the groups of schools distinguished in step 1, we estimate the educational production frontier using a conditional input-oriented robust FDH model. We compute the efficiency scores for each school under analysis following equation (2.5), where inputs and outputs are considered together with the assignment variable, namely the share of disadvantaged students. 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 presented in Appendix D), without accounting for relevant contextual characteristics (imbalanced variables have 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 1.2 percentage points higher for treated schools and also the average school-specific efficiency is slightly higher for this group of schools. The overall inefficiency level among the treated schools is almost 7.5% (obtained as 1-0.9253) versus 8.7% (obtained as 1-0.9131) among the control group. However, this difference is not remarkable and the treated group presents a bigger variability in the efficiency scores, denoted by a lower minimum value.

To explore the role of the policy, we look at the program efficiency. A program efficiency score for the control schools lower than 1 denotes that the control-specific frontier is further from the overall frontier compared to the treated-specific frontier. The average program efficiency of the treated schools amounts to 1.0027, suggesting that the treated schools are mainly constituting the metafrontier.[16] This puts forth the notion that treated schools might have successfully convert more resources into more outputs around the threshold. In the ideal RDD setting where all the plausible relevant contextual variables are balanced, these estimates would represent the local average treatment effect. As found in step 1, this is not the case for few of them. Accordingly, we caution the reader that to provide causal inference we have to resort on step 3.

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	
Conditional model without covariates									
Overall efficiency	0.9131	0.0858	0.7094	1.0000	0.9253	0.0944	0.5874	1.0000	0.1916
School efficiency	0.9207	0.0861	0.7128	1.0000	0.9233	0.0972	0.5823	1.0000	0.4654
Program efficiency	0.9919	0.0181	0.8871	1.0000	1.0027	0.0153	0.9063	1.0658	0.0000
Observations	68				71				

Note: Results for **6%-discontinuity sample** (8%-discontinuity sample in Appendix D). 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. **Outputs**: i) Share of students without problems of absenteeism, ii) Share of students without grade retention, iii) Share of students progressing to next school year without restrictions, iv) Share of students enrolled in higher education

4.3. Step 3: a conditional metafrontier approach including covariates

Following the insights from the RDD literature and the evidence from Table 1, we include the relevant covariates (the imbalanced variables) in the model specification, together with the assign-

[16] Recall that efficiency scores > 1 point to 'super-efficient' observations, which is due to the resampling technique discussed in Section 2. A score of 1.0027 can be interpreted as the schools are performing 0.3% better than expected.

ment variable. Table 5 shows that the addition of imbalanced variables in the frontier estimation does play a role with respect to the findings incurred in step 2. The mean of the overall and school efficiency is higher for the treated schools, pointing to the fact that at least some of them manage to make use of the expanded possibility production set. However, the minimum is lower for the treated schools, denoting that there is an unexploited production capacity among some of them. When including the imbalanced variables, the average difference in program efficiency between the control and treated groups almost vanishes, although the variation in the program efficiency scores is larger in the treated schools. This points to the fact that not all the schools successfully managed to convert more resources into more output (and this can explain also why the policy has been found ineffective in a previous study - see De Witte et al., 2017). The absence of significant differences between the treated and the control schools suggests that the additional resources have not been organised in a different way and have not stimulated a better management of the resources across all the treated schools. There are some schools that do expand their production possibility set and reorganise their managerial practises, however on average around the threshold this impact is not that remarkable and is not statistically significant. In other words, we look at the control schools as a reference to observe what would have had happened if the treated schools had not received additional resources. No difference in the program efficiency between treated and control units indicates that in the treated sample there are units that act as if no more resources were given, hence missing the potential opportunity to work differently. 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 no significant differences for the program efficiency.

Table 5: Descriptive statistics of the efficiency scores.

	Below threshold				*Above threshold*				p-value
	mean	sd	min	max	mean	sd	min	max	
Conditional model with relevant covariates									
Overall efficiency	0.9147	0.0801	0.7321	1.0000	0.9449	0.0786	0.5752	1.0000	0.0466
School efficiency	0.9260	0.0759	0.7317	1.0000	0.9601	0.0704	0.6063	1.0000	0.0013
Program efficiency	0.9878	0.0275	0.8984	1.0347	0.9849	0.0548	0.7781	1.2014	0.2681
Observations	68				71				

Note: Results for **6%-discontinuity sample** (8%-discontinuity sample in Appendix D). 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.
Outputs: i) Share of students without problems of absenteeism, ii) Share of students without grade retention, iii) Share of students progressing to next school year without restrictions, iv) Share of students enrolled in higher education
The conditional model with relevant covariates includes the following variables imbalanced between the treated and control group **at 5% statistical level**: School track (General education), Previously treated school, % 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. 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. On the contrary, the further we move away from the threshold (within the optimal bandwidth), the higher is the potential to gather resources to implement anything that can have an impact on efficiency. As a matter of fact, exploring the results for a larger discontinuity sample, we found on average a higher program efficiency for the treated group and the difference with respect to the control group is statistically significant (we provide the analysis for the 8% discontinuity sample -upper bound- in Appendix D).[17]

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 included in the estimation, 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 curse of dimensionality, only the imbalanced variables have been included in line with step 3. The other observed characteristics might be included if more data were available for the estimation procedure.

