10. Impact evaluation and frontier methods in education: a step forward

*Daniel Santín and Gabriela Sicilia**

'Getting something wrong is not a crime.
Failing to learn from past mistakes because you are not monitoring and evaluating, is'.
Shapiro (2011, p. 5).

Monitoring and evaluation toolkit.
CIVICUS: World Alliance for Citizen Participation.

1 INTRODUCTION

Targets and tools for the monitoring and evaluation of educational policies and interventions in the economics of education have changed rapidly in the last 20 years. Most previous works in this field have focused on running multivariate analysis models to find statistical associations between variables, controlling for the presence of other covariates and factors that also influence the dependent variables. Behind this traditional approach lies the strong assumption that all covariates related to the dependent variable are exogenously determined. In other words, we would say that unobserved variables are equally distributed among the population to be analysed.[1] Given modern estimation technology it is no longer reasonable to make this assumption.

Parents' decisions regarding the education of their children (for example, choice of school, pre-primary education attendance, extra-curricular activities, support at home, choice of teacher and so on) are strongly related to the so-called unobserved heterogeneity, or simply endogeneity. This unobserved heterogeneity is mainly rooted in the difficulty in measuring certain dimensions – such as parents' motivation, expectations, incentives, non-cognitive traits, religious values and so on – which exert a big influence on both the aforementioned key academic decisions and on educational achievements, and this leads the researcher to confound the true causes of the observed results.

The same reasoning can be applied when we aim to evaluate the performance of teachers, principals or schools in many public education systems where they are not randomly distributed into schools. Highly qualified and more motivated teachers tend to self-select into better schools with higher academic results, better facilities and a better peer group.

* We are grateful to participants at the 4th Workshop on the Efficiency in Education held in Milano for useful comments and suggestions. We are also grateful to Tommaso Agasisti, Juan Aparicio, Geraint Johnes and Mika Kortelainen for helpful discussions.

[1] This 'ideal for analysis' education system would be equivalent to allocating students, teachers and principals to schools by holding a big lottery at the age of school entry. This unrealistic strategy is based on the idea that randomisation only creates small non-significant differences in unobserved variables. Of course, school choice in the real world is far away from this system.

Another example could be when better schools are able to select high-achieving students or those with the most motivated parents to support their children in the education production process. An unfair management comparison between schools will conclude that regardless of observed physical input quantities, the latter schools obtain better achievements due to better management. However, this result will likely be biased because of the non-observed factors.

Likewise, the endogeneity problem in the education sector can also have an effect in the opposite direction when there is a direct negative feedback from low educational achievement to resources. This applies, for example, when an educational intervention allocates more resources to schools with a greater proportion of disadvantaged students with poorer academic results (Levačić and Vignoles, 2002). In this case, a programme may incorrectly be diagnosed as ineffective in a standard multivariate analysis identifying a negative relationship between the intervention and the results.

Based on this background and the insights of statistics and econometrics, the impact of an evaluation in education literature has concluded that the best way to measure the true impact of educational interventions (the treatment) would be to observe the average performance of exactly the same population group, both with and without the analysed treatment. As this is impossible, the best solution to overcome this problem is to carry out randomised experiments by selecting an untreated or counterfactual population group with similar average characteristics to the treated group to compare both results after the intervention. Nevertheless, most public programmes are not yet designed to be evaluated using a counterfactual group. In these contexts, as we will discuss in Section 2, a set of so-called quasi-experimental evaluation techniques are usually used to look for appropriate counterfactual groups.

The impact evaluation of educational interventions can be carried out for programmes devoted to individuals (students, parents, families, teachers and so on) or to organisations (schools, districts, municipalities and so on). Although impact evaluation in terms of average output differences is the mainstream for evaluating educational programmes targeted at individuals (for example, scholarships for more disadvantaged students), when programmes are intended for organisations it also becomes relevant to measure educational efficiency and productivity differences by means of production frontiers (Worthington, 2001; Johnes, 2015; De Witte and López-Torres, 2015; Thanassoulis et al., 2016).

When public educational interventions are devoted to organisations they can be carried out to improve results through input-oriented interventions (raising school budgets, reducing teacher-pupil ratio, increasing teacher salary, facilities or instruction material, among others) but also by promoting the organisations' productivity, *via* technological change and/or improvements in technical efficiency. These practices are currently very common (external exams, schools' autonomy, teaching-learning practices in the classrooms or instruction time), so it becomes equally relevant to evaluate their potential impacts on the output to unravel the channels through which the intervention operates (inputs, technological change and/or efficiency). Surprisingly, to date in the economics of education, both fields of research, impact evaluation and production frontiers, run as parallel lines of research with no relationship, or very little, between them.

In recent years there has been an emerging and growing interest in the production

frontiers literature in addressing the endogeneity issue in the estimation of technical efficiency from a theoretical approach.[2] Furthermore, some works have started to relate production frontiers with impact evaluation insights into empirical educational problems. For example, Perelman and Santín (2011) address the endogeneity problem of school choice in Spain using instrumental variables, Crespo-Cebada et al. (2014) apply propensity score matching to compare performance across different school ownership arrangements, Santín and Sicilia (2014) exploit a natural experiment to evaluate teachers' performance in primary schools and Van Klaveren and De Witte (2014) relate conditional efficiency scores to the matching strategy.

This chapter is concerned with two main goals. First, Section 2 introduces the basics of impact evaluation in education for those readers not familiar with this approach. Secondly, in Section 3 we develop a theory to relate impact evaluation and production frontiers using the education production function framework. Section 4 describes a Monte Carlo simulation run to show how production frontiers can help to enhance the traditional impact evaluation approach regarding not only mean differences in outputs but also mean differences in productivity changes caused by technological and/or efficiency changes. Finally, Section 5 concludes and proposes the main lines and challenges for future research.

2 BASICS ON IMPACT EVALUATION

Behind any intervention to improve academic achievements lies a results chain in which the policymakers define the targets of the programme and the indicators to be used to measure whether it has been successful or not. The results chain contributes to clarifying all the steps necessary to reach the objectives and facilitate the evaluation. A typical results chain is made up of the following dimensions.

- *Inputs:* resources needed in the production function to achieve the outputs. They include teachers, other staff, school resources and budget.
- *Activities:* include instruction tasks carried out to transform inputs into outputs. For example, instruction time, homework, classroom organisation and so on.
- *Outputs:* results produced and delivered to the beneficiaries (students, teachers, principal and so on). Depending on the intervention and the final beneficiaries the outputs can be measured from test scores to the skills obtained in the trained dimensions.
- *Results:* short-to-medium term effects achieved by the beneficiaries.
- *Outcomes:* medium-to-long term goals of the intervention.

Once the results chain is defined and the indicators to measure the success of the programme are clear, the next stage, and the central challenge in carrying out effective impact

[2] Orme and Smith (1996); Bifulco and Bretschneider (2001, 2003); Ruggiero (2003, 2004); Cordero et al. (2015); Cazals et al. (2016); Mayston (2016), Simar et al. (2016) and Santín and Sicilia (2017) deal with endogeneity in the estimation of technical efficiency using non-parametric techniques; and Greene (2010); Mayston (2015); Amsler et al. (2016) and Griffiths and Hajargasht (2016) use parametric approaches.

evaluations, is to *identify* the *causal relationship* between the intervention and the outcomes (Gertler et al., 2016).[3] Impact evaluation is the technical approach that economists involved in education use to demonstrate causality.

Basically, an impact evaluation in education consists of a procedure to measure the causal effect of a programme or intervention (the treatment) on educational outputs, such as academic achievement, success rates, raise of non-cognitive skills and so on. To do this, the impact evaluation assess the average changes that can be *attributed* to this particular treatment in the well-being (effects and outcomes) of individuals or organisations receiving the treatment (the treated group) with respect to another group not receiving the programme (the counterfactual or control group).

Technically, the grounds for measuring causality are, in principle, quite simple. Let N be a population of individuals or schools that may receive a programme or treatment or not (D). We define, $D = 1$ if individuals received the treatment; $D = 0$ otherwise. After the treatment we observe the following outcomes: $E[Y_1| D = 1]$; the expected outcome[4] (the average) of treated individuals Y_1 in the treated group $D = 1$; and $E[Y_0| D = 0]$; the average outcome of non-treated individuals Y_0 in the non-treated group $D = 0$.

The theoretical, but impossible to obtain, target would be to measure the impact of the treatment on the treated group over exactly the same group of population (the treated individuals) without the treatment:

$$E[\Delta| D = 1] = E[Y_1| D = 1] - E[Y_0| D = 1] \tag{10.1}$$

where the unobserved potential average outcome of non-treated individuals in the treated group $E[Y_0| D = 1]$ constitutes the identification problem of impact evaluation. Several methodological strategies; randomised controlled trials (RCT), regression discontinuity designs, instrumental variables and differences in differences, have been developed in statistics and econometrics to deal with this problem. Table 10.1 provides a rough description of these techniques.

Additionally, the propensity score matching (PSM) technique (Heckman and Navarro-Lozano, 2004) has been used in education economics to account for causal effects. When there was no randomisation it is possible to match beneficiaries in the treated group with non-beneficiaries in the control group using observed (before the treatment) variables. Although this method can mitigate the problem of self-selection, because we can assume that estimations are done just with similar individuals, it is unlikely that the assumption of no unobserved differences between the treated and empirically derived control group, essential for the propensity score strategy, held. For this reason, we think that PSM lags behind the ones described in Table 10.1 in terms of its potential ability to identify causal evidence. In the following, we shall describe the quasi-experimental techniques showed in Table 10.1.

[3] In the absence of a specific intervention, the focus is mainly on monitoring the standard school activity, that is, the transformation of input into outputs through the educational activities. In this case, the most accurate toolbox for analysing and comparing the school's management performance is to carry out a productivity and/or efficiency analysis.

[4] For simplification, we will use the term expectation and the average result as equivalent expressions.

Table 10.1 Description of the impact evaluation methods most used in education economics

Approach	Description	Advantages	Drawbacks
Randomised Controlled Trials (RCTs)	Individuals are randomly assigned to the treated and control groups through a social experiment.	Both groups will be distributed identically in all variables but in receiving the treatment. When is well designed, the results are highly robust.	Is more expensive than the other alternatives to guarantee external validity. In occasions, it raises ethical problems to run a social experiment.
Regression Discontinuity Designs (RDD)	Participation is decided through an exogenous cut-off point, normally a law requirement.	The cut-off point reproduces a random experiment. Is cheap, easy to apply and provides robust results. It suits well with educational policies based on rules such grants, entry criteria and so on.	Results are local average treatment effects in the sense that they could not be generalised for individuals far away from the cut-off-point.
Instrumental Variables (IV)	The nature or the legal framework originates exogenous sources of variation correlated with receiving the treatment but uncorrelated with the dependent variable.	The method exploits a partial random assignment that reproduces a natural experiment. It provides even robust results than their counterparts.	Most of the time is quite difficult to find a good instrument. Finding an outstanding instrument is really hard.
Difference in Differences (DiD)	The treatment is exogenous for the treated group. The treated and counterfactual groups would have the same trend in the absence of the treatment.	The method is easy to apply and provides robust results.	Data demanding in terms of 'before' and 'after' the treatment. It is necessary to run different test and models to demonstrate the equal trends assumption that guarantees that the method provides causal results.

2.1 Randomised Trials

The gold standard identification strategy used to deal with the identification problem and determining causality is randomisation (Angrist, 2004). Assigning the treatment through a random method guarantees that both groups will be distributed identically in all observed and non-observed relevant variables. Summarising, randomisation ensures that: $E[Y_0|\, D = 1] = E[Y_0|\, D = 0]$ so the causal impact of the treatment can now be estimated as:

$$E[\Delta|\, D = 1] = E[Y_1|\, D = 1] - E[Y_0|\, D = 0] \tag{10.2}$$

In the last decade several randomised trials have been carried out in the context of education with the aim of obtaining robust impact evaluations, for instance those focused on the evaluation of early childhood education programmes (Heckman et al., 2010; Jensen et al., 2013), charter schools (Angrist et al., 2016), changes in the size of the classroom (Chetty et al., 2011; Dynarski et al., 2013), the use of ICT (Angrist and Lavy, 2002; Banerjee et al., 2007), the effectiveness of extended day programmes (Meyer and Van Klaveren, 2013) or the implementation of incentives for teachers or students (Dee and Keys, 2004; Kane and Staiger, 2008; Angrist et al., 2009: Fryer, 2010;; Fryer et al., 2012; Duflo et al., 2011, 2012, 2015).

