Causal Inference on Education Policies: A Survey of Empirical Studies Using PISA, TIMSS and PIRLS

Cordero, José Manuel and Cristobal, Victor and Santín, Daniel

Universidad de Extremadura, Spain, Forschungsinstitut für Bildungs- und Sozialökonomie (FIBS), Germany, Universidad Complutense de Madrid, Spain

17 January 2017

Causal Inference on Education Policies: A Survey of Empirical Studies Using PISA, TIMSS and PIRLS

José M. Cordero[1]*, Víctor Cristóbal[2], Daniel Santín[3]

[1] *University of Extremadura, Spain*
[2] *Forschungsinstitut für Bildungs- und Sozialökonomie (FIBS), Germany*
[3]*Complutense University of Madrid, Spain*

Abstract

The identification of the causal effects of educational policies is the top priority in recent education economics literature. As a result, a shift can be observed in the strategies of empirical studies. They have moved from the use of standard multivariate statistical methods, which identify correlations or associations between variables only, to more complex econometric strategies, which can help to identify causal relationships. However, exogenous variations in databases have to be identified in order to apply causal inference techniques. This is a far from straightforward task. For this reason, this paper provides an extensive and comprehensive overview of the literature using quasi-experimental techniques applied to three well-known international large-scale comparative assessments, such as PISA, PIRLS or TIMSS, over the period 2004-2016. In particular, we review empirical studies employing instrumental variables, regression discontinuity designs, difference in differences and propensity score matching to the above databases. Additionally, we provide a detailed summary of estimation strategies, issues treated and profitability in terms of the quality of publications to encourage further potential evaluations. The paper concludes with some operational recommendations for prospective researchers in the field.

Keywords: Literature review, Education, Causal Inference, Selection-bias, International assessments.

* Corresponding author: jmcordero@unex.es

1. Introduction

Large-scale assessment surveys in the educational research and policy landscape have played a growing role over the last two decades (Gustafsson, 2008; Kamens, 2009). Broadly defined, large-scale assessments are surveys of knowledge, skills, or behaviors in a given domain that provide comparable data about many different educational systems around the world. Researchers can use this information to analyze differences in achievement between and within countries and to investigate the effects of various educational and societal factors on educational achievement, as well as the impact of skills on economic and social outcomes (Creemers and Kyriakides, 2008; Hanushek and Woessman, 2011). Likewise, such international comparisons are particularly useful for evaluating the impact of educational reforms, especially with respect to some specific institutional features for which the variation can only be observed across countries (Strietholt et al., 2014).

Historically, most empirical analyses using these comparative data have been based on regressions in the form of educational production functions that link resource inputs with educational outcomes after controlling for various background features (Hanushek, 1979; Todd and Wolpin, 2003). However, this approach may fail to produce convincing estimates when the treatment, an explanatory variable in the model, is not exogenous due to the well-known endogeneity problem. In education, the main source of endogeneity is self-selection. For example, schools with better academic outcomes tend to attract relatively more motivated parents seeking the best education for their children. When this unobserved heterogeneity is correlated with receiving the treatment, the econometric estimation of the causal effect of this treatment is likely to be biased. Reverse causality is a second major source of endogeneity that arises, for example, when poor test scores for some students or schools lead to the implementation of a reform (treatment) to boost the results. In this case, the direct comparison between treated and untreated schools will be biased because the treatment is correlated with the unobserved reason behind the poor performance of these schools.

Therefore, the estimation of causal effects in the presence of endogeneity often biases results (Webbink, 2005). This limitation has led to the development of more sophisticated techniques that allow valid causal inference based on defining the

counterfactual group through a quasi-experiment on observational data (Morgan and Winship, 2007, Gertler et al., 2016). Such econometric techniques in education economics are mainly represented by instrumental variables, regression discontinuity designs, difference in differences and propensity score matching.

The aim of this paper is to review empirical studies applying such methods to observational data from three well-known large-scale assessments and explain the specific estimation strategies employed by educational researchers with these databases in order to identify the causal impact of different educational policies on outcomes. The databases are the Programme for International Student Assessment (PISA), launched by the Organization of Economic Cooperation and Development (OECD), and the two surveys conducted by the International Association for the Evaluation of Educational Achievement (IEA), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). PISA has tested 15-year-old students in math, science, and reading performance every three years since 2000. TIMSS has assessed the mathematics and science achievements of fourth- and eighth-grade students every four years since 1995, whereas PIRLS focuses on the reading literacy achievement of fourth-grade students, who have been surveyed every five years since 2001.

This survey describes the estimation strategies used by educational researchers and highlights the potential of these databases for analyzing the causal effects of multiple key issues in education policy (class size, instructional time, maturity and so on) on students' results. The aim is to inspire new empirical applications using these databases with insight from the research developed to date. Additionally, we summarize the trends for this line of research regarding different issues and also provide a snapshot of the scientific journals in which the papers were published.

The remainder of the paper is organized as follows. Section 2 discusses the literature review strategy followed to retrieve the analyzed papers. Section 3 briefly explains methodological aspects related to the econometric approaches applied in empirical studies in order to facilitate their interpretation. Section 4 presents the results of the literature review conducted considering four different categories corresponding to the employed econometric approaches and distinguishing several research topics. Section 5

summarizes the contents of the empirical studies surveyed in the previous section, including an overview of the journals in which they were published. Finally, Section 6 concludes.