Table 6 summarizes the main findings obtained for the included contextual variables, listing their median influence 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. 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. As revealed from the nonparametric regression plot, being a school which had received additional resources in the previous three-year cycle has an unfavourable influence, signifying the aspects of a lack of learning effect in management of these extra resources. 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. Following the same reasoning, the share of disadvantaged students plays an unfavourable role on the performance assessment.

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

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

Table 6: Influence direction of the variables.

	Influence	p-value	
School characteristics			
General education	Favorable	0.1945	
Previously treated	Unfavorable	0.146	
Student characteristics			
Share of students with grade retention in primary school	Unfavorable	0.0000	***
Share of students with special needs in primary school	Unfavorable	0.0985	*
Assignment variable			
Share of disadvantaged students	Unfavorable	0.0000	***

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

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

In this model specification only the variables imbalanced between the treated and control group at 5% statistical level have been included, together with the assignment variable.

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, the difference of the program efficiency between the treated and the control group is not statistically significant. 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.

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 suggest how to account for the heterogeneity in the estimation of the production frontier of step 2 and how to include the environmental variables. 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, causal 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*), together with the assignment variable, namely the share of disadvantaged students. Whereas, a number of contextual variables were chosen among schools, teachers and students characteristics.

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 do not display a remarkable difference in the program efficiency compared to their counterfactual control schools. These results seem to be very robust to several specifications (e.g. by different output combinations and by education track). By design, the treated group is made by "local compliers" (see for example Frölich and Huber, 2019). As a consequence, the conclusions drawn from this empirical application can be considered (local) average treatment effect and they have a high internal validity, but not an external one (we refer to Wing and Bello-Gomez, 2018, for a recent methodological research overview to improve external validity). This piece of information is anyway interesting because it can suggest whether the set threshold and the intensity treatment is effective or not at least in its proximate neighbourhood.

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.

References

Abadie, A., Cattaneo, M.D., 2018. Econometric Methods for Program Evaluation. *Annual Review of Economics* 10 (1), 465–503.

Afriat, S.N., 1972. Efficiency Estimation of Production Functions. *International Economic Review* 13 (3), 568–598.

Amsler, C., Prokhorov, A., Schmidt, P., 2016. Endogeneity in stochastic frontier models. *Journal of Econometrics* 190 (2), 280–288.

Angrist, J.D., Lavy, V., 1999. Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics* 114 (2), 533–575.

Angrist, J.D., Pischke, J.S., 2009. Mostly harmless econometrics: An empiricist's companion. Princeton university press.

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. *European Journal of Operational Research* 256 (2), 533–542.

Aparicio, J., Santín, D., 2018. A note on measuring group performance over time with pseudo-panels. *European Journal of Operational Research* 267 (1), 227–235.

Bădin, L., Daraio, C., Simar, L., 2012. How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research* 223 (3), 818–833.

Bartelsman, E.J., Doms, M., 2000. Understanding productivity: Lessons from longitudinal microdata. *Journal of Economic literature* 38 (3), 569–594.

Battese, G.E., Rao, D.S.P., 2002. Technology Gap , Efficiency , and a Stochastic Metafrontier Function. *International Journal of Business and Economics* 1 (2), 87–93.

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

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

Calonico, S., Cattaneo, M.D., Farrell, M.H., Titiunik, R., 2017. rdrobust: Software for regression-discontinuity designs. *The Stata Journal* 17 (2), 372–404.

Calonico, S., Cattaneo, M.D., Farrell, M.H., Titiunik, R., 2019. Regression discontinuity designs using covariates. *Review of Economics and Statistics* 101 (3), 442–451.

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

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

Camanho, A.S., Dyson, R.G., 2006. Data envelopment analysis and Malmquist indices for measuring group performance. *Journal of Productivity Analysis* 26 (1), 35–49.

Cattaneo, M.D., Frandsen, B.R., Titiunik, R., 2015. Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference* 3 (1), 1–24.

Cattaneo, M.D., Jansson, M., Ma, X., 2018. rddensity: Manipulation Testing based on Density Discontinuity. *Stata Journal* 18 (1), 234–261.

Cazals, C., Fève, F., Florens, J.P., Simar, L., 2016. Nonparametric instrumental variables estimation for efficiency frontier. *Journal of econometrics* 190 (2), 349–359.