Although the number of social experiments in education that uses randomisation in the real world is growing rapidly, until now most educational public programmes and interventions to boost academic results do not implement randomisation when they are designed. In these cases, we cannot guarantee that $E[Y_0|\, D = 1] = E[Y_0|\, D = 0]$ holds and what we actually estimate is:

$$E[\Delta|\, D = 1] = E[Y_1|\, D = 1] - E[Y_0|\, D = 0] + \{E[Y_0|\, D = 1] - E[Y_0|\, D = 1]\} \qquad (10.3)$$

where by rearranging the terms we have:

$$E[\Delta|\, D = 1] = E[Y_1|\, D = 1] - E[Y_0|\, D = 1] + \{E[Y_0|\, D = 1] - E[Y_0|\, D = 0]\} \qquad (10.4)$$

or:

$$E[\Delta|\, D = 1] = E[Y_1|\, D = 1] - E[Y_0|\, D = 1] + B \qquad (10.5)$$

The lack of randomisation causes a potential bias B in the measurement of the causal effect of D on the difference in expected results for both the treated and the control group. As the term $E[Y_0|\, D = 1]$ is unknown, it is impossible to determine the magnitude of this bias, which results in a misleading estimation and which is hardly conducive to taking decisions.

Therefore, organising random experiments is not always possible, nor desirable. Instead, in many practical circumstances, analysts should rely upon data generated for other purposes, such as surveys, exercises or even administrative data. To tackle the endogeneity problem in the absence of randomisation, the econometric literature proposes a growing toolbox of causal inference methods for evaluating the impact of interventions (the treatment) using observational data, the so-called 'quasi-experimental' evaluation methods. The root idea is to look for a good counterfactual group to try to avoid or at least minimise selection bias derived from the lack of intended randomisation (Khandker et al., 2010). It is out of the scope of this chapter to review all these techniques[5] in depth. Instead we provide a basic overview of the main and more robust approaches used in education economics; instrumental variables, regression discontinuity designs and the difference in differences.

[5] There are many references in the literature that introduce impact evaluation and causal inference methods. For the interested reader, we suggest Angrist and Pischke (2008, 2014) for a technical point of view and Schlotter et al. (2011) and Webbink (2005) for a non-technical point of view applied to the education sector.

2.2 Instrumental Variables

The basic idea behind this approach is to find an exogenous source of variation, the instrumental variable (IV), correlated with having received the treatment but uncorrelated with the outcome. Exogenous sources of variation are difficult to find and so this approach requires creativity on the part of the researcher, the availability of rich databases and a profound knowledge about the intervention and the circumstances under it was developed. Frequently, a starting point for finding an instrument is to check for legal or natural variations during the period analysed (see Angrist and Krueger, 1991; Angrist and Lavy, 1999; Hoxby, 2000; West and Woessmann, 2010; Jensen and Rasmussen, 2011; Kearny and Levine, 2015, for example). Once an IV is found, we can keep the part of the treatment that is exogenous and free from endogeneity bias. Remember that the basic regression for analysing the impact of a programme is:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i \tag{10.6}$$

where D_i is a dummy variable that indicates whether the individual belongs to the treated group ($D = 1$) or the control group ($D = 0$); X_i is a set of covariates and Y_i corresponds to the outcome. In this equation β_1 provides the impact of this treatment only if it is exogenous, that is $corr(D, u_i) = 0$. When this condition does not hold, an exogenous variable Z helps us to isolate the exogenous part of D. The instrument may fulfil two conditions: being correlated with D so that $corr(D, Z) \neq 0$ (instrument relevance) but uncorrelated with u, and hence $corr(Z, u_i) = 0$ (instrument exogeneity). While the first condition can be examined using the observed data, the second cannot be tested directly because the error term is not observed in empirical settings. In this case 'we must maintain this condition by appealing to economic behaviour or introspection' (Wooldridge, 2012 p. 514). To apply the IV method, we proceed in two stages (two stage least squares).

1. Estimate a predictor for D using the instrument and the rest of exogenous covariates:

$$\hat{D} = \hat{\pi}_1 + \hat{\pi}_2 Z + \hat{\pi}_3 X + \varepsilon \tag{10.7}$$

2. Substitute $\hat{D}$ for D in Equation (10.6) to obtain a robust estimation of β_1, the impact of the treatment:

$$Y = \beta_0 + \beta_1 \hat{D} + \beta_2 X + \varepsilon \tag{10.8}$$

   β_1 in (10.8) provides the measure of the impact of the treatment on the output considered.

2.3 Regression Discontinuity Design

Regression Discontinuity Design (RDD) was introduced in the evaluation literature by Thistlethwaite and Campbell (1960) when they tried to study the effect of a scholarship only granted to those students who obtained specific test scores above a threshold. This method has been widely applied in education to evaluate diverse issues as the effect of

class size on students' performance (Angrist and Lavy, 1999), the impact of university financial aid awards on college enrolment (Van der Klaauw, 2002), the influence of grade retention on educational attainment (Jacob and Lefgren, 2004), the impact of the Head Start programme on childrens' life chances (Ludwig and Miller, 2007), the effect of attending a mandatory summer school on test scores in the following year (Matsudaira, 2008), the impact of the month of birth on cognitive and non-cognitive skills (Crawford et al*.,* 2014), the effect of the IMPACT programme (a performance-based incentives system based on rigorous teacher evaluations) on teachers' retention and performance (Dee and Wykoff, 2015), among others.

The RDD is the most appropiate method for programmes in which participation is decided through a cut-off point so that whether or not an individual is treated depends on their position relative to a threshold on some continuous variable. As the cut-off is usually decided arbitrarily by an external rule, normally to adjust the available budget to the expected population, around the cut-off gives rise to a natural random experiment in which individuals are comparable in all respects other than that those on one side of the threshold receive the treatment while those on the other do not. Therefore, differences in outcomes can be entirely attributed to the intervention itself (Gertler et al., 2016).

Let us assume a programme which has a continuous eligibility index, X_i, with a strictly defined cut-off point, $\overline{x}$, to determine who is eligible and who is not. Then, if D_i denotes the treatment then:

$$D_i \begin{cases} 1 \ if \ X_i \leq \overline{x} \rightarrow Treated \\ 0 \ if \ X_i > \overline{x} \rightarrow Non-treated \end{cases} \tag{10.9}$$

There are two main general settings within the RDD. The *sharp regression discontinuity design* is applied when a running variable X_i which defines the treatment and control group precisely by running the following equation:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon \tag{10.10}$$

where D_i indicates whether the individual belongs to the treated group or the control group; X_i is the running variable and Y_i corresponds to the outcome. On the other hand, in the *fuzzy regression discontinuity design* (FRDD) the running variable does not determine the treatment group perfectly but creates a discontinuity in the probability of receiving the treatment (Schlotter et al*.,* 2011). This applies when the eligibility rules are not strictly adhered to as some unobserved variables rule the assignment to treatment (Hahn et al*.,* 2001).

FRDD can be analysed in an instrumental variables framework, defining a simple indicator, denoted by I_i, to determine whether the running variable X_i is below or above the cut-off point and using it as an instrument for treatment variable D_i in the estimation of the outcome equation (Angrist and Pischke, 2008). FRDD is estimated using the following equations:

$$\text{First stage or treatment equation:} \quad D_i = \gamma_0 + \gamma_1 I_i + \gamma_2 X_i + \varepsilon \tag{10.11}$$

$$\text{Second stage or outcome equation:} \quad Y_i = \beta_0 + \beta_1 \hat{D}_i + \beta_2 X_i + \varepsilon \tag{10.12}$$

where D_i is the estimated treatment variable in the first stage and denotes the probability of receiving the treatment.

There are several concerns to consider when RDD is applied. First, the running variable should not be manipulated to ensure assignment to treatment. Second, the specification may be sensitive to the functional form used in modelling the relationship between the assignment variable and the outcome variable (Gertler et al., 2016). On the other hand, RDD produces local average treatment effects that cannot necessarily be generalised to units far away from the cut-off point (Kandher et al., 2010). Finally, it is not always possible to find enough observations close enough to the threshold – the method involves 'throwing away' observations that are far from the cut-off point. To solve the problem related to the limited sample size, the interval around the cut-off point can be increased, but as we move further away from the eligibility threshold, the eligible and ineligible units will become more different, which can bias the comparison (Schlotter et al., 2011). Including more covariates in the aforementioned equations may eliminate some bias resulting from the higher bandwidths (Imbens and Lemieux, 2008).

2.4 Difference in Differences

The Difference-in-Differences approach (*DiD*) estimates the impact of an intervention by comparing the average outcomes of the treated and control groups both before and after the treatment. The key identifying assumption of the *DiD* approach is that the trend in the outcomes would be the same in both treatment and control groups in the absence of treatment. In other words, both groups of individuals may be observationally different before the treatment, but these differences are time invariant in the absence of treatment, so the difference between both groups after the treatment can be attributable to the intervention. This assumption holds when one group of individuals in the sample have been exogenously exposed to the treatment (treated units).[6] Some examples of *DiD* applications in education can be found in Pischke (2007), Bellei (2009); Schlotter et al. (2011), Graves (2011), Felfe et al. (2015) and Anghel et al. (2015), Pedraja et al. (2016).

To estimate the impact of the treatment in the simplest scenario with just two periods and two groups it is necessary to take two differences into account. First, the average difference in outcomes over the analysed period separately for both the treated and control groups (first difference); and then, an additional difference between the average changes in outcomes for these two groups (second difference). The average *DiD* treatment effect can be calculated as:

$$DiD = [E(Y_1^T - Y_0^T|T = 1)] - [E(Y_1^C - Y_0^C|T = 0)] \tag{10.13}$$

where Y_1^T and Y_0^T represent the average outcome for the treated group both after and before the treatment respectively, Y_1^C and Y_0^C represent the average outcome for the control group after and before the treatment respectively.

[6] This assumption is sometimes unlikely to hold for example, if the treatment was not exogenous. To overcome this problem, it is worth, for example, running a propensity score matching (PSM) in the pre-treatment year to obtain similar treated and control groups in terms of observable characteristics before the programme starts. Alternatively, if a panel database with some periods before and after the treatment is available, another key identifying assumption is to verify that trends in the treatment and the control groups are equal in the absence of treatment.

The *DiD* estimator is usually solved using a linear regression by estimating the following equation:

$$Y_{it} = \alpha + \beta T_i t + \rho T_i + \gamma t + \varepsilon_i \tag{10.14}$$

where T is the treatment variable (which takes the value 1 if the individual belongs to the treatment group, and 0 if not), t is the time dummy variable (1 denotes the value for the periods after the treatment and 0 before) and the coefficient β associated to the interaction between T and t, represents the estimated impact of the treatment (the *DiD* effect in Equation 10.13).[7] The model can be generalised in many ways to adapt the estimation to our empirical problem. For example, the time variable can be replaced for a set of dummies representing each year after the treatment. These time effects allow following the changes produced by the treatment over time. Another strategy is to introduce fixed unit effects, or even unit time trends (Angrist and Pischke, 2008).

3 A STEP FORWARD: IMPACT EVALUATION AND PRODUCTION FRONTIERS

Traditionally, educational policy has sought to change outcomes through an increase in input. However, although theoretically public policies oriented to giving more resources to schools should work, they are not a guaranteed success. Moreover, there is no reason to suppose that each type of intervention would provide the same bang per buck, and thus it is important to know what are the most (and least) effective interventions so that finite resources can be invested most effectively. For this reason, most educational programmes and research in education nowadays are not devoted to increasing the budget but to improving the school's management and the educational practices inside the classrooms. These policies do not imply a significant change in the observed inputs but an alternative way to improve the schools' productivity.

This chapter contributes to the impact evaluation literature by proposing a new approach, based on production frontiers, not only to compare the final average results between treated and control schools to evaluate the causal impact of an intervention, but also to analyse how a treatment implemented in the schools can influence the production activity of a group of treated schools in comparison to the control schools. The novelty of this approach is to show that the causal influence of a successful programme on a group of treated units with respect to a control group cannot only be visible in the average output difference but also can be detected by measuring the total factor productivity changes (*TFPC*) caused by technology and/or efficiency changes, that can occur throughout the treatment.

Let us assume that the educational production function for a group of schools can be defined using a vector of inputs $x = (x_1, \ldots, x_k) \in \Re^{K+}$ and outputs $y = (y_1, \ldots, y_l) \in \Re^{L+}$. A feasible production technology can be defined using the output possibility set

[7] We can rewrite Equation (10.13) using Equation (10.14) as $DiD = [E(Y_1^T - Y_0^T|T = 1)] - [E(Y_1^C - Y_0^C|T = 0)] = [(\alpha + \beta + \rho + \gamma + \varepsilon) - (\alpha + \rho + \varepsilon)] - [(\alpha + \gamma + \varepsilon) - (\alpha + \varepsilon)]$ from which we have $DiD = [(\beta + \gamma) - \gamma] = \beta$.