2. Literature review and search strategy

The literature addressing the econometric techniques available for developing causal inference on impact evaluation problems in depth is vast (Angrist and Pischke, 2008, 2014; Gertler et al., 2016). Likewise, there are also some papers providing helpful guidelines for practitioners interested in implementing causal inference econometric approaches in education economics problems (Webbink, 2005; Schlotter et al., 2011). In addition, Hanushek and Woessman (2014) provide an extensive review of studies using international survey data to analyze different institutional features as part of a cross-country approach. However, they address several papers using traditional econometric methods, such as least squares, whose estimated effects are very unlikely to reveal causal implications.

Taking this literature as a reference, our target here is to review empirical applications for four causal inference techniques: instrumental variables, regression discontinuity designs, difference in differences and propensity score matching on the three best-known international databases: PISA, TIMSS and PIRLS. In order to conduct our search for empirical studies, we used three main search engines: ERIC (Educational Resources Information Center), Scopus and ISI Web of Science (WoS). ERIC is an online digital library of educational research and information and is sponsored by the Institute of Education Sciences of the United States Department of Education. It is the largest educational database worldwide, providing access to about 1,000 scientific journals. It provides a comprehensive, searchable, Internet-based bibliographic and full-text database of education research and information for educators, researchers, and the general public. Scopus is a bibliographic database maintained by Elsevier, which contains abstracts and citations for academic journal articles, books and conference proceedings in many different fields of research. Finally, ISI WoS is the world's leading academic citation indexing database and search service, which is provided by Thomson Reuters. It covers the sciences, social sciences, arts and humanities. Likewise, it provides bibliographic content and tools to access, analyze and manage research

information. Finally, we rounded out our search by consulting other well-known databases like Econlit (American Economic Association), ABI/Inform Global and Google Scholar to add any articles that we possibly missed to our results.

Our literature search was performed from June to October 2016 and was restricted to studies written in English language. We included empirical papers starting from 2004 up to the year 2016. We performed a computerized systematic search using a wide range of search terms or keywords merged into two groups. The first one included terms related to the methodological approach applied (causal inference, identification strategy, exogenous variation, instrumental variables, regression discontinuity, propensity score matching, difference in differences, fixed effects), and the second one was referred to the database employed (PISA, TIMSS, PIRLS, large-scale assessment, cross-country, comparative study, student performance and achievement). Our initial search identified more than 180 papers. After a careful review of their content, however, this number was reduced significantly because some of the studies were not in fact using causal inference methods or employed national databases instead of the three international large-scale assessments considered. The final selection included 66 studies.

The studies can be classified according to different criteria (e.g. chronological order, topic studied or database employed). However, we decided to organize them according to the identification strategy applied to deal with the common problem of endogeneity bias in data since this is the main focus of this paper. Section 3 roughly explains each approach pointing out their main advantages and drawbacks with respect to international databases. Section 4 describes the empirical studies applying each causal inference approach on international databases to evaluate the effects of different educational programs or interventions. The discussion or comparison of the results is beyond the scope of this survey.

In order to facilitate the identification of the main characteristics of each empirical study, similarly to Hanushek and Woessman (2014), we built a table listing their main details (see Table A1 in the Annex). Each record includes the year of publication, the dataset/s employed, type of data (cross-sectional or pooled data), the country/countries studied, the estimation method and an overview of the analyzed research question. From this information, we found that the authors of almost half of the studies adopt a cross-

country approach in order to leverage the often much larger variation existing across countries (Woessman, 2007). Nevertheless, we also came across multiple studies analyzing data about a single nation, especially among European countries.

3. Methods

The estimation of causal effects is now the top priority of current educational research. Both researchers and policy makers are interested in having empirical evidence suitable for guiding decision-making on effective educational policies and practices. The foundations of causal inference derive from the work of Rubin (1974; 2008). Rubin developed the fundamental pillars of the counterfactual theory of causation with respect to the estimation of treatment effects. The basic idea is that, ideally, researchers would like to know what would have happened if an individual exposed to a treatment condition (T) had instead been in the control group (C). With this definition of the potential outcomes, the causal effect (δ) of treatment for individual i is defined as the difference in the outcome (Y) for individual i when he or she receives T versus C, all else being equal: $\delta_i = Y_i^T - Y_i^C$

In practice, we cannot estimate the causal effect because each individual is in either the treatment group or the control group. Thus, we can observe only one of these potential outcomes. This is often referred to as the fundamental problem of causal inference (Holland, 1986). Therefore, causal inference is basically a missing data problem, where at least half of the values of interest (the potential outcomes) are missing (Stuart, 2007). In this context, researchers need to make assumptions in order to approximate what they would have observed if individuals were in the alternative condition (counterfactuals). The gold standard approach for dealing with this problem and estimating the effects of treatments or interventions on outcomes is the randomized control trial (RCT). Randomization guarantees that individuals belonging to the treated and counterfactual groups are equal with respect to all observed and unobserved characteristics except for treatment reception.

In RCT designs participants are randomly assigned to treatment and control groups, ensuring that treatment status will not be confounded with either measured or unmeasured baseline characteristics. Therefore, the effect of treatment on outcomes can

be estimated over time by comparing average outcomes directly between the two groups. Nevertheless, RCTs are often difficult conduct in the education sector because of high implementation costs, ethics or political differences. In such circumstances, researchers are forced to rely on secondary observational data sourced from large-scale assessments (Schneider et al., 2007). Over the past four decades, different statistical procedures have been designed to deal with potential endogeneity when making comparisons between treatment and control groups (e.g., Heckman, 1976, 1979; Rosenbaum, 1986).

Note, at this point, that we do not intend to provide a detailed explanation of the research methods applied in such empirical studies. As mentioned above, descriptions are available in several manuals and handbooks specifically designed for this purpose[1]. However, we do provide a brief non-technical description of the basic ideas underlying each method in order to give interested readers a feeling for each approach. The four quasi-experimental approaches included in this survey are instrumental variables, regression discontinuity designs, difference in differences and propensity score matching.