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

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

Cherchye, L., De Witte, K., Perelman, S., 2019. A unified productivity-performance approach applied to secondary schools. *Journal of the Operational Research Society* 70 (9), 1522–1537.

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

Cordero, J.M., Santín, D., Sicilia, G., 2015. Testing the accuracy of dea estimates under endogeneity through a monte carlo simulation. *European Journal of Operational Research* 244 (2), 511–518.

Dahl, G.B., Lochner, L., 2012. The Impact of Family Income on Child Achievement : Evidence from the Earned Income Tax Credit. *American Economic Review* 102 (5), 1927–1956.

Daneshvary, N., Clauretie, T.M., 2001. Efficiency and costs in education: Year-round versus traditional schedules. *Economics of Education Review* 20 (3), 279–287.

Daraio, C., Simar, L., 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis* 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 non-convex technologies: A unifying approach. *Journal of Productivity Analysis* 28 (1-2), 13–32.

De Witte, K., Hindriks, J., 2017. L'École de la Réussite. *Itinera Institue - Skribis* (September).

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. *Applied Economics* 45 (17), 2401–2412.

De Witte, K., López-Torres, L., 2017. Efficiency in education: A review of literature and a way forward. *Journal of the Operational Research Society* 68 (4), 339–363.

De Witte, K., Rogge, N., 2011. Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review* 30 (4), 641–653.

De Witte, K., Smet, M., Van Assche, R., 2017. The impact of additional funds for schools with disadvantaged students. *Steunpunt Onderwijsonderzoek, Gent*.

De Witte, K., Titl, V., Holz, O., Smet, M., 2019. Financing Quality Education for All : The Funding Methods of Compulsory and Special Needs Education. Leuven University Press, Leuven.

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

D'Inverno, G., Carosi, L., Ravagli, L., 2018. Global public spending efficiency in tuscan municipalities. *Socio-Economic Planning Sciences* 61 (March), 102–113.

Duflo, E., Dupas, P., Kremer, M., 2015. School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics* 123 (March), 92–110.

European Commission, 2017. Education and Training Monitor 2017 - Belgium.

Frölich, M., Huber, M., 2019. Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics* 37 (4), 736–748.

Gibbons, S., McNally, S., Viarengo, M., 2018. Does Additional Spending Help Urban Schools? An Evaluation Using Boundary Discontinuities. *Journal of the European Economic Association* 16 (5), 1618–1668.

Grenet, J., 2013. Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws. *Scandinavian Journal of Economics* 115 (1), 176–210.

Grosskopf, S., Hayes, K.J., Taylor, L.L., 2014. Efficiency in education: Research and implications. *Applied Economic Perspectives and Policy* 36 (2), 175–210.

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

Grosskopf, S., Valdmanis, V., 1987. Measuring hospital performance: A Non-parametric Approach. *Journal of Health Economics* 6 (2), 89–107.

Hanushek, E., Woessmann, L., 2015. Universal Basic Skills, What Countries Stand to Gain.

Hanushek, E.A., 1979. Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources* 14 (3), 351–388.

Hanushek, E.A., 2002. Publicly provided education. *Handbook of public economics* 4, 2045–2141.

Imbens, G., Kalyanaraman, K., 2012. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies* 79 (3), 933–959.

Jackson, C.K., Johnson, R., Persico, C., 2016. The Effects of School Spending on Educational and Economic Outcomes. *The Quarterly Journal of Economics* 131 (1), 157–218.

Johnes, G., Johnes, J., Agasisti, T., López-Torres, L., 2017a. Handbook of Contemporary Education Economics.

Johnes, J., 2015. Operational research in education. *European Journal of Operational Research* 243 (3), 683–696.

Johnes, J., Portela, M., Thanassoulis, E., 2017b. Efficiency in education. *Journal of the Operational Research Society* 68 (4), 331–338.

Johnson, A.L., Ruggiero, J., 2014. Nonparametric measurement of productivity and efficiency in education. *Annals of Operations Research* 221 (1), 197–210.

Jones, P., Waguespack, A., 2011. Grade Retention. Springer US, Boston, MA. pp. 708–709. URL: https://doi.org/10.1007/978-0-387-79061-9_1272, doi:10.1007/978-0-387-79061-9_1272.

Kerstens, K., O'Donnell, C., Van de Woestyne, I., 2019. Metatechnology frontier and convexity: A restatement. *European Journal of Operational Research* 275 (2), 780–792.

Kneip, A., Simar, L., Wilson, P.W., 2015. When bias kills the variance: Central Limit Theorems for DEA and FDH efficiency scores. *Econometric Theory* 31 (2), 394–422.