$P(x)$, which can be produced using the input vector x: $P(x) = \{y: x \text{ can produce } y\}$, which is assumed to satisfy the set of axioms described in Färe and Primont (1995). Following this scheme, the well-known educational production function proposed by Levin (1974) and Hanushek (1979) for a set of S schools $s = 1, \ldots, S$ is:

$$y_s = A.F(x_s)\cdot u_s \tag{10.15}$$

where sub index s refers to school, and y_s represents the educational output vector while the vector of educational inputs x_s capture the average student's family, social, cultural and economic background together with school educational resources, and A accounts for changes in total output growth relative to changes in the technology.

As Levin (1974) comments, one of the major assumptions derived from market theory that tacitly underlies the estimation of educational production functions is that schools are technically efficient, that is, that they are maximising output given the input mix that they have selected. However, in real life, it is quite frequent to detect inefficient behaviours. According to Leibenstein (1966) the source of inefficiency mainly comes from light competition pressure. Efficiency in schools may be due to multiple factors related with management, incentives structure, clear targets and factors related to the motivation of the agents involved in the educational process. Levin (1974) provides and develops six sources of inefficiency in education; (i) managerial knowledge of the technical production process; (ii) substantial managerial discretion over input mix; (iii) a basic competitive environment with all of its attendant assumptions (freedom of entry, many firms, perfect information); (iv) managerial knowledge of prices for both inputs and outputs; (v) an objective function that is consistent with maximising output such as profit maximisation and (vi) clear signals of success or failure (profits, losses, sales, costs, rate of return, share of market). Although all these factors are not direct inputs, they may significantly affect student performance.

Another important issue to be underlined here is that the education service produces several outputs, although educational achievement is the one that has concentrated major attention in the literature (Hoxby, 1999). The multiple dimensions of cognitive (for example, mathematics, reading or science test scores) and non-cognitive outputs (for example, the *big five* personality characteristics (Heckman, 2011)), raise into consideration the relationship between inputs and outputs jointly with the trade-off between the different outputs that is feasible to produce with a vector of inputs.

For these two reasons, the educational production function is frequently estimated in education economics through production frontiers considering a multi-output multi-input framework that incorporates the possible existence of inefficient behaviours in schools (for a review see Worthington, 2001; Johnes, 2015 and De Witte and López-Torres, 2015). In Equation 10.15, u_s captures the efficiency level of school s and is distributed over the interval $0 < u_s \leq 1$. Values of $u_s = 1$ imply that the school is fully efficient, meaning that, given the initial input endowment and the existing technology, this school is maximising its outputs and correctly managing the school inputs available given existing technology. Values of $u_s < 1$ indicate that the school is inefficient, and therefore the efficiency rate, $\theta_s = 1/u_s$ indicates the amount by which the actual output should be multiplied to reach the frontier.

In short, when the policymaker applies a treatment to raise the educational outputs, y_s, the programme modifies one of the following factors: the educational inputs x_s, the

technology A or the technical efficiency u_s. From now on we discuss the implications for the analysis of this framework in a randomised trial. We are aware that although randomised interventions are considerable growing worldwide in the last decade, they are not the most common way to carry out and evaluate educational public policies yet. However, as this chapter aims to introduce and illustrate a new approach we try to make it from a simple viewpoint, if before the intervention both groups are equal on the average inputs, technology and efficiency. Soon, a fruitful contribution will be to extend this approach to the case where before implementing the treatment, both groups may differ in terms of input, technology or efficiency.

3.1 The Randomised Trial in the Simplest Production Setting

To introduce and illustrate these ideas briefly, let us parameterise Equation 10.15 using a Cobb-Douglas specification as follows:

$$y_s = A.\prod_{k=1}^{K} x_{ks}^{\beta_k} \cdot u_s \tag{10.16}$$

where x_{ks} represent the $k = 1, \ldots, K$ inputs to produce a unique output y_s. From now on we assume a single input, single output setting[8] in which we carry out a randomised trial to introduce an educational treatment. Figure 10.1 illustrates this situation. The left-hand panel plots a set of schools *before the treatment* and constitutes the baseline scenario in which all schools share the same technology and have different inefficiency levels represented by the distance of each dot to the production frontier (in terms of the output). The right-hand panel shows how schools are randomly divided into two groups where the white and black circles represent the schools assigned to the treated and counterfactual groups respectively.

Randomisation assigns N decision making units (DMUs) – represented by white circles – to the treated group, whereas the M black circles represent the DMUs in the control group, $N + M = S$. Points T (the white diamond) and C (the black diamond) in the right-hand panel in Figure 10.1 represent, for illustration purposes, the theoretical average production activity observed for the treated and control groups respectively, where:

$$\bar{y}_T = \frac{1}{N}\sum_{i=1}^{N} y_i;\ \bar{x}_T = \frac{1}{N}\sum_{i=1}^{N} x_i;\ \bar{u}_T = \frac{1}{N}\sum_{i=1}^{N} u_i \text{ and } \bar{y}_C = \frac{1}{M}\sum_{j=1}^{M} y_j;\ \bar{x}_C = \frac{1}{M}\sum_{j=1}^{M} x_j;$$

$$\bar{u}_C = \frac{1}{M}\sum_{s=j}^{M} u_j.$$

Randomisation guarantees that mean differences in inputs $\bar{x}_T \cong \bar{x}_C$ and outputs $\bar{y}_T \cong \bar{y}_C$ will not be statistically significant different from zero when both groups are compared

[8] The Cobb-Douglas approximation is employed in most of regressions run in impact evaluation. Although frontier analysis can deal with multiple outputs and multiple inputs at the same time, for the sake of simplicity and to be able of representing the technology in a graph we assume the simplest specification of Equation 10.15, a technology with just one input K = 1 and one output L = 1 under constant returns to scale. This framework is drawn in Figure 10.1 assuming $\beta = 1$.

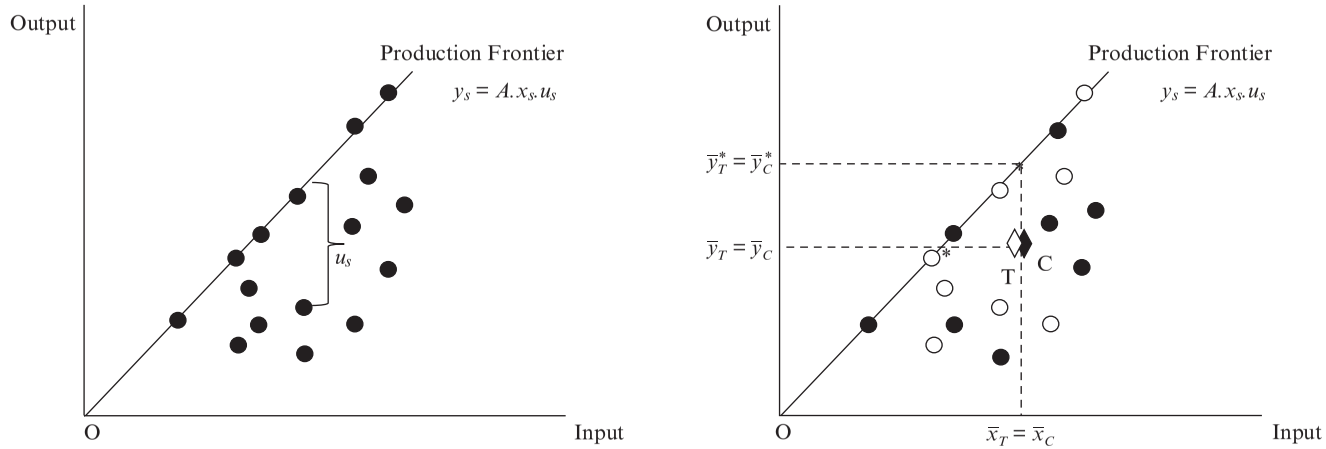

Figure 10.1 Productivity frontier of a set of schools before the treatment starts

after the randomised trial but before the treatment starts.[9] Likewise, the production activity information can be used to estimate both production frontiers and efficiency scores for all schools using DEA. Then, we could also estimate the average efficiency of each group projecting each theoretical average school T and C upwards, following an output orientation, up to the respective production frontier. In Figure 10.1 the mean efficiency of the treated and control groups is $\bar{u}_T < 1$ and $\bar{u}_C < 1$ respectively, where again randomisation guarantees that $\bar{u}_T \cong \bar{u}_C$.

3.2 Educational Treatments Introduced by a Randomised Trial

Here we simply set out four theoretical potential changes that the treatment could bring about in the production function.[10] In all cases, we will assume that the treatment produces a positive change in the treated group although in real life changes due to an intervention programme could result in positive, negative or no changes with respect to the control group. The left-hand panel in Figures 10.2, 10.3, 10.4 and 10.5 reproduces the right-hand panel in Figure 10.1, that is, the initial situation just after the randomised trial in which we decided which schools will be treated and which will be the controls but the treatment had not yet started.

3.2.1 A treatment that changes the input level in the treated group

An intervention that changes one input is, basically, a policy that increases one of the controllable inputs to the treated schools $\bar{x}_T' > \bar{x}_T$. The clearest example is one in which the number of teachers in the school is increased. The dots T' and C in the right-hand panel in Figure 10.2 denote the theoretical average production activity for the treated and the control groups after the treatment. We now observe a difference in the average output $(\bar{y}_T' - \bar{y}_C > 0)$ concluding that the treatment, the change in the input level $(\bar{x}_T' - \bar{x}_C > 0)$, was effective. Nevertheless, we can also see that both technologies $A_T = A_C$ and efficiency levels $\bar{u}_T = O\bar{y}_T'/O\bar{y}_T^{*\prime} \cong O\bar{y}_C/O\bar{y}_C^{*} = \bar{u}_C$ remain without significant differences.

3.2.2 A treatment that improves the efficiency level

In this case, we assume a treatment that only changes the managerial efficiency of the treated schools $\bar{u}_T' > \bar{u}_T$. Some examples of these interventions are to adjust educational content to the individual student's needs, to offer intensive teacher training programmes, to sort students into classrooms by prior achievement level or to introduce new teacher-learning practices, among others. Once the treatment finishes in the right-hand panel in Figure 10.3 we again observe a difference in the average output $(\bar{y}_T' - \bar{y}_C > 0)$ concluding that the treatment was effective. However, in this case there is no change in the input levels $\bar{x}_T \cong \bar{x}_C$ so differences in outputs are due to total factor productivity change (*TFPC*). As both technologies are coincident $A_T = A_C$, efficiency levels $\bar{u}_T = O\bar{y}_T'/O\bar{y}_T^{*\prime}$ and $\bar{u}_C = O\bar{y}_C/O\bar{y}_C^{*}$ are the only aspects responsible for this *TFPC* because the treatment caused a better management of the treated school giving rise to $\bar{u}_T' > \bar{u}_C$.

[9] In finite samples, little but not significant differences can arise by chance between both groups.

[10] Certainly, we could derive many other scenarios but the aim of this chapter is to illustrate the main sources that originate changes in outputs. Then, any combination could be easily replicable.