Instrumental variables (IV)

The so-called IV method is a standard econometric approach applied to overcome omitted variable problems in estimating causal relationships. Only that part of the variation in the predictor that is not related to unobservable factors affecting both predictor and outcome can be used in this technique. It relies on finding an additional variable that is related to the decision rule but not correlated with the outcome. This variable, known as the 'instrument', introduces some randomness into the assignment. This reproduces the effect of an experiment. Such a procedure allows researchers to isolate the exogenous variation in the treatment to get unbiased estimates of the causal relationship between the outcome and the predictor (Schlotter et al., 2011; Pokropek, 2016).

The key issue in the implementation of the IV approach is, therefore, the choice of a valid instrument. In this respect, the researcher has to attempt to find a variable that is correlated with the treatment determining the probability of treatment, but causally

uncorrelated with the dependent variable. This means that it should not be correlated with the error term (Wooldridge, 2010). When a convincing instrument is found, causal effects can be identified with cross-sectional observations. Thus the implementation of this econometric approach is becoming increasingly frequent in empirical studies using data from large-scale international assessments.

In practice, this effect is usually estimated by implementing the two-stage least squares (2SLS) approach proposed by Heckman (1979)[2]. The first stage consists of a regression where the dependent variable is the treatment and the covariates are the IVs and other exogenous variables that are used in the second stage. The inclusion of covariates in this model helps to fulfill the assumption that there is no direct relationship between the instrument and the analyzed outcome. Finally, the second stage estimates a regression replacing the original treatment variable by the treatment prediction estimated in the previous model whilst maintaining the same set of covariates.

Regression Discontinuity Design (RDD)

This approach can be applied in specific settings when the participation in an intervention or treatment changes discontinuously with some continuous or running variable. Thus, the key point of this method is that the probability of participating is determined by a certain cut-off value of a running variable[3]. The basic idea of the method is that the comparison of students or schools within a fairly small range above and below this cut-off point guarantees that the characteristics of both groups are statistically similar, but only some of them receive the treatment. This scenario is very close to an experimental design with random assignment, since we have a control group (below the cut-off) and a treatment group (above the cut-off) that can be compared. In this framework, the jump or discontinuity in outcomes that can be observed at the threshold can then be interpreted as the causal effect of the program.

In most cases, however, the cut-off threshold does not always divide the sample into two groups, since it is sometimes possible to find control and treatment observations below and above the cut-off. In this framework, the usual estimation strategy is a fuzzy regression discontinuity design. This exploits discontinuities in the probability of treatment using the legal cut-off point as the instrumental variable[4]. The most common

problem for implementing the RDD approach using data from international comparative studies is to find enough observations around the cut-off point[5].

Difference in differences (DiD)

The idea behind this approach is simple. We need two groups of individuals or schools observed in two different periods. If one group is exogenously exposed to a treatment or policy shift and the other is not, then the effect of the treatment can be easily measured taking the differences between the average results for the two groups before and after the educational policy is implemented. Subsequently, the impact or causal effect of the treatment is calculated as the difference between those two differences. The main benefit of this approach is that it accounts for changes within units of interest only. This limits the bias caused by unobserved or uncontrolled differences between these units. The key assumption required to identify the effect of the treatment is that the trends in the outcome of interest would be identical in both groups in the absence of treatment.

For this reason, this approach is normally performed with a panel or pseudo-panel database that can be used to test the equal trends hypothesis assuming that any existing heterogeneity is constant over time (McCaffrey et al., 2003). In a panel data framework, we can control for fixed effects. In this manner, we can account for an indicator variable that takes out mean differences between units so that the effect of the evaluated program or policy can be identified by the changes experienced by the other variables over time. Note also that such fixed effects can be introduced in the model at different levels (students, teachers or schools) or even combining some of them in more complex settings (e.g. Rivkin et al., 2005; Clotfelter et al., 2007).

In principle, this approach cannot be implemented when data are retrieved from large-scale international assessments since they do not provide longitudinal information at individual or school level. However, this methodology can be adapted to a single dimension of time when there are at least two observations for the same evaluated unit (e.g. test scores for different subjects or students enrolled in different grades) or, alternatively, when the units have very similar characteristics (e.g. evaluating the impact on twins). Another possibility would be to use several international waves as a pseudo-panel database to account for differences at regional or country level.

Propensity Score Matching (PSM)

Rosenbaum and Rubin (1983) proposed propensity score analysis as a practical tool for reducing selection bias by balancing treatment and control groups with respect to observed covariates. This method is an extension of the non-parametric matching approach. This approach aims to reproduce the treatment group among the non-treated to emulate the experimental conditions in a non-experimental setting with observational data. In order to implement this method, the unobserved variables have to be assumed to be equally distributed in treated and control groups. In other words, the underlying assumption is that the set of observables contains all the information that determined the probability to be treated.

Heckman and Navarro (2004) recommend the selection of variables describing the information available at the time of treatment assignment and simultaneously explaining the outcome of interest. Thus this estimation strategy usually requires access to an extensive dataset. Fortunately, this is not a problem in empirical studies using whose data are sourced from international comparative studies, since most of them include information about multiple aspects that might have influence on educational outcomes. As a result, the implementation of the propensity score matching approach in empirical papers using data from international comparative studies has increased notably in recent years.