Kneip, A., Simar, L., Wilson, P.W., 2016. Testing Hypotheses in Nonparametric Models ofProduction. *Journal of Business and Economic Statistics* 34 (3), 435–456.

Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. *Journal of economic literature* 48 (2), 281–355.

Leithwood, K., Jantzi, D., 2009. A review of empirical evidence about school size effects: A policy perspective. *Review of educational research* 79 (1), 464–490.

Leuven, E., Lindahl, M., Oosterbeek, H., Webbink, D., 2007. The Effect of Extra Funding for Disadvantaged Pupils on Achievement. *The Review of Economics and Statistics* 89 (4) (November), 721–736.

Levin, H.M., 1974. Measuring efficiency in educational production. *Public Finance Quarterly* 2 (1), 3–24.

Li, Q., Racine, J.S., 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. *Journal of Transport Economics and Policy* 30 (1), 83–93.

Mayston, D.J., 2003. Measuring and managing educational performance. *Journal of the Operational Research Society* 54 (7), 679–691.

McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142 (2), 698–714.

Muralidharan, K., Sundararaman, V., 2015. The Aggregate Effect of School Choice: Evidence from a two-stage experiment in India Karthik. *The Quarterly Journal of Economics* 130 (3), 1011–1066.

Nusche, D., Miron, G., Santiago, P., Teese, R., 2015. OECD Reviews of School Resources: Flemish Community of Belgium 2015, in: OECD Reviews of School Resource. OECD Publishing, Paris.

O'Donnell, C.J., 2016. Using information about technologies, markets and firm behaviour to decompose a proper productivity index. *Journal of Econometrics* 190 (2), 328–340.

O'Donnell, C.J., Fallah-Fini, S., Triantis, K., 2017. Measuring and analysing productivity change in a metafrontier framework. *Journal of Productivity Analysis* 47 (2), 117–128.

O'Donnell, C.J., Rao, D.S., Battese, G.E., 2008. Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. *Empirical Economics* 34 (2), 231–255.

OECD, 2017a. Educational Opportunity for All: Overcoming Inequality throughout the Life Course.

OECD, 2017b. The Funding of School Education: Connecting Resources and Learning. OECD Publishing, Paris.

Olesen, O.B., Petersen, N.C., Podinovski, V.V., 2015. Efficiency analysis with ratio measures. *European Journal of Operational Research* 245 (2), 446–462.

Olesen, O.B., Petersen, N.C., Podinovski, V.V., 2017. Efficiency measures and computational approaches for data envelopment analysis models with ratio inputs and outputs. *European Journal of Operational Research* 261 (2), 640–655.

Palmaccio, S., Schiltz, F., De Witte, K., 2020. The effect of information shocks in primary schools. Evidence from inspectorate data. *Mimeo.*

Pischke, J.S., von Wachter, T., 2008. Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation. *Review of Economics and Statistics* 90 (3), 592–598.

Poesen-Vandeputte, M., Nicaise, I., 2012. Tien jaar gok-decreet, balans van het evaluatieonderzoek van het gelijke onderwijskansenbeleid in vlaanderen.

Rosenfeld, M.J., 2010. Nontraditional Families and Childhood Progress Through. *Demography* 47 (3), 755–775.

Santín, D., Sicilia, G., 2017a. Dealing with endogeneity in data envelopment analysis applications. *Expert Systems with Applications* 68 (February), 173–184.

Santín, D., Sicilia, G., 2017b. Impact evaluation and frontier methods in education: a step forward. *Handbook of Contemporary Education Economics*, 211–245.

Santín, D., Sicilia, G., 2018. Using DEA for measuring teachers' performance and the impact on students' outcomes: evidence for spain. *Journal of Productivity Analysis* 49 (1), 1–15.

Schlotter, M., Schwerdt, G., Woessmann, L., 2011. Econometric methods for causal evaluation of education policies and practices: A non-technical guide. *Education Economics* 19 (2), 109–137.

Shepard, R.W., 1970. Theory of Cost and Production Function. NJ: Princeton University Press, Princeton.

Silva, M.C., Camanho, A.S., Barbosa, F., 2019. Benchmarking of secondary schools based on students results in higher education. *Omega*, 102119.

Simar, L., Vanhems, A., Van Keilegom, I., 2016. Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *Journal of Econometrics* 190 (2), 360–373.

Stephens, M.J., Yang, D.Y., 2014. Compulsory Education and the Benefits of Schooling. *American Economic Review* 104 (6), 1777–1792.

Vaz, C.B., Camanho, A.S., 2012. Performance comparison of retailing stores using a Malmquist-type index. *Journal of the Operational Research Society* 63 (5), 631–645.

Wing, C., Bello-Gomez, R.A., 2018. Regression discontinuity and beyond: Options for studying external validity in an internally valid design. *American Journal of Evaluation* 39 (1), 91–108.