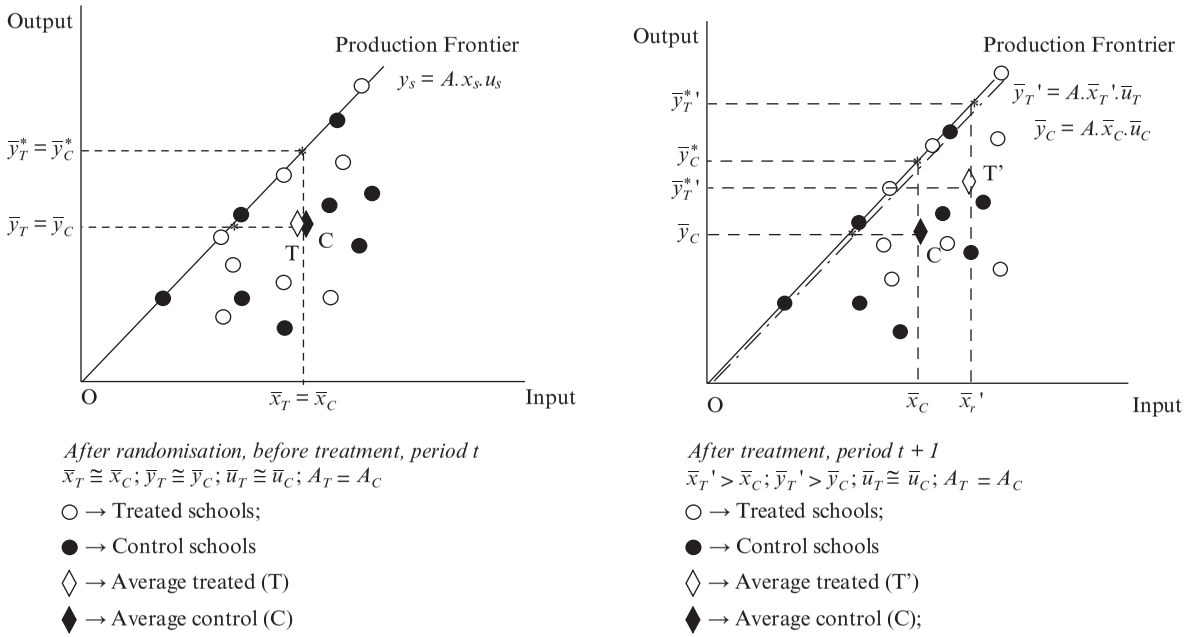

Figure 10.2 Production frontier of treated and control school before and after a treatment that brings about a rise in input in the treated schools

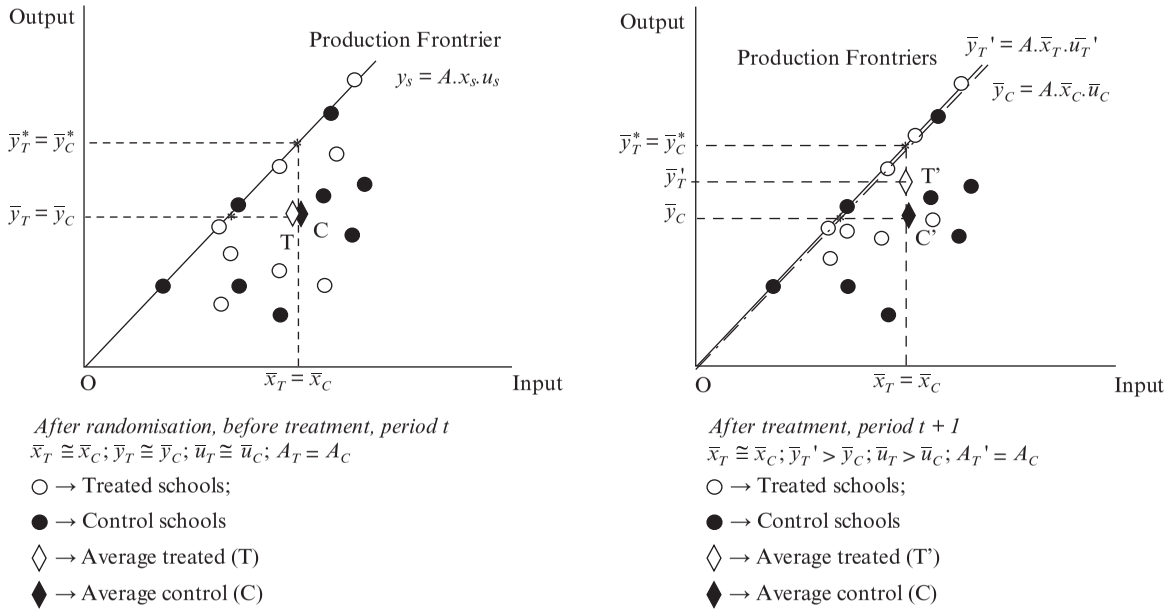

Figure 10.3 Production frontier of treated and control school before and after a treatment that brings about a rise in efficiency in the treated schools

3.2.3 A treatment that brings about a technological change

The final pure illustrated intervention is a treatment that gives rise to a positive shift in the technology of schools belonging to the treated group $A_T' > A_T$. Implementing or publicising the results of standardised external tests, introducing changes in the educational curriculum or tracking students are just a few of the illustrative interventions that could bring about a technological change in the education system. This situation is represented in the right-hand panel in Figure 10.4 through the production frontier drawn with the dashed line while the straight line displays the production technology for the control group. Again, once the treatment is finished we observe a positive difference in the average output $(\bar{y}_T' - \bar{y}_C > 0)$ concluding that the treatment was effective. In this case, there is neither a change in the input level nor in the efficiency levels $\bar{u}_T = O\bar{y}_T'/O\bar{y}_T^{*\prime} \cong O\bar{y}_C/O\bar{y}_C^{*} = \bar{u}_C$ leading us to conclude that differences in outputs originate from differences in the technologies used to transform inputs into outputs $A_T' > A_C$.

3.2.4 A treatment that brings about both a technological and efficiency change

In this last scenario we define a treatment that simultaneously positively shifts the production frontier $A_T' > A_T$ and increases the managerial efficiency in the treated group $\bar{u}_T' > \bar{u}_T$. Figure 10.5 provides this framework. The difference between the two outputs $(\bar{y}_T' - \bar{y}_C > 0)$ brought about by the treatment is caused by a positive *TFPC* driven by a technological gap as the treated units are now more productive in terms of technology $A_T' > A_C$; but also by an efficiency gap because the treated schools are more efficient than the control ones $\bar{u}_T = O\bar{y}_T'/O\bar{y}_T^{*\prime} > O\bar{y}_C/O\bar{y}_C^{*} = \bar{u}_C$.

3.3 A New Approach: Impact Evaluation Through Production Frontiers

As we discussed earlier, impact evaluations carried out in randomised trials generally estimate the average treatment effect on the welfare, outcomes or outputs of treated schools with respect to the counterfactual group. To do this, researchers normally carry out mean differences tests in these variables between both groups. However, we also wonder through which channels the treatment is changing the average output in the treated schools. According to Equation 10.15, changes in outputs can be explained through three channels in the production process:

1. A change in inputs: the treatment can consist of increasing the endowment of one or more inputs in the treated group with respect to the control group. After the intervention differences in inputs x_s can explain the differences in observed outputs.
2. A technological change: because of the treatment, the production technology A may increase and result in a positive change in outputs in the treated group.
3. An efficiency change: a treatment could influence the managerial activity in the treated schools leading to an improvement in the average technical efficiency u_s in the treated group.

In other words, after the treatment we allow the production frontiers of each group to vary in different ways. For the treatment group the production function is now $y_{sT} = A_T \cdot F(x_{sT}) \cdot u_{sT}$ while for the control group the new production function is $y_{sC} = A_C \cdot F(x_{sC}) \cdot u_{sC}$.

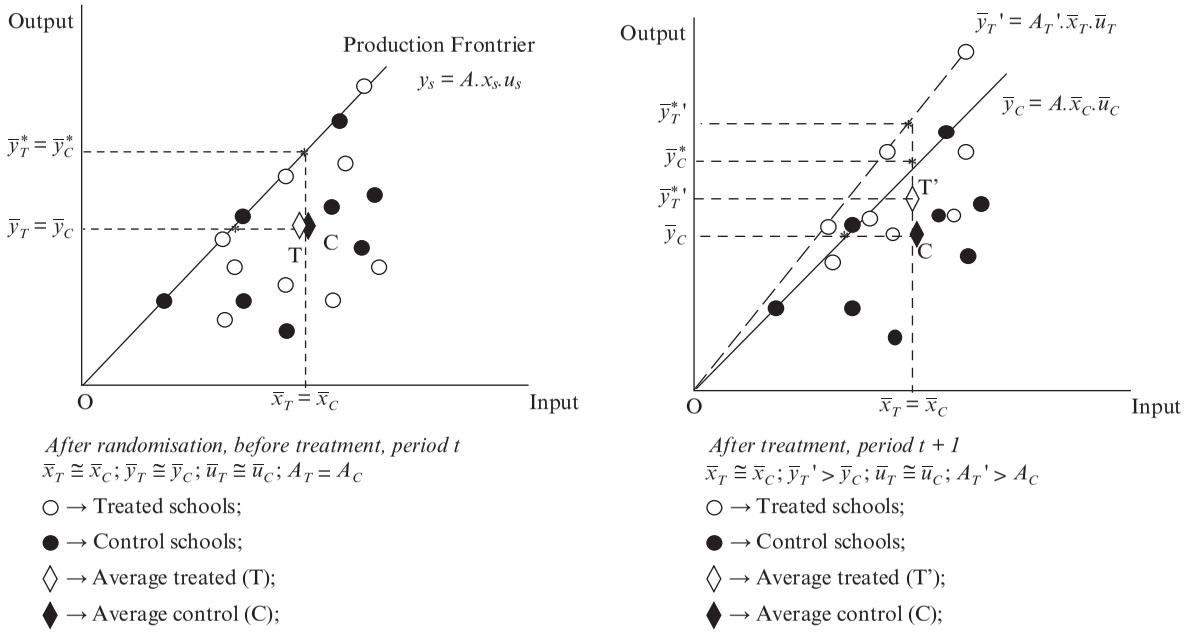

Figure 10.4 Production frontier of treated and control school before and after a treatment that brings about a technological change in the treated schools

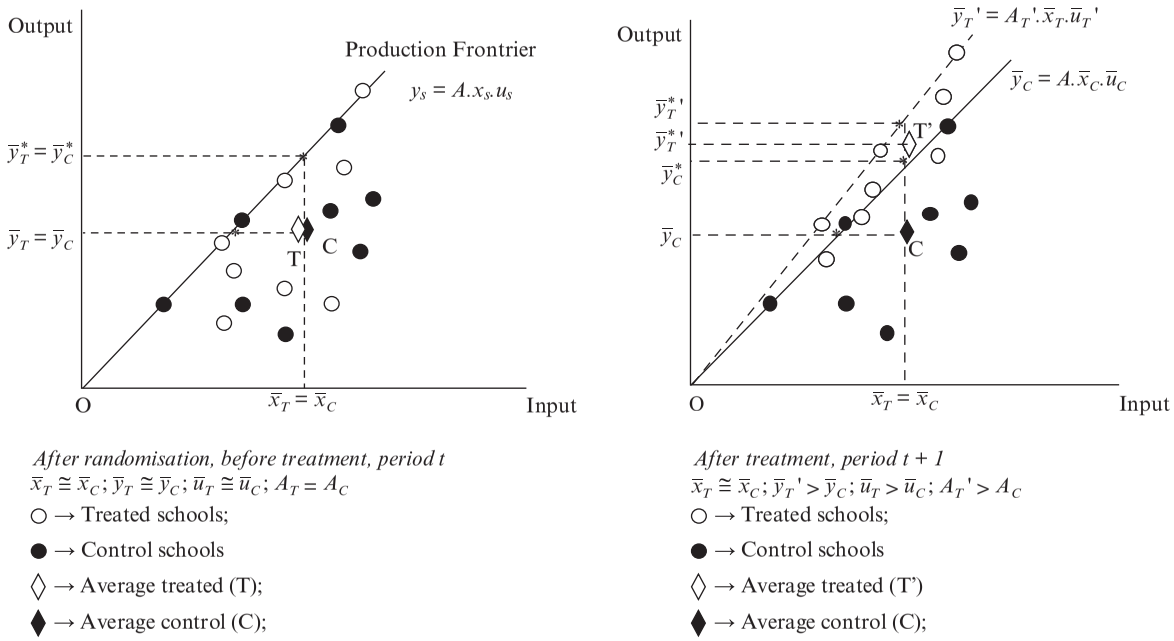

Figure 10.5 Production frontier of treated and control school before and after a treatment that brings about both a technological and an efficiency change in the treated schools

By averaging the production information for all treated and control schools in the post-treatment period we can obtain the following average production functions: $\overline{y}_T = A_T.F(\overline{x}_T) \cdot \overline{u}_T$ and $\overline{y}_C = A_C.F(\overline{x}_C) \cdot \overline{u}_C$. Operating by dividing both expressions we obtain the variation in each component of the production process as follows:

$$\frac{\overline{y}_T}{\overline{y}_C} = \frac{A_T.F(\overline{x}_T) \cdot \overline{u}_T}{A_C.F(\overline{x}_C) \cdot \overline{u}_C} = \frac{A_T}{A_C} \cdot \frac{F(\overline{x}_T)}{F(\overline{x}_C)} \cdot \frac{\overline{u}_T}{\overline{u}_C} \tag{10.17}$$

where changes in the average output between both groups $\frac{\overline{y}_T}{\overline{y}_C}$ after the treatment can be decomposed into changes in the technology $\frac{A_T}{A_C}$ and/or in inputs $\frac{F(\overline{x}_T)}{F(\overline{x}_C)}$ and/or in efficiency $\frac{\overline{u}_T}{\overline{u}_C}$.