PSM is implemented in two stages. In the first stage, the researcher calculates the probability, known as the "propensity score", of each individual receiving the treatment. This reduces the matching problem to a single dimension, thus significantly simplifying the matching procedure (Wilde and Hollister, 2007). The idea behind this estimator is that if two students or schools have the same propensity score but are in different treatment groups, the assignment can be assumed to be random. When using propensity score matching, the comparison group for each treated individual is chosen using a predefined matching criterion of proximity between the propensity scores for treated and controls. Likewise, after defining a neighborhood for each treated observation, it is necessary to select the appropriate weights to associate observations in the treatment and control group and drop treatment and control observations whose propensity score

is greater than the maximum or less than the minimum of the controls. This ensures a common support for all matched observations.

PSM is a non-experimental technique. Thus, although this method can mitigate the problem of self-selection, the assumption of no unobserved differences between the treated and empirically derived control group, essential for the propensity score strategy, is unlikely to hold. For this reason, PSM is probably the worst choice for improving estimations with respect to the use of all untreated individuals as controls as long as unobservable variables correlate with observables, leading to a reduction in the endogeneity bias.

To conclude this section, Table 1 summarizes the main characteristics of these four econometric techniques, as well as their main strengths and weaknesses for their use with international databases.

Table 1: Causal inference methods applied on international educational databases

Approach	Description	Strengths	Weaknesses
Instrumental Variables (IV)	Sometimes nature or the legal framework leads to exogenous sources of variation correlated with the treatment but uncorrelated with the dependent variable.	The method exploits a partial random assignment that reproduces a natural experiment. It provides even more robust results than other methodological approaches.	It is mostly quite difficult to find a good, endogeneity-free instrument for international databases.
Regression Discontinuity Designs (RDD)	Participation is decided by an exogenous cut-off point, normally defined by an education law requirement.	The cut-off point reproduces a random experiment. It is easy to apply and provides robust results. It works well with educational policies based on rules, such as grants, entry criteria, etc.	Results are average local treatment effects in the sense that they could not be generalized for individuals that are far from the cut-off point.
Differences in Differences (DiD)	"Before" and "after" information is required for the treated and the counterfactual groups. The treatment should be exogenous for the treated group.	Once the information is available and the equal trends assumption is verified before applying the treatment, the method is easy to apply and provides robust results.	Data demanding in terms of 'pre' and 'post' periods. It is crucial to demonstrate the equal trends assumption. For international databases, this probably requires the linkage of different waves.
Propensity Score Matching (PSM)	Beneficiaries are matched with control individuals using prior-to-treatment observed covariates. This requires an estimation of the probability of belonging to the treated group for all individuals. Then, the estimated probabilities are used to match pairs of treated	PSM improves causal estimations with respect to using all untreated individuals as a control as long as unobservable variables correlate with observables. Whenever this assumption holds and treated and control individuals have	PSM is a non-experimental approach because there is no randomization in the treatment assignment. It is mostly unreliable to assume that the unobservable variables of students or parents affecting both the treatment and the results

	individuals and control individuals that have a similar probability of being treated but are in the control group.	the same distribution on unobservable variables, PSM mitigates the endogeneity problem.	will be equally distributed in the treated and untreated groups.

4. Empirical studies review

In this section, our goal is to review the empirical studies in which the above methods have been applied to estimate the causal effect of different educational practices or treatments using observational data from PISA, TIMSS or PIRLS or a combination of databases. To organize the results, we classify the surveyed studies according to the estimation strategy applied and the issue covered.

4.1. Instrumental Variables

Exogenous sources of variation are difficult to find. Therefore, this approach requires researcher creativity, the availability of a valid instrument and a profound knowledge of the intervention and the circumstances under which it was developed. The most frequent topics analyzed using this approach are the private-public school debate or the effects of class size, school entry age and immigrant concentration in schools. Nevertheless, there are some studies using this strategy covering other issues.

Public vs. private schools

Vandenberghe and Robin (2004) pioneered the application of the IV approach (compared with other alternative methodologies like PSM) to deal with selection bias in their analysis of the effect of private school attendance on educational achievement using data about different countries participating in PISA 2000. The instrument that they used in their attempt to control for the potential endogeneity of the treatment was the location of the school defined by a dummy whose value is one if the school is located in a big city (more than 100,000 inhabitants) and 0 otherwise. The same instrument was also selected by Pfeffermann and Landsman (2011) in their empirical analysis of private and public schools in Ireland using PISA 2000 data, as well as Perelman and Santin (2011) in their research about Spanish public and private schools participating in PISA 2003. As a novelty, Perelman and Santín (2011) applied this strategy to estimate efficiency measures using parametric stochastic frontier methods.

Cornelisz (2013) again employs a similar instrument to analyze this phenomenon in the Netherlands, although his indicator is sourced from the school principal's response to the question of whether parental endorsement of the instructional or religious philosophy of the school is taken into consideration at the time of admission.

Another potential way of analyzing this issue is to consider whether historical differences lead to persistent differences in the size of the private school sector. First, West and Woessmann (2010) study the relationship between private school competition and student performance in a cross-country setting. They use the share of each country´s Catholic population in 1900 as an instrument for measuring the effect of contemporary private school competition. Similarly, Falck and Woessman (2013) also used the percentage of a country's Catholic population in 1900 in interaction with an indicator that Catholicism was not official state religion in the country as an instrument for explaining the country's share of students attending private schools today. Both studies analyze the effect of that variable on student achievement using PISA data (2003 in the former and 2006 in the latter).

Class size

Another topic of research studied by applying this method is the effect of class size and class composition on student performance using the rule indicating the maximum number of students per classroom established by states or countries. With the aim of identifying size effects (controlling for within school sorting), Jürges and Schneider (2004), Woessmann and West (2006) and West and Woessman exploit available data about 13-year-old students in TIMSS 1995, combining school fixed effects and instrumental variables to identify random variation between two adjacent grades (seven and eight)[6]. The variable used as an instrument for students' actual class size is the average class size at different grade levels according to the questionnaire responses given by school principals. Denny and Oppedisano (2013) analyze this question for the United States and the United Kingdom using PISA 2003 data and also select the average class size at the respective grade level in the school as an instrument. Konstantopoulos and Traylor (2014) and Konstantopoulos and Shen (2016) examine this relationship for public schools in Greece and Cyprus using data from PIRLS 2001 and TIMSS 2003 and 2007, respectively. Their instrument is an index representing the

average class size, which should be independent of unobserved student, teacher, or school variables. Likewise, Li and Konstantopoulos (2016) use the same instrument to estimate class size effects on fourth-grade mathematics achievement in 14 out of the 25 European countries participating in TIMSS 2011, since they selected countries that had known clear rules about maximum limits on class size only.

Age at school entry

The IV approach has also been applied by Bedard and Dhuey (2006) to examine the impact of maturity differences on student performance. Since the relative age evaluated at any point in the educational process is endogenous, they base their estimation strategy on birth date, which is arguably exogenous. To do this, they pool data from different datasets (mainly TIMSS 1995 and TIMSS 1999) and compare the test scores of children with older and younger assigned relative ages at the fourth and eighth grade levels. The estimation strategy relies on using the birth month relative to the school cut-off date as an instrument representing the observed age. Puhami and Weber (2008) also exploit the exogenous variation in month of birth to estimate the effect of age at school entry on educational outcomes using data about German students participating in PIRLS 2011. They adopt an instrumental variable identification strategy in which the instrument for the endogenous age of school entry is the theoretical age of school entry as prescribed by the state institution.

García-Pérez et al. (2014) selected the students' quarter of birth as an instrument to examine the effect of grade retention on academic performance, although they used cross-sectional data about Spanish students participating in PISA 2009 only. Ponzo and Scoppa (2014) also exploit the exogenous variations in the month of birth coupled with the school entry cut-off date to investigate whether the age at school entry affects Italian students´ performance at the fourth, eighth and tenth grade levels using data from PIRLS 2006, TIMSS 2007 and PISA 2009.

Immigrant concentration

Jensen and Rasmussen (2011) adopt an IV estimation strategy to study the effect of immigrant concentration in schools on the educational outcomes of both immigrant and native children in Denmark. The empirical data used in their empirical analysis is a

combination of the Danish subsample of the PISA study from the year 2000 and a special Danish PISA study from 2005 in which there is an oversampling of children from immigrant backgrounds. In order to deal with the potential selection problem deriving from the fact that a school may have a high immigrant concentration because the parents of the immigrant children have decided to settle in a neighborhood with many immigrants, Jensen and Rasmussen use immigrant concentration in a larger geographical area as an instrument in their empirical analysis.

Moreover, Isphording et al. (2016) analyze the causal effect of immigrant students´ reading performance on their math performance using an IV approach in an attempt to overcome endogeneity issues related to the unobserved ability of students. To do this, they pool data from four different PISA waves (2003, 2006, 2009, 2012) and exploit variation in different ages at arrival and linguistic distance between origin and destination country languages. Such variables cannot be used as instruments because both have a direct effect on migrants´ math performance, but the interaction between such variables can be considered as a good identifying variable in order to isolate variation that only affects language performance.

Other topics

Lee and Fish (2010) examine the extent and sources of variation in value-added academic growth patterns in mathematics applying hierarchical linear models with an instrumental variable method. In their empirical analysis they use data about different states in the US and six nations in which there is an established cut-off birth date for student enrollment at school. Specifically, Lee and Fish merge samples from TIMSS 1995 fourth-grade with 1999 eighth-grade math assessment data and samples from the National Assessment of Educational Progress (NAEP) 1996 fourth-grade with 2000 eighth-grade math assessment data. In order to avoid potential problems of endogeneity with some variables (e.g. age and grade), they use the relative age at which children should be observed on the basis of their birth date relative to the school cut-off, as well as the grade in which the students would be expected to be enrolled based on their birth date relative to the school cut-off date as the instruments in their estimations.

Choi et al. (2012) employed the IV approach in a multilevel framework to evaluate the impact of time spent on private tutoring on the performance of Korean students in mathematics and reading using PISA 2006 data. Using this estimation strategy, Choi et al. were able to avoid potential data endogeneity since families whose children are more capable of achieving better results can be assumed to be more willing to invest more in tutoring. The instrument used is the number of hours of private tutoring in science received per week.

Gamboa et al. (2013) analyze the effect of pupils' self-motivation on academic achievement in science in a panel of countries using PISA 2006 data. In order to reduce the potential endogeneity bias, they construct an instrument representing students´ perceptions about the importance of science in their lives and for society based on their responses to a set of specific questions related to this topic included in the questionnaire. In their empirical analysis, they use instrumental variable quantile regression models to evaluate the effect of independent variables on different points of the science score conditional distribution.

Gustafsson (2013) also uses the IV approach to investigate the effects of time spent doing homework on mathematics achievement. Using data from 22 countries participating in TIMSS 2003 and TIMSS 2007, they constructed two different measures of the total number of minutes spent on mathematics homework per week according to the information provided by students and teachers. In their empirical analysis, they used the variable based on teachers´ responses as an instrument for the time reported by students. The IV regressions were conducted separately for each country in the two datasets.