This approach allows a step forward because, by combining causal inference and production frontiers, we can evaluate the impact of the treatments and disentangle the causes that give rise not only to a change in the average outputs, but also whether the programme brings about changes in terms of the total factor productivity due to technological and/or efficiency changes. The usefulness of this approach is also relevant in several scenarios frequently observed when educational policies are implemented. Owing to space limitations, however, we cannot illustrate that in this chapter. An example would be when an educational programme initially causes an upward shift of the production frontier led by a reference set of schools that constitutes the best practices applying the programme. In this case, receiving feedback from monitoring and evaluating the best performers is relevant to translate best practices into a second stage for the remaining schools. This process will enhance the programme in a second step leading to an effective catching-up process. However, a priori, if were only to evaluate the programme in terms of the average output we would conclude that it has no effect on the schools treated. Another example could be a treatment that is effective only for the most productive units but where this effect is not converted into a significant difference in average outputs due to the existence of unaccounted losses in efficiency.

As technology and efficiency are not *observed* in empirical educational applications they must be *estimated* from the observed data sample. We propose an adaptation of the Malmquist index methodology to estimate changes in total factor productivity *TFPC*, in both technology and efficiency after the treatment.

3.3.1 The estimation of total factor productivity changes

The Malmquist index was proposed by Caves et al. (1982) with the aim of measuring the *TFPC* between two data points within two time periods as the ratio of the distances of each data point relative to a common frontier. The index may be built and decomposed using several data envelopment analysis (DEA) programmes to compute different distances between the evaluated production unit and the frontier for each period. Following Färe et al. (1994) the output-orientated Malmquist productivity index for two periods of time t and $t + 1$ under a constant returns to scale technology[11] can be written as:

[11] There are other studies that also consider variable returns to scale. However, it is well known (Pastor and Lovell, 2005) that infeasibilities can arise when DEA is used to compute the distance functions constituting the Malmquist decomposition when the scale component is considered in the productivity growth. For the sake of simplicity in this paper we follow Camanho and Dyson (2006) assuming a constant returns to scale technol-

$$MI(x^{t+1}, y^{t+1}, x^t, y^t) = \frac{D^{t+1}(x^{t+1}, y^{t+1})}{D^t(x^t, y^t)} \cdot \left[\left(\frac{D^t(x^{t+1}, y^{t+1})}{D^{t+1}(x^{t+1}, y^{t+1})} \right) \cdot \left(\frac{D^t(x^t, y^t)}{D^{t+1}(x^t, y^t)} \right) \right]^{1/2} = TEC^*TC \tag{10.18}$$

Where the super-index indicates the time period and $D^{t+1}(x^t, y^t)$ represents the distance from the period t observation (x^t, y^t) to the period $t + 1$ technology $D^{t+1}(\cdot)$. A Malmquist index higher (lower) than one implies productivity improvements (losses) from period t to period $t + 1$. Furthermore, Equation (10.18) includes two components. The first ratio reflects the technical efficiency change (*TEC*), which captures the efficiency improvements (reductions) in period $t + 1$ with respect to period t if $TEC > 1$ ($TEC < 1$), whereas $TEC = 1$ indicates no changes in technical efficiency. The second measure (in squared brackets) represents the technological change (*TC*) in period $t + 1$ with respect to period t, whose value may be analysed in a similar way to *TEC*, ($TC > 1$ now implies technological progress). The two measures may of course go in different directions. As De Witte and López-Torres (2015) state 'the decomposition in a Malmquist index can help to open the black box of effect studies as it shows what exactly is driving the results'.

The standard Malmquist index methodology requires observing a group of DMUs in two different periods. To compare the treated and the control group under this evaluation framework we need to calculate two different Malmquist indices, one for each group, using an unbalanced panel data in which data in the baseline period t are shared by the two groups. The explanation is easy to demonstrate. Before a randomised trial starts we know that all schools, treated and control, constitute the production set sharing the same common technology as shown in the left-hand panel in Figure 10.1. This implies that the whole sample of schools will be used to define the initial production technology in period t. After the treatment, and as we discuss above, if the technology may change differently in both groups, for period $t + 1$ and subsequent periods we will use only the data for the treated and control groups separately to estimate each new by-group production frontier technology.

Therefore, we run this By-Group Malmquist Index (BGMI) method for treated (control) schools using an unbalanced panel database in which period t contains all school information, the whole sample made up of treated and control schools,[12] but in period $t + 1$ we only use the treated (control) schools. The BGMI for the treated and the control groups can be defined as:

ogy. In the case of empirical applications in education, scale is not a major problem and variables are usually normalised to avoid school size issues.

[12] If the production technology is the same in period t an alternative strategy would be to run the Malmquist indexes separately for the treated and the control groups using two balanced panel data comparing the mean results. However, we know that in finite samples the estimated production frontier in period t could vary due to exogenous differences or to random noise as the frontier is built with the information available. This fact could lead to estimating different technologies in period t for both groups although this problem is reduced, as long as sample size increases.

$$BGMI_T(x_T^{t+1}, y_T^{t+1}, x_{TC}^t, y_{TC}^t) = \underbrace{\frac{\left(\prod_{n=1}^{N} D_T^{t+1}(x_T^{t+1}, y_T^{t+1})\right)^{1/N}}{\left(\prod_{s=1}^{S} D_{TC}^t(x_{TC}^t, y_{TC}^t)\right)^{1/S}}}_{TEC_T} \cdot \underbrace{\left[\frac{\left(\prod_{n=1}^{N} D_{TC}^t(x_T^{t+1}, y_T^{t+1})\right)^{1/N}}{\left(\prod_{n=1}^{N} D_T^{t+1}(x_T^{t+1}, y_T^{t+1})\right)^{1/N}} \cdot \frac{\left(\prod_{s=1}^{S} D_{TC}^t(x_{TC}^t, y_{TC}^t)\right)^{1/S}}{\left(\prod_{s=1}^{S} D_T^{t+1}(x_{TC}^t, y_{TC}^t)\right)^{1/S}}\right]^{1/2}}_{TC_T} = TEC_T * TC_T \tag{10.19}$$

$$BGMI_C(x_C^{t+1}, y_C^{t+1}, x_{TC}^t, y_{TC}^t) = \underbrace{\frac{\left(\prod_{m=1}^{M} D_C^{t+1}(x_C^{t+1}, y_C^{t+1})\right)^{1/M}}{\left(\prod_{s=1}^{S} D_{TC}^t(x_{TC}^t, y_{TC}^t)\right)^{1/S}}}_{TEC_C} \cdot \underbrace{\left[\frac{\left(\prod_{m=1}^{M} D_{TC}^t(x_C^{t+1}, y_C^{t+1})\right)^{1/M}}{\left(\prod_{m=1}^{M} D_C^{t+1}(x_C^{t+1}, y_C^{t+1})\right)^{1/M}} \cdot \frac{\left(\prod_{s=1}^{S} D_{TC}^t(x_{TC}^t, y_{TC}^t)\right)^{1/S}}{\left(\prod_{s=1}^{S} D_C^{t+1}(x_{TC}^t, y_{TC}^t)\right)^{1/S}}\right]^{1/2}}_{TC_C} = TEC_C * TC_C \tag{10.20}$$

where now subscripts T and C denote schools in the treated and control groups respectively while subscript TC indicates schools in treated and control groups (all schools in the sample).

The estimated BGMI and its components can be used to empirically decompose the estimated change in the average output between both groups $\frac{\overline{y}_T}{\overline{y}_C}$ after the treatment into changes in the technology, efficiency and/or inputs (Equation 10.17). Changes in the technology $\frac{A_T}{A_C}$ can be estimated by $\frac{\overline{TC}_T}{\overline{TC}_C}$ and changes in efficiency $\frac{\overline{u}_T}{\overline{u}_C}$ can be estimated by $\frac{\overline{TEC}_T}{\overline{TEC}_C}$.

Changes in the production function $\frac{F(\overline{x}_T)}{F(\overline{x}_C)}$ after the treatment cannot be directly estimated in empirical samples, but it can be computed as the residual by:

$$\frac{F(\overline{x}_T)}{F(\overline{x}_C)} = \frac{\left(\frac{\overline{y}_T}{\overline{y}_C}\right)}{\left[\left(\frac{\overline{TC}_T}{\overline{TC}_C}\right)\right]\left(\frac{\overline{TEC}_T}{\overline{TEC}_C}\right)} \tag{10.21}$$

Thus, the decomposition of the estimated change in the average output between both groups $\frac{\overline{y}_T}{\overline{y}_C}$ after the treatment can be expressed as:

$$\frac{\overline{y}_T}{\overline{y}_C} \cong \left(\frac{\overline{TC}_T}{\overline{TC}_C}\right) \cdot \left(\frac{F(\overline{x}_T)}{F(\overline{x}_C)}\right) \cdot \left(\frac{\overline{TEC}_T}{\overline{TEC}_C}\right) \tag{10.22}$$

4 MONTE CARLO SIMULATIONS

To illustrate the theoretical ideas discussed earlier we use synthetic data generated in a Monte Carlo experiment. First, we use a data generation process (DGP) to create a baseline dataset with no intervention. Second, we simulate the four alternative educational treatments illustrated in Section 3 with different intensities of the treatment and we measure the impact on a set of treated schools compared with the control schools.

4.1 Data Generation Process and Experimental Design

To emulate the educational production technology of schools we assume a Cobb-Douglas production function in a single output setting with three inputs and constant returns to scale:

$$y = A.x_1^{\beta_1}.x_2^{\beta_2}.x_3^{\beta_3}.u.v \tag{10.23}$$

where y represents the output, and x_1, x_2 and x_3 are the observed inputs, A measures the Hicks-neutral technological change and u represents the efficiency level. The three inputs are randomly and independently drawn from a uniform distribution $U[5, 50]$ with weights $\beta_1 = 0.4$ and $\beta_2 = \beta_3 = 0.3$ defining the contribution of each input to produce the output. This production function exhibits constant returns to scale because the elasticity of scale, the sum of output elasticities, is equal to one. In the baseline scenario, there is no technological change, then $A = 1$. To compute the efficiency component u, we generate a random term θ assumed to be independently distributed from a half-normal distribution $|N(\varphi; \sigma^2)|$ where in our case $|N(0; 0.30)|$ and $u = e^{-\theta}$. It is also assumed that around 15 per cent of the schools belong to the production frontier, that is, $u = 1$, so these DMUs are defined as fully efficient. To do this, a Bernoulli distribution $B(p = 0.15)$ is used to decide which schools in the sample are defined as fully efficient. The generated average efficiency level in each experiment ranges from 0.779 to 0.864 with a standard deviation of from 0.124 to 0.166, respectively. To represent a more realistic set, we simulate a small two-sided random statistical perturbation ν drawn from a normal distribution $N(0; \sigma_\varepsilon^2)$ to account for statistical noise. In our case ε distributes $N(0; 0.025)$ and $\nu = e^{\varepsilon}$. Finally, we compute the observed educational output using Equation (10.21) for a set of $N = 200$ DMUs.

We now assume that a public agency runs a randomised experimental trial where half of the 200 schools are randomly selected to receive the treatment. To do this we randomly assign half of schools to the treated group $N_T = 100$, the rest of schools being the control group $N_C = 100$. We simulate the four scenarios illustrated in Section 3.

In the first scenario $S1$ (*a treatment that changes the input level*), we assume that the intervention plan to produce a change in the output through a change in input x_1. We simulate five intensities of the increment of input x_{1T} in the treated schools: 10 per cent, 20 per cent, 30 per cent, 50 per cent and 75 per cent.

In the second scenario $S2$ (*a treatment that improves the efficiency level*), we assume that the intervention produces a positive efficiency change in the treated schools. We simulate three levels of improvement in efficiency u'_T in the treated schools through reducing the variance of the θ' term: $\sigma^2_{\theta' T} = 0.25$; $\sigma^2_{\theta' T} = 0.20$ and $\sigma^2_{\theta' T} = 0.15$. These reductions in the variance θ' are translated into an increment in u_T of 2.5 per cent, 5.4 per cent and 8.2 per

cent respectively. The new average efficiency in the treated group $\bar{u}'_T$ in each simulated scenario now being 0.848, 0.873 and 0.897 respectively. We generate the new efficiency after the treatment $\bar{u}'_T$ assuming a correlation coefficient of 0.80[13] with the efficiency before the treatment $\bar{u}_T$, and maintaining unchanged the 15 per cent of the schools which belonged to the production frontier. It is worth to note here that although the new after treatment efficiency is randomly generated with the new efficiency distribution it is necessary to maintain certain degree of correlation with the initial inefficiency term to observe the efficiency change.

In the third scenario *S*3 (*a treatment that brings about a technological change*), we assume that the intervention produces a positive change in the treated schools' technology. Thus, we simulate four levels of a positive Hicks-neutral technological change A'_T to the treated schools: $A'_T = 1.01$; $A'_T = 1.025$; $A'_T = 1.05$; and $A'_T = 1.10$.