Edwards and Garcia-Marin (2015) examine whether the inclusion of educational rights in political constitutions has an influence on student performance using data from 61 countries participating in PISA 2012. In their empirical analysis, Edwards and Garcia-Marin selected two different instruments: the historical origins of legislation protecting minority investors in a score of countries and the year of independence of each country.

4.2. Regression discontinuity designs

There are very few empirical studies using this estimation strategy on international databases, although we can find several studies covering topics such as the effects of class size, schooling or tracking.

Class size

Woessmann (2005) uses data from TIMSS 1995 to estimate class-size effects by exploiting discontinuities in class size induced by the maximum class size rule (see Angrist and Lavy, 1999). The idea here is that many countries have a rule establishing a maximum class size. Therefore, whenever grade enrollment is greater than this value, the school will create a second class. As a result, the average class size drops discontinuously. Therefore, the rule-prescribed class size based on grade enrollment may be a valid instrument for identifying exogenous variations in class size. If student performance is found to be different in classes differing in size due to this treatment, this gap can be attributed to a causal effect of class size. More recently, Kostantopoulos and Shen (2016) used the same approach to compute the average class size in fourth and eighth grade classes in Cyprus using data from TIMSS 2003 and 2007, as well as Li and Kostantopoulos (2016) for a sample of European countries using data from TIMSS 2011.

Effect of schooling

Luyten (2006) studies the absolute effect of schooling based on empirical data using the regression discontinuity approach. The estimation strategy exploits the availability of data about two adjacent grades in TIMSS 1995 combined with students´ date of birth. In this framework, the effect of age on achievement is estimated for each grade, where there is expected to be a discontinuity between the oldest students in the lower grade and the youngest students in the higher grade. This discontinuity reflects the effect of having received an extra year of schooling (i.e. being in the higher grade), assuming the average level of achievement is similar across cohorts. In order to obtain the cut-off points, the original variable representing the date of birth is transformed into a continuous variable with 12 potential values (one for each month)[7].

Luyten et al. (2008) also adopt a RD approach to assess the effect of one year´s schooling on student performance in reading, engagement in reading, and reading activities outside school. They use data from UK students participating in PISA 2000, because there are very low retention rates in this country. Therefore, the criterion for assigning students to the lower or upper grade according to their age can be assumed to be strictly adhered to. In this context, the effect of schooling is estimated as the difference between both grades minus the effect of age. Tiumeneva and Kuzmina (2015) also estimate the effectiveness of one year of schooling in seven countries using PISA 2009 data. Their approach is based on the determination of a particular threshold date and takes into account the distribution of students around this threshold point. Moreover, the empirical analysis was performed for both regular and vocational training programs.

Tracking

Kuzmina and Carnoy (2016) rely on a fuzzy regression discontinuity design based on school system age of entrance rules to examine the relative labor market value of vocational and academic education. In particular, they exploit the variation in a student's age relative to age cut-offs for entering primary school in each country to compare the gain for students in vocational and academic tracks using data from three European countries (Austria, Croatia and Hungary) with early tracking systems.

4.3. Difference in differences

The implementation of this method requires longitudinal data, where the same individuals are followed over time, or repeated cross-sectional data[8], where samples are drawn from the same population before and after the intervention. Unfortunately, this type of information is not available in comparative international datasets at individual or school level, since they only provide cross-sectional data referred to different population (fourth- or eighth-grade students in TIMSS and PIRLS or 15-year-old pupils in PISA). However, it is possible to take advantage of the strength of longitudinal designs in international studies when data are aggregated at country level, as Gustafsson (2007) claims. Thus we can find a large number of empirical studies adopting a DiD approach pooling data from different databases to assess the effects of multiple aspects, such as

tracking, peers, instructional time, preschool attendance, central examinations or different questions related to teaching.

Tracking

This approach has been applied by several authors to evaluate the effect of early tracking on performance by comparing differences in achievement between students attending primary school (when there is no tracking in any country) and secondary school (when some countries use tracking and others do not) across countries with and without tracked school systems. This idea was first explored by Hanushek and Woessman (2006) who implemented a DiD method to analyze country-level results from PIRLS, PISA, and TIMSS. Subsequently, Jakubowski (2010) tested the robustness of this approach by including controls for mean age differences between samples and countries and extended the empirical analysis using micro data. Likewise, Lavrijsen and Nicaise (2015) also adopted a similar approach. However, they attempted to account for the fact that part of the social origin effect already exists before tracking. Thus they apply the DiD analysis to social origin and reading achievement data from PIRLS 2006 (primary education) and PISA 2012 (secondary education). Ruhose and Schwerdt (2015) also analyzed the effect of tracking using DiD in a cross-country framework (45 countries), but they control for unobserved differences in relevant characteristics of the migrant and native student populations that remain constant across educational stages. They also exploit variation in migrant-native test score gaps between primary and secondary schools after pooling data from all cycles of TIMSS, PIRLS and PISA conducted between 1995 and 2012. Finally, Lavrijsen and Nicaise (2016) also adopted a DiD approach to examine the effects of the age at which tracking occurred on student achievement in a comparative perspective using data from PIRLS (2001, 2006 and 2011), TIMSS (2007 and 2011) and PISA (2006 and 2009). In addition, they distinguish the effects on different groups in the achievement distribution.

We can also find empirical studies in the literature that focus on a single country and evaluate some specific educational policies. For instance, Piopiunik (2014) studied the effects of early tracking exploiting a school reform implemented in the German region of Bavaria. He estimates a triple-differences model in which students in elementary and middle schools in Bavaria are compared with the respective changes of students in the

non-gymnasium tracks in the control states using data from PISA 2003 and 2006. Then, the performance of gymnasium students is added to the double-differences model as an additional control group to compute the triple differences estimator.

Peer group

Another interesting topic that can be studied using this approach is the impact of schoolmates on students' academic outcomes, i.e. the so-called peer effect. Schneeweis and Winter-Ebmer (2008) study this issue using PISA 2000 and 2003 data from Austria, where lower and upper secondary education is highly segregated. In order to address the potential self-selection of students into schools and peer-groups, they use two specifications: school type fixed effects and school fixed effects. Vardardottir (2015) also used PISA data about a highly segregated schooling system (Switzerland), although he controls for student heterogeneity by using track-by-school fixed effects to mitigate problems of self-selection in the type of students across schools. Ammermuller and Pischke (2009) exploit variation across classes within schools using PIRLS 2001 data about fourth-grade students attending a single-tracked primary school from school enrollment to at least fourth grade in six European countries. They also include school fixed effects in their econometric model in order to avoid potential bias due to self-selection.

Instructional time

Other authors have estimated the effects of instructional time on academic achievement. Specifically, Lavy (2015) studies a sample of students from 50 countries participating in PISA 2006, while Rivkin and Schiman (2015) gather data about 72 countries participating in PISA 2009. The estimation approach in both studies is based on exploiting the existence of test scores in three different subjects (reading, math and science) for each student and a relatively large variation in instructional time across subjects within schools. Thus it is possible to apply student fixed effects to control for individual time invariant characteristics that affect performance across subjects equally (innate abilities, previous achievements or family background). Moreover, Rivkin and Schiman (2015) also control for variations in the quality of instruction and classroom environment across schools for specific subjects. This is possible thanks to the existence of data for multiple grades in many schools (mainly ninth and tenth grade), thus they

can include school-by-subject fixed effects in the model (panel data structure). Therefore, they estimate a model that accounts for both school-by-grade and school-by-year fixed effects. This can be viewed as a difference in difference in differences model, where the difference between mathematics and reading scores for tenth grade minus the difference in ninth grade is related to the difference between mathematics and reading instruction time for tenth grade minus the difference in ninth grade. Finally, they also propose a model including a country-by-subject-by-grade term to account for national differences in the curriculum and other institutional features that might affect student performance.

Cattaneo et al. (2016) also use the variance of subject-specific instruction time to determine the causal impact of instruction time on student test scores in Switzerland using data from PISA 2009. However, they refined the empirical analyses performed in the previous papers by controlling for extra time spent on specific subjects either during school or after school (enrichment, remedial courses or paid private tutoring). Likewise, they performed separate empirical analyzes for different school tracks, since tracking starts in primary school in Switzerland.

Preschool participation

Schultz (2009) uses data from a single database (PISA 2003) to analyze the impact of pre-primary institution attendance on student performance at age 15. Her estimation strategy relies on the assumption that pre-elementary enrollment follows the same rules in all countries, thus the interaction of pre-primary attendance with structural quality measures resembles an international difference in differences approach. In particular, Schultz exploits within-country variation in pre-primary attendance and achievement, controlling for differences in various student, family, and school characteristics. This model yields reliable results when country fixed effects are included in the model. This implies that the remaining cross-country heterogeneity is unrelated to the effect of pre-primary attendance.

Felfe et al. (2015) evaluate whether the introduction of high-quality public childcare for three-year-olds has an influence on their cognitive performance by the end of compulsory schooling. In particular, they compare the educational outcomes of children

(at age 15) who were three years old before and after the reform in states where public childcare expanded substantially and states with a less pronounced increase in public childcare in the years immediately after the reform. Using this estimation strategy, they can control for all average time-constant differences between children living in different locations (by including a dummy for the treatment areas) and in different years (by including a dummy for the different cohorts).

Central examinations

Some researchers have also applied DiD using data from a single period. The application of this strategy is, however, subject to the adaptation of the method to other dimensions, such as the consideration of different subjects or grade levels. Jurges et al. (2005) pioneered the development of this idea to identify the effect of central exit examinations (CEE) on student performance in some German states. They exploit the fact that the dataset provides test scores for both mathematics and science, whereas only mathematics is tested in central exams. Therefore, their first difference is the difference between subjects and the second one is the difference between students in states with and without CEE. The key assumption required to identify the causal effect is that the difference in both outcome variables would be identical in the absence of treatment. Therefore, the excess on the difference in the mathematics test in CEE states should reflect the causal effect of interest. The key strength of this approach is that each student is serving as his or her control group. Thus it is possible to control for most of the heterogeneity at the individual level.

Anghel et al. (2015) study the effects of conducting and publishing the results of standardized tests in primary schools by exploiting the fact that this policy has only been implemented by one region in Spain (Madrid) since 2005. Therefore, their estimation strategy consists of setting up the treatment group before the treatment (students from Madrid who took the PISA 2000 reading exam and the PISA 2003 mathematics test) and after the treatment (students who took the 2009 PISA reading exam or the 2012 mathematics test), where the control group is composed of students from other Spanish regions where there was no primary school exam before (PISA 2000 or 2003) and after the treatment (PISA 2009 or 2012).

Pupil-teacher gender interaction

Several different researchers have used this approach to examine a number of aspects related to teaching activities. For instance, Ammermuller and Dolton (2006) investigated the potential existence of pupil-teacher gender interaction effects on performance, i.e. whether boys perform better when they are taught by male teachers and girls perform better when taught by female teachers. They use data from different waves of TIMSS (1995, 1999 and 2003) and PIRLS (2001) for only two countries (England and United States). Their strategy consists of considering two performance measures for the same student in different subjects and including student fixed effects in their econometric model to avoid potential bias in the estimation of the treatment effects because the assignment of class teacher gender may not be random. Subsequently, Cho (2012) extended this empirical analysis to a sample of students from 15 OECD countries using a similar approach.