In the last scenario *S*4 (*a treatment that brings about a technological change and efficiency change*), we assume that the intervention will not produce a change in inputs, but it may (or indeed may not) change the level of the output depending on the intensity of the intervention over the efficiency and the technological change. Thus, we simulate three combinations of improvements in the level of efficiency $\bar{u}'_T$ and a positive Hicks-neutral technological change A'_T to the treated schools: $\sigma^2_{\theta'T} = 0.25$ & $A'_T = 1.05$; $\sigma^2_{\theta'T} = 0.20$ & $A'_T = 1.025$; and finally, $\sigma^2_{\theta'T} = 0.15$ & $A'_T = 1.025$.

Summarising, 15 scenarios were simulated using a Monte Carlo experiment.[14] In each replication, we aim to evaluate if the intervention has significant impacts on the output y, the input x_1, the efficiency u and the technology A of the treated group compared with the control group.

4.2 Results

Table 10.2 summarises the average of the means and standard deviations obtained in each scenario after the 100 simulated replications for the simulated input x_1, the simulated efficiency u, the output y, and the estimated By-Group Malmquist Index (*BGMI*) and its decomposition into technical efficiency change (*TEC*) and technological change (*TC*). After each loop in every scenario we computed the mean t-test differences for the treated and control groups after the treatment in the aforementioned variables. Table 10.3 provides the rejection rate for the 100 means t-test differences run after the Monte Carlo simulations in each scenario. For example, a rejection rate of 0.30 means that in 30 out of the 100 replications the mean t-test difference for the considered variable between the treated and the control group was statistically significant at 99 per cent.

As in a real empirical estimation we only have one database, MC replications can be interpreted as a robust bootstrap set of samples to build confidence intervals. For each replication in the MC simulation we calculate the average value of the variable of interest. Then we compute the confidence intervals at 98 per cent discarding the lowest and highest 1 per cent values for each variable distribution. In an empirical application, we

[13] To generate the new efficiency variable with the desired correlation coefficient with the efficiency before the treatment we follow the procedure used in Cordero et al. (2015) to generate two correlated variables.

[14] Each experiment was replicated 100 times in MATLAB R2013b using the DEA Toolbox developed by Álvarez et al. (2016) available at http://www.deatoolbox.com/.

Table 10.2 Main descriptive of simulated input, efficiency and output variables by groups

	x_1					u					$y = A.\ F(x).u$				
	T		C		Diff.	T		C		Diff.	T		C		Diff.
	Mean	S.D.	Mean	S.D.	T – C	Mean	S.D.	Mean	S.D.	T – C	Mean	S.D.	Mean	S.D.	T – C
Baseline	27.59	13.00	27.43	12.95	0.159	0.828	0.144	0.828	0.143	0.000	20.90	7.80	20.76	7.65	0.139
S1 10%	30.16	14.33	27.51	12.93	2.653	0.830	0.142	0.830	0.141	0.000	21.67	7.93	20.87	7.69	0.802
20%	32.66	15.61	27.64	12.89	5.021	0.827	0.142	0.829	0.142	−0.002	22.22	8.15	21.02	7.71	1.193
30%	35.76	16.98	27.62	12.80	8.136	0.830	0.142	0.829	0.142	0.001	23.17	8.66	20.91	7.62	2.267
50%	41.40	19.48	27.52	13.06	13.884	0.830	0.141	0.829	0.142	0.000	24.62	9.07	20.83	7.67	3.790
75%	48.51	22.74	27.49	13.05	21.017	0.830	0.140	0.830	0.142	−0.001	26.11	9.61	20.82	7.70	5.285
S2 σ^2=0.25	27.51	13.02	27.44	13.03	0.073	0.850	0.123	0.829	0.142	0.021	21.35	7.60	20.73	7.59	0.621
σ^2=0.20	27.60	13.05	27.49	12.97	0.117	0.876	0.104	0.831	0.141	0.045	21.98	7.68	20.82	7.65	1.157
σ^2=0.15	27.50	12.90	27.74	13.05	−0.236	0.898	0.084	0.831	0.141	0.067	22.65	7.70	21.00	7.73	1.646
S3 AT = 1.01	27.59	12.91	27.60	13.12	−0.003	0.828	0.142	0.831	0.141	−0.004	21.13	7.78	21.00	7.74	0.130
AT = 1.025	27.19	12.86	27.49	13.03	−0.298	0.827	0.143	0.829	0.142	−0.002	21.26	7.84	20.82	7.59	0.437
AT=1.05	27.45	12.95	27.45	13.05	0.004	0.832	0.141	0.827	0.141	0.005	21.92	8.00	20.76	7.69	1.162
AT =1.1	27.56	12.96	27.51	12.99	0.044	0.830	0.140	0.829	0.142	0.002	22.95	8.38	20.75	7.61	2.195
S4 σ^2=0.25 & AT = 1.05	27.50	12.95	27.39	12.99	0.111	0.857	0.125	0.831	0.142	0.026	22.77	8.04	20.80	7.60	1.966
σ^2=0.20 & AT = 1.025	27.40	12.99	27.65	12.91	−0.244	0.881	0.106	0.831	0.142	0.051	22.70	7.82	20.87	7.69	1.837
σ^2=0.15 & AT = 1.025	27.71	13.04	27.56	13.02	0.148	0.910	0.081	0.829	0.141	0.080	23.45	7.82	20.93	7.76	2.519

Table 10.2 (continued)

	BGMI					*TEC*					*TC*				
	T		C		Diff.	T		C		Diff.	T		C		Diff.
	Mean	S.D.	Mean	S.D.	T – C	Mean	S.D.	Mean	S.D.	T – C	Mean	S.D.	Mean	S.D.	T – C
Baseline	1.001	0.035	1.001	0.036	0.000	1.021	0.046	1.020	0.045	0.001	0.982	0.033	0.982	0.032	0.000
S1 10%	1.000	0.036	1.001	0.036	0.000	1.019	0.045	1.022	0.047	−0.002	0.983	0.032	0.981	0.033	0.002
20%	1.000	0.040	1.001	0.035	−0.001	1.021	0.046	1.020	0.046	0.002	0.980	0.035	0.983	0.033	−0.003
30%	1.000	0.045	1.001	0.035	−0.001	1.019	0.046	1.021	0.047	−0.001	0.982	0.041	0.982	0.033	0.000
50%	1.000	0.056	1.000	0.036	0.000	1.021	0.046	1.020	0.045	0.001	0.980	0.049	0.982	0.031	−0.001
75%	1.000	0.069	1.000	0.035	−0.001	1.021	0.046	1.020	0.045	0.001	0.980	0.064	0.981	0.032	−0.002
S2 σ^2=0.25	1.037	0.121	1.001	0.035	0.037	1.058	0.127	1.021	0.047	0.037	0.982	0.032	0.981	0.033	0.000
σ^2=0.20	1.074	0.132	1.001	0.036	0.073	1.091	0.138	1.020	0.047	0.071	0.985	0.030	0.982	0.033	0.003
σ^2=0.15	1.109	0.158	1.000	0.035	0.108	1.125	0.165	1.018	0.047	0.106	0.987	0.029	0.983	0.034	0.004
S3 AT = 1.01	1.010	0.036	1.000	0.035	0.010	1.020	0.046	1.019	0.046	0.001	0.991	0.032	0.983	0.032	0.009
AT = 1.025	1.027	0.036	1.000	0.035	0.027	1.020	0.047	1.020	0.045	0.001	1.007	0.034	0.982	0.031	0.026
AT=1.05	1.051	0.037	1.001	0.035	0.049	1.018	0.046	1.022	0.046	−0.004	1.034	0.034	0.981	0.032	0.052
AT =1.1	1.101	0.039	1.001	0.035	0.100	1.021	0.046	1.021	0.047	−0.001	1.079	0.036	0.981	0.033	0.099
S4 σ^2=0.25 & AT = 1.05	1.097	0.127	1.001	0.035	0.096	1.057	0.125	1.020	0.044	0.037	1.039	0.038	0.982	0.031	0.057
σ^2=0.20 & AT = 1.025	1.109	0.139	1.001	0.036	0.108	1.090	0.140	1.020	0.045	0.069	1.019	0.035	0.982	0.032	0.037
σ^2=0.15 & AT = 1.025	1.147	0.159	1.001	0.035	0.146	1.124	0.160	1.021	0.046	0.103	1.022	0.034	0.981	0.032	0.040

Note: Mean values after 100 replications.

Table 10.3 Rejection rate of the mean differences t-test for relevant variables

Treatment / variables		x_1	u	y	*BGMI*	*TEC*	*TC*
Baseline		0.04	0.00	0.03	0.00	0.27	0.30
S1	10%	0.11	0.01	0.04	0.00	0.24	0.29
	20%	0.43	0.01	0.05	0.01	0.27	0.30
	30%	0.89	0.03	0.27	0.02	0.25	0.30
	50%	1.00	0.00	0.70	0.05	0.16	0.30
	75%	1.00	0.00	0.96	0.03	0.12	0.22
S2	σ=0.25	0.01	0.05	0.02	0.63	0.54	0.29
	σ=0.20	0.00	0.49	0.07	0.99	0.91	0.30
	σ=0.15	0.02	0.98	0.15	1.00	1.00	0.30
S3	AT = 1.01	0.02	0.01	0.00	0.25	0.21	0.51
	AT = 1.025	0.00	0.01	0.00	1.00	0.29	0.83
	AT=1.05	0.01	0.00	0.04	1.00	0.23	1.00
	AT =1.1	0.00	0.00	0.24	1.00	0.25	1.00
S4	σ=0.25 & AT = 1.05	0.00	0.14	0.21	0.99	0.55	1.00
	σ=0.20 & AT = 1.025	0.02	0.63	0.21	1.00	0.91	0.97
	σ=0.15 & AT = 1.025	0.01	0.99	0.41	1.00	0.99	0.99

Note: The rejection rate indicates the proportion of times that the null hypothesis (equal means between both groups) has been rejected in 100 replications. In grey rejection rates higher than 30 per cent.

propose to follow the same strategy by drawing a bootstrap sample with replacement from the original one (Simar and Wilson, 1999). Then, we test whether the confidence intervals built for the treated and the control groups overlap. If they overlap, we cannot reject the null mean equality differences for the relevant variable. Table 10.4 reports the results from the confidence interval overlap analysis after the MC, where 'Non-Reject' (NR hereafter) denotes that after the replications the 98 per cent confidence intervals of both groups overlap. Alternatively, 'Reject' (R hereafter) denotes that the intervals do not overlap at 98 per cent and consequently, we reject that the means of both groups are equal. In this last case, we consider that the treatment had a positive impact on the treated schools.

From Tables 10.2 to 10.4, we first verify that after the randomised trial but before the intervention (baseline scenario), the mean values of all variables are not significantly different between the treated and control groups. The only exception appears in Table 10.3 for TEC and TC components where around 30 per cent of replications were statistically significantly different in the baseline scenario. This result is due to the deterministic nature of the non-parametric frontier estimations in finite sample sizes; however, building confidence intervals is a way of correcting this spurious difference as shown in Table 10.4 in which the differences are not significant.

Second, it is worth noting that results obtained in Tables 10.3 and 10.4 correspond to that expected. In scenario 1, we only find significant differences in mean outputs once that the increase in inputs is equal to or higher than 30 per cent. An input increase does not produce a TFPC because in this scenario the technology and the efficiency components remain unchanged. In scenarios 2 and 3 we corroborate that a moderate higher efficiency and a neutral technology progress in the treated group leads to a positive TFPC explained

Table 10.4 Confidence intervals overlaps from MC distributions

Treatment / variables		x_1	y	*BGMI*	*TEC*	*TC*
Baseline		NR	NR	NR	NR	NR
S1	10%	NR	NR	NR	NR	NR
	20%	R	NR	NR	NR	NR
	30%	R	R	NR	NR	NR
	50%	R	R	NR	NR	NR
	75%	R	R	NR	NR	NR
S2	σ^2=0.25	NR	NR	NR	NR	NR
	σ^2=0.20	NR	NR	R	R	NR
	σ^2=0.15	NR	NR	R	R	NR
S3	AT = 1.01	NR	NR	NR	NR	NR
	AT = 1.025	NR	NR	R	NR	R
	AT=1.05	NR	NR	R	NR	R
	AT =1.1	NR	NR	R	NR	R
S4	σ^2=0.25 & AT = 1.05	NR	NR	R	NR	R
	σ^2=0.20 & AT = 1.025	NR	NR	R	R	R
	σ^2=0.15 & AT = 1.025	NR	R	R	R	R

Note: NR = indicates that the null hypothesis (equal mean between treated and control groups) is not rejected at 98 per cent of confidence, i.e. the central 98% of the distributions of the means values in each group overlap. R = indicates that the null hypothesis (equal mean between treated and control groups) is rejected at 98 per cent of confidence, i.e. confidence interval for the means do not overlap. In grey when null hypothesis is rejected.

by the efficiency and technological components respectively. Finally, the scenario 4 results show that the By-Group Malmquist Index can disentangle the channels used by the programme to improve TFP in the treated group correctly.

Third, it is remarkable that when an input change does not occur, the treatment effect is measured better as regards TFPC than through output changes in the treated and control groups. Remember that TFPC can measure the efficiency and technology changes determined by best practices detected through the production frontier instead of using a simple difference in average results. Although in the case of input movements the average output difference is the only way of detecting improvements in, for other policies and programmes it is clear that improvements in technology can be hidden by random noise and/or inefficiency if we only measure the treatment effect be estimating the mean outputs differences between the treated and the control groups. For example, in all simulations included in scenarios 2 and 3 we do not observe significant differences in average outputs but we are able to find these differences in terms of efficiency and technology respectively. The conclusion is that the programme has carried out a TFPC in the treated group with respect to the control one and now we need to learn from best practices to transfer their behaviour to less productive schools.

Finally, Table 10.5 reports the decomposition of the ratio between the estimated average output in the treated and control groups $\frac{\bar{y}_T}{\bar{y}_C}$ into the components of the production function (Equation 10.15). We provide the true simulated ratios (Equation 10.17) and the estimated ratios for each component (Equation 10.22) respectively. These results

Table 10.5 Average output and components of the production function variation between treated and control groups

Treatment / variables		Simulated variables					Estimated variables			
		$\frac{\bar{y}_T}{\bar{y}_C}$	$\frac{\overline{TFPC_T}}{\overline{TFPC_C}}$	$\frac{\bar{u}_T}{\bar{u}_C}$	$\frac{A_T}{A_C}$	$\frac{F(\bar{x}_T)}{F(\bar{x}_C)}$	$\frac{\overline{BGMI_T}}{\overline{BGMI_C}}$	$\frac{\overline{TEC_T}}{\overline{TEC_C}}$	$\frac{\overline{TC_T}}{\overline{TC_C}}$	$\frac{F(\bar{x}_T)}{F(\bar{x}_C)}$
Baseline		1.007	1.000	1.000	1.000	1.007	1.000	1.001	1.000	1.007
S1	10%	1.038	1.000	1.000	1.000	1.038	1.000	0.998	1.002	1.039
	20%	1.057	0.998	0.998	1.000	1.059	0.999	1.001	0.997	1.058
	30%	1.108	1.001	1.001	1.000	1.107	0.999	0.999	1.000	1.109
	50%	1.182	1.000	1.000	1.000	1.181	1.000	1.001	0.999	1.182
	75%	1.254	0.999	0.999	1.000	1.255	0.999	1.001	0.998	1.255
S2	$\sigma^2 = 0.25$	1.030	1.025	1.025	1.000	1.005	1.037	1.036	1.000	0.993
	$\sigma^2 = 0.20$	1.056	1.054	1.054	1.000	1.001	1.073	1.070	1.003	0.984
	$\sigma^2 = 0.15$	1.078	1.081	1.081	1.000	0.998	1.108	1.104	1.004	0.973
S3	AT = 1.01	1.006	1.005	0.995	1.010	1.001	1.010	1.001	1.009	0.996
	AT = 1.025	1.021	1.022	0.997	1.025	0.999	1.027	1.001	1.026	0.994
	AT=1.05	1.056	1.056	1.006	1.050	1.000	1.049	0.996	1.053	1.006
	AT =1.1	1.106	1.102	1.002	1.100	1.003	1.100	0.999	1.101	1.00
S4	$\sigma^2 = 0.25$ & AT = 1.05	1.095	1.083	1.032	1.050	1.010	1.096	1.036	1.058	0.99
	$\sigma^2 = 0.20$ & AT = 1.025	1.088	1.088	1.061	1.025	1.000	1.108	1.068	1.037	0.982
	$\sigma^2 = 0.15$ & AT = 1.025	1.120	1.124	1.097	1.025	0.996	1.146	1.101	1.041	0.978

allow us to account for the ability of the proposed BGMI method to correctly estimate the magnitude of the treatment's impacts on productivity, efficiency and technology.

Table 10.5 shows that not only is the proposed approach able to detect significant impacts of the treatment on total factor productivity, technical efficiency and technology correctly, but also the magnitude of these estimated effects is also accurately estimated. For example, under the treatment of a positive technological change of 5 per cent (S3; $AT = 1.05$) the estimated ratio between the estimated technological change of the treated and controls school $\frac{\overline{TC_T}}{\overline{TC_C}}$ is 5.3 per cent. Additionally, the decomposition computed in Table 10.5 allows us to account for the contribution of each component on the production function to the average output improvement in treated schools. For example, in the last simulated scenario (S4, $\sigma^2 = 0.15$ & $AT = 1.05$) in which the average output of the treated schools is 12 per cent greater than in control schools, improvements in treated schools' efficiency rises to 10.1 per cent and the estimated positive technological change in the treated schools is estimated as 4.1 per cent. In other words, improvements estimated in technical efficiency accounts for almost 85 per cent of the positive impact found on the output.

4.3 Discussion

As it was showed throughout this section production frontier analysis can notably complement and reinforce impact evaluation and vice versa. On one hand, the traditional causal

inference analysis evaluates the impacts of an educational treatment conducted on schools on the average. However, as we have discussed in the previous sections, average effects may be highly influenced by inefficient behaviours and do not allow to observe the best performers. The novelty of the proposed approach is to assume that the treatment effects that could be unobserved on the mean output, can detected by measuring TFPC through BGMI using production frontiers that allow decomposing efficiency and technological changes between the treated and the control groups. Moreover, the production frontier framework allows to easily deal with multiple inputs and multiple outputs which allows to evaluate the treatment effects over all considered outputs at the same time. For the sake of simplicity, in our simulation study we have only considered just one output but it will be worth in future research to extend the impact evaluation through production frontiers gathering all outputs together. On the other hand, to date, production frontiers were estimated for evaluating school efficiency without considering the endogeneity problem (Cordero et al., 2015). Obviously, it will be necessary more research to incorporate this issue in standard efficiency analysis to benchmark schools.

The main conclusion to be noted from our simulations is the large applicability of the approach proposed in this chapter. First, when educational interventions do not translate into significant increments in the mean observed outputs but they actually do in terms of schools' productivity. Second, when educational treatments do significantly impact on the average observed output level, our approach results notably helpful for disentangling these impacts between improvements in inputs or in total factor productivity changes.

Evaluating educational treatments impacts in terms of productivity changes beyond the effects on average outputs seems to be of great relevance in most simulated scenarios where the impacts on output are not significant.[15] In these contexts, if we only evaluate the treatments in terms of the average output improvements we will probably arrive at inaccurate conclusions about the effectiveness of the interventions. But, if we also measure the impacts of the treatments in terms of TFPC, that is, efficiency and technology, we can find significant impacts. In this sense, scenarios 2 and 3 reveal that a hypothetical treatment that only affected efficiency levels or provoked a technology shift would be only detected by using production frontiers. An illustrative intervention could be a programme that promote the implementation of new innovative teaching methods inside the classrooms. It seems likely that inefficiency may arise at the beginning of a new treatment because some schools could be reluctant to apply new procedures. However, some schools can take advantage of the intervention and rapidly incorporate the new education policy to boost students' results. If this were the case, an evaluation on the average could mask the effectiveness of the treatment for some schools in terms of increasing efficiency or improving the technology. This would allow policymakers to learn from the best practices giving an opportunity to enhance the programme when applied to the rest of schools.

Furthermore, even when the treatment significantly impacts on the average output, it is crucial for policymakers to know whether these impacts are driven only by increasing input allocation or by their better management, that is, through schools' efficiency improvements or an overall technological progress in the educational sector. This is

[15] Most of the simulated average impacts on outputs represent less than 0.25 standard deviations which are relatively moderate effects in the context of educational interventions to be found as significant impacts.

evident if we compare two treatments with similar impacts in terms of the average output. For example, by comparing an intervention that only increases the input by around 30 per cent (S1; 30 per cent) with a treatment that increases total factor productivity by 12 per cent (S4; $\sigma = 0.15$ & $AT = 1.05$). When we estimate the effect in terms of average outputs, a priori both scenarios show similar impacts, but of course, the causes behind these improvements are vastly different with very different consequences in terms of budget and performance-based policy recommendations.

5 CONCLUSIONS

At present, the impact evaluation literature provides the most robust available approach for evaluating and enhancing public policies in education. This is because it allows the causal relationship between the intervention and the changes in educational outcomes of individuals receiving the treatment to be identified with respect an appropriately selected counterfactual group of people not participating in the programme. On the other hand, production frontier research allows the production activity of a set of schools to be evaluated by estimating their technical efficiency level and pointing out best practices. Although both fields of research provide complementary information to evaluate programmes and policies comprehensively, to date the two methods have not been linked.

In this chapter, we give an overview of the basics of impact evaluation on education together with a new framework based on production frontiers to analyse the impact of public programmes implemented on schools through a randomised trial. This strategy proposes to measure TFPC (due to efficiency and/or technology changes), through the estimation of a By-Group Malmquist Index. This methodology is illustrated in a Monte Carlo experiment demonstrating that the proposed approach accurately identifies the impacts of the treatments simulated in all scenarios.

From the analysis, we find that in those scenarios where the treatment consists of providing more input to schools, the average output difference is an accurate way of detecting output improvements. However, for policies and programmes devoted to enhancing schools' productivity, technology improvements can be hidden if we only calculate mean output differences between the treated and the control groups. In these cases, the treatment effects are better measured regarding TFPC because it allows us to measure efficiency and technology changes determined by best practices detected through the production frontier. If we only evaluate the treatments in terms of the average outputs improvement, we might not find significant impacts concluding that the intervention had no effect and leading to imprecisely evidence-based policy recommendations. Even when we find significant impacts on the average output, it is crucial for policymakers to reveal the channels through which the treatment operates (additional inputs, schools' efficiency improvements and/or an overall technological progress in the educational sector).

In summary, this chapter highlights the potential of combining a production frontiers framework with the traditional impact evaluation approach to enhance impact evaluations in education. Many lines of research can be addressed in the near future. First, it is necessary to run new Monte Carlo experiments with alternative data generation processes to confirm the robustness of the results found in this chapter. Additionally, alternative treatments should be simulated to test the usefulness of the proposed approach in a

wider context – for example, when we only take advantage of the treatment of the most inefficient schools. Finally, it would be a fruitful contribution to develop a theoretical framework to carry out impact evaluations using production frontiers in the absence of randomised trials, that is, when only quasi-experimental data are available by relating standard quasi-experimental approaches (DiD, IV, RDD, PSM) with production frontiers. First, it seems straightforward to combine this frame with sharp RDD and DiD methods. For instance, as De Witte and López-Torres (2015) suggest, to relate the DiD technique to a metafrontier framework. An alternative approach to run DiD through production frontiers could be to use the Aparicio et al. (2016) methodology, which allows technology and/or efficiency between the treated and control groups to be controlled for differences in input before implementing the educational programme.

REFERENCES

Álvarez, I., J. Barbero and J.L. Zofio (2016). 'A data envelopment analysis toolbox for Matlab'. Universidad Autónoma de Madrid (Spain), Department of Economic Analysis (Economic Theory and Economic History).

Amsler, C., A. Prokhorov and P. Schmidt (2016). 'Endogeneity in stochastic frontier models'. *Journal of Econometrics,* **190**(2), 280–8.

Anghel, B., A. Cabrales and J.M. Carro (2015). 'Evaluating a bilingual education programme in Spain: the impact beyond foreign language learning'. *Economic Inquiry,* **54**(2), 1202–23.

Angrist, J.D. (2004). 'American education research changes tack'. *Oxford Review of Economic Policy,* **20**(2), 198–212.

Angrist, J.D. and A.B. Keueger (1991). 'Does compulsory school attendance affect schooling and earnings?' *The Quarterly Journal of Economics,* **106**(4), 979–1014.

Angrist, J.D. and V. Lavy (1999). 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement'. *The Quarterly Journal of Economics,* **114**(2), 533–75.

Angrist, J.D. and V. Lavy (2002). 'New evidence on classroom computers and pupil learning'. *The Economic Journal,* **112**(482), 735–65.

Angrist, J.D. and J.S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Angrist, J.D. and J.-S. Pischke (2014). *Mastering 'Metrics: The Path from Cause to Effect.* Princeton: Princeton University Press.

Angrist, J.D., S.R. Cohodes, S.M. Dynarski, P.A. Pathak and C.R. Walters (2016). 'Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice'. *Journal of Labor Economics,* **34**(2), 275–318.

Aparicio, J., E. Crespo-Cebada, F. Pedraja-Chaparro and D. Santín (2016). 'Comparing school ownership performance using a pseudo-panel database: A Malmquist-type index approach'. *European Journal of Operational Research,* **256**, 533–42.

Banerjee, A.V., S. Cole, E. Duflo and L. Linden (2007). 'Remedying education: Evidence from two randomized experiments in India'. *The Quarterly Journal of Economics,* **122**(3), 1235–64.

Banker, R.D., A. Charnes and W.W. Cooper (1984). 'Some models for estimating technical and scale inefficiencies in data envelopment analysis'. *Management Science,* **30**(9), 1078–92.

Bellei, C. (2009). 'Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile'. *Economics of Education Review,* **28**(5), 629–40.

Bifulco, R. and S. Bretschneider (2001). 'Estimating school efficiency: A comparison of methods using simulated data'. *Economics of Education Review,* **20**(5), 417–29.

Bifulco, R. and S. Bretschneider (2003). 'Response to comment on estimating school efficiency'. *Economics of Education Review,* **22**(6), 635–8.

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

Caves, D.W., L.R. Christensen and W.E. Diewert (1982). 'The economic theory of index numbers and the measurement of input, output, and productivity'. *Econometrica: Journal of the Econometric Society*, 1393–414.

Cazals, C., F. Fève, J.-P. Florens and L. Simar (2016). 'Nonparametric instrumental variables estimation for efficiency frontier'. *Journal of Econometrics,* **190**(2), 349–59.

Charnes, A., W.W. Cooper and E. Rhodes (1978). 'Measuring the efficiency of decision making units'. *European Journal of Operational Research,* **2**(6), 429–44.
Chetty, R., J.N. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach and D. Yagan (2011). 'How does your kindergarten classroom affect your earnings? Evidence from Project Star'. *Quarterly Journal of Economics,* **126**(4).
Cordero, J.M., D. Santín and G. Sicilia (2015). 'Testing the accuracy of DEA estimates under endogeneity through a Monte Carlo simulation'. *European Journal of Operational Research,* **44**(2), 511–18.
Crawford, C., L. Dearden and E. Greaves (2014). 'The drivers of month-of-birth differences in children's cognitive and non-cognitive skills'. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* **177**(4), 829–60.
Crespo-Cebada, E., F. Pedraja-Chaparro and D. Santín (2014). 'Does school ownership matter? An unbiased efficiency comparison for regions of Spain'. *Journal of Productivity Analysis,* **41**(1), 153–72.
De Witte, K. and L. López-Torres (2015). 'Efficiency in education: a review of literature and a way forward'. *Journal of the Operational Research Society*, **68**(4), 339–63.
Dee, T.S. and B.J. Keys (2004). 'Does merit pay reward good teachers? Evidence from a randomized experiment'. *Journal of Policy Analysis and Management,* **23**(3), 471–88.
Dee, T.S. and J. Wyckoff (2015). 'Incentives, selection, and teacher performance: Evidence from IMPACT'. *Journal of Policy Analysis and Management,* **34**(2), 267–97.
Duflo, E., P. Dupas and M. Kremer (2015). 'School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools'. *Journal of Public Economics,* **123**, 92–110.
Duflo, E., R. Hanna and S.P. Ryan (2012). 'Incentives work: Getting teachers to come to school'. *The American Economic Review*, **102**(4) 1241–78.
Dynarski, S., J. Hyman and D.W. Schanzenbach (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion'. *Journal of Policy Analysis and Management,* **32**(4), 692–717.
Färe, R. and D. Primont (1995). *Multi-output production and duality: theory and applications*. Dordrecht, Netherlands: Kluwer Academic Publishers.
Färe, R., S. Grosskopf, M. Norris and Z. Zhang (1994). 'Productivity growth, technical progress, and efficiency change in industrialized countries'. *American Economic Review*, **84**(1), 66–83.
Felfe, C., N. Nollenberger and N. Rodríguez-Planas (2015). 'Can't buy mommy's love? universal childcare and children's long-term cognitive development'. *Journal of Population Economics,* **28**(2), 393–422.
Fryer, R.G. (2010). 'Financial incentives and student achievement: Evidence from randomized trials'. National Bureau of Economic Research No. w15898.
Fryer Jr, R.G., S.D. Levitt, J. List and S. Sadoff (2012). 'Enhancing the efficacy of teacher incentives through loss aversion: A field experiment'. National Bureau of Economic Research, No. w18237.
Gertler, P.J., S. Martinez, P. Premand, L.B. Rawlings and C.M. Vermeersch (2016). *Impact evaluation in practice*, *second edition*. Washington, DC: IDB and World Bank.
Graves, J. (2011). 'Effects of year-round schooling on disadvantaged students and the distribution of standardized test performance'. *Economics of Education Review,* **30**(6), 1281–305.
Greene, W. (2010). A stochastic frontier model with correction for sample selection. *Journal of Productivity Analysis,* **34**(1), 15–24.
Griffiths, W.E. and G. Hajargasht (2016). 'Some models for stochastic frontiers with endogeneity'. *Journal of Econometrics,* **190**(2), 341–8.
Hahn, J., P. Todd and W. Van der Klaauw (2001). 'Identification and estimation of treatment effects with a regression-discontinuity design'. *Econometrica,* **69**(1), 201–9.
Hanushek, E.A. (1979). 'Conceptual and empirical issues in the estimation of educational production functions'. *Journal of Human Resources*, 351–88.
Heckman J. (2011). 'Integrating personality psychology into economics (No. w17378)'. National Bureau of Economic Research.
Heckman, J. and S. Navarro-Lozano (2004). 'Using matching, instrumental variables, and control functions to estimate economic choice models'. *Review of Economics and Statistics,* **86**(1), 30–57.
Heckman, J.J., S.H. Moon, R. Pinto, P.A. Savelyev and A. Yavitz (2010). 'The rate of return to the HighScope Perry Preschool Programme'. *Journal of Public Economics,* **94**(1), 114–28.
Hoxby, C.M. (1999). 'The productivity of schools and other local public goods producers'. *Journal of Public Economics*, **74**, 1–30.
Hoxby, C. (2000). 'Peer effects in the classroom: Learning from gender and race variation'. National Bureau of Economic Research, No. w7867.
Imbens, G.W. and T. Lemieux (2008). 'Regression discontinuity designs: A guide to practice'. *Journal of Econometrics,* **142**(2), 615–35.
Jacob, B. A. and L. Lefgren (2004). 'Remedial education and student achievement: A regression-discontinuity analysis'. *Review of Economics and Statistics,* **86**(1), 226–44.

Jensen, B., A. Holm and S. Bremberg (2013). 'Effectiveness of a Danish early year preschool programme: A randomized trial'. *International Journal of Educational Research,* **62**, 115–28.
Jensen, P. and A.W. Rasmussen (2011). 'The effect of immigrant concentration in schools on native and immigrant children's reading and math skills'. *Economics of Education Review,* **30**(6), 1503–15.
Johnes, J. (2015). 'Operational research in education'. *European Journal of Operational Research,* **243**(3), 683–96.
Kane, T. J. and D.O. Staiger (2008). 'Estimating teacher impacts on student achievement: An experimental evaluation'. National Bureau of Economic Research, No. w14607.
Kearney, M.S. and P.B. Levine (2015). 'Early childhood education by MOOC: Lessons from Sesame Street'. National Bureau of Economic Research, No. w21229.
Khandker, S.R., G.B. Koolwal and H.A. Samad (2010). *Handbook on Impact Evaluation: Quantitative Methods and Practices*. World Bank Publications.
Leibenstein, H. (1966). 'Allocative efficiency vs. "X-efficiency"'. *The American Economic Review*, **56**(3), 392–415.
Levačić, R. and A. Vignoles (2002). 'Researching the links between school resources and student outcomes in the UK: A review of issues and evidence'. *Education Economics,* **10**(3), 313–31.
Levin, H.M. (1974). 'Measuring efficiency in educational production'. *Public Finance Quarterly,* **2**(1), 3–24.
Ludwig, J. and D.L. Miller (2007). 'Does Head Start improve children's life chances? Evidence from a regression discontinuity design'. *The Quarterly Journal of Economics*, **122**(1), 159–208.
Matsudaira, J.D. (2008). 'Mandatory summer school and student achievement'. *Journal of econometrics,* **142**(2), 829–50.
Mayston, D.J. (2015). 'Analysing the effectiveness of public service producers with endogenous resourcing'. *Journal of Productivity Analysis,* **44**(1), 115–26.
Mayston, D.J. (2016). 'Data envelopment analysis, endogeneity and the quality frontier for public services'. *Annals of Operations Research*, in press.
Orme, C. and P. Smith (1996). 'The potential for endogeneity bias in data envelopment analysis'. *Journal of the Operational Research Society*, **47**(1), 73–83.
Pastor, J.T. and C.K. Lovell (2005). 'A global Malmquist productivity index'. *Economics Letters,* **88**(2), 266–71.
Pedraja-Chaparro, F., D. Santín and R. Simancas (2016). 'The impact of immigrant concentration in schools on grade retention in Spain: a difference-in-differences approach'. *Applied Economics,* **48**(21), 1978–90.
Perelman, S. and D. Santín (2011). 'Measuring educational efficiency at student level with parametric stochastic distance functions: an application to Spanish PISA results'. *Education Economics,* **19**(1), 29–49.
Pischke, J.S. (2007). 'The impact of length of the school year on student performance and earnings: Evidence from the German short school years'. *The Economic Journal,* **117**(523), 1216–42.
Rosenbaum, P.R. and D.B. Rubin (1983). 'The central role of the propensity score in observational studies for causal effects'. *Biometrika,* **70**(1), 41–55.
Ruggiero, J. (2003). 'Comment on estimating school efficiency'. *Economics of Education Review,* **22**(6), 631–4.
Ruggiero, J. (2004). 'Performance evaluation when non-discretionary factors correlate with technical efficiency'. *European Journal of Operational Research,* **159**(1), 250–7.
Santín, D. and G. Sicilia (2014). 'The teacher effect: An efficiency analysis from a natural experiment in Spanish primary schools', paper presented at the Workshop on Efficiency in Education, Lancaster University Management School, 19 September.
Santín, D. and G. Sicilia (2017). 'Dealing with endogeneity in data envelopment analysis applications'. *Expert Systems with Applications*, **68**, 173–84.
Schlotter, M., G. Schwerdt and L. Woessmann (2011). 'Econometric methods for causal evaluation of education policies and practices: a non-technical guide'. *Education Economics,* **19**(2), 109–37.
Schultz, T.P. (2004). 'School subsidies for the poor: evaluating the Mexican Progresa poverty programme'. *Journal of Development Economics,* **74**(1), 199–250.
Shapiro, J. (2011). 'Monitoring and evaluation toolkit'. CIVICUS: World Alliance for Citizen Participation. Available at http://www.civicus.org/view/media/Monitoring%20and%20Evaluation.pdf, accessed 23 July 2017.
Simar, L., A. Vanhems and I. Van Keilegom (2016). 'Unobserved heterogeneity and endogeneity in nonparametric frontier estimation'. *Journal of Econometrics,* **190**(2), 360–73.
Simar, L. and P.W. Wilson (1999). 'Estimating and bootstrapping Malmquist indices'. *European Journal of Operational Research,* **115**(3), 459–71.
Thanassoulis, E., K. Witte, J. Johnes, G. Johnes, G. Karagiannis and C.S. Portela (2016). 'Applications of data envelopment analysis in education'. In J. Zhu (ed.), *Data Envelopment Analysis: A Handbook of Empirical Studies and Applications* (pp. 367–438). Boston, MA: Springer US.
Thistlethwaite, D.L. and D.T. Campbell (1960). 'Regression-discontinuity analysis: An alternative to the ex post facto experiment'. *Journal of Educational Psychology,* **51**(6), 309.
Van der Klaauw, W. (2002). 'Estimating the effect of financial aid offers on college enrolment: A regression–discontinuity approach'. *International Economic Review,* **43**(4), 1249–87.

Van Klaveren, C. and K. De Witte (2014). 'How are teachers teaching? A nonparametric approach'. *Education Economics,* **22**(1), 3–23.
Webbink, D. (2005). 'Causal effects in education'. *Journal of Economic Surveys,* **19**(4), 535–60.
West, M.R. and L. Woessmann (2010). '"Every Catholic child in a Catholic school": Historical resistance to state schooling, contemporary private competition and student achievement across countries'. *The Economic Journal,* **120**(546), F229–F255.
Wooldridge, J. (2012). *Introductory econometrics: A modern approach*. Cengage Learning.
Worthington, A.C. (2001). 'An empirical survey of frontier efficiency measurement techniques in education'. *Education Economics,* **9**(3), 245–68.