Teaching practices

Schwerdt and Wuppermann (2011) use information provided by teachers and students about US eighth-grade students participating in TIMSS 2003 to study the effect of different teaching strategies on student achievement. In particular, they compare two teaching practices (lecture style presentations vs. in-class problem solving) exploiting between-subject variation to control for unobserved student traits. Focusing on a variable representing the teaching time spent on lecture style presentation relative to problem solving, they also apply school fixed effects to eliminate the effects of between-school sorting and exclude any systematic between-school variation in performance or teaching practice.

Similarly, Bietenbeck (2014) uses data about US students participating in TIMSS 2007 to analyze the effects of traditional and modern teaching practices on students´ cognitive skills. He also exploits the existence of two different observations for each student from two different subjects and includes student fixed effects in the empirical model to account for the sorting to teaching practices across schools and classrooms. Moreover, he also controls for a rich set of teacher and class characteristics in order to account for potential bias derived from unobserved teachers´ characteristics.

Other topics

Ammermuller (2012) merges micro data from two different datasets (PIRLS 2001 and PISA 2000) to investigate whether cross-country differences in educational opportunities are related to the institutional features of schooling systems using a DiD estimation approach. The schooling systems are analyzed at grade four and grade nine/ten, and the features studied are as follows: the use of streaming in school systems, annual instruction time, proportion of students in private schools and school autonomy. The identification strategy uses the difference in the dependence between social status and educational outcomes across grades between countries whose institutions have changed between grades and countries with no institutional changes across grades. Therefore, this by and large controls for country-specific factors, aside from the schooling system, assuming they are identical for students of different ages. Therefore, the DiD approach consists of eliminating the country-specific factors in order to estimate the changes in educational opportunities between grades for each country.

Kiss (2013) examines grade discrimination using data about German primary and secondary schools from PIRLS 2001 and PISA 2003, respectively. Specifically, Kiss studies whether second-generation immigrants and girls are graded worse in math than comparable natives or boys by applying class fixed effects regressions to control for the average teacher effect. Additionally, he applies a matching approach that accounts for nonlinear relationships between grades and teacher characteristics.

Hanushek et al. (2013) study the effect of school autonomy on student achievement or, more specifically, whether altering the degree of local school decision-making autonomy might have an impact on performance. For this purpose, they propose using a cross-country panel analysis covering the 42 countries that participated in at least three of the four waves of PISA (2000, 2003, 2006 and 2009). Being a panel analysis at country level, their model can include country fixed effects to exploit international variation in policy initiatives focused on autonomy, while accounting for cross-country divergences in institutional features.

Hanushek et al. (2014) combine the use of student fixed effects and an IV approach to investigate the role of teacher cognitive skills in explaining student outcomes. The data

used for estimating teacher numeracy and literacy skills was the Programme for the International Assessment of Adult Competencies (PIAAC). Subsequently, this dataset was merged with PISA micro data for 23 countries to estimate international education production functions. Their identification strategy exploits information about the performance of students and teachers in two different subjects, thus they can control for unobserved student-specific characteristics that similarly affect math and reading performance, as well as for all differences across countries that are not subject specific. Subsequently, they also exploit exogenous variation in teacher cognitive skills using international differences in relative wages of non-teacher public sector employees as an instrument.

Green and Pensiero (2016) also use a similar approach to assess the contribution of upper secondary education and training to inequalities in skills opportunities and outcomes using data about literacy and numeracy skills in PISA 2000 and the Survey of Adult Skills (SAS) conducted by the OECD in 2011-12. Their estimation strategy is based on comparing the variations in literacy and numeracy skills demonstrated by students at different ages across countries, using a pseudo-cohort derived from the 15-year-olds participating in PISA 2000 and the SAS (2011/12) sample of 25- to 29-year-olds who represent the PISA sample 12 years later.

Finally, Pedraja-Chaparro et al. (2016) assess whether the concentration of immigrant students in Spanish schools during the period 2003-2009 has affected student performance. Their estimation strategy consists of identifying schools without sampled immigrants in all the datasets (control group) and schools hosting immigrants throughout this period (treatment group) and calculating the average difference in outcomes separately for each group over the period. Likewise, as the percentage of immigrants varies across schools, the DiD approach is adapted to deal with a dose treatment, where the dose is the percentage of immigrants at each school belonging to the treated group.

4.4. Propensity Score Matching

Although weaker than other methods, PSM has been widely applied with international data in order to obtain more accurate estimates when performing comparisons between public and private schools or students in different tracks, for example.

Public vs. private schools

The first authors to use the PSM approach were Vandenberghe and Robin (2004). They analyzed the effect of attending a private school on students´ achievement in different countries using alternative approaches. Specifically, propensity score matching is implemented by matching pupils attending private schools (treated) and students attending public schools (control). Similarly, Dronkers and Avram (2010) also use this method to estimate the effectiveness of private schools on reading achievement in 26 countries using a pooled sample of data from three waves of PISA (2000, 2003 and 2006).

In addition to such cross-country studies, we can also find empirical studies dealing with this issue in a national context for countries with a high proportion of students enrolled in private schools. For example, Cornelisz (2013) uses data from two different waves of PISA (2006 and 2009) to analyze the case of the Netherlands, where this proportion is nearly two-thirds of all students. Crespo-Cebada et al. (2014) also apply this technique to analyze the case of Spanish schools, using PISA 2006 data about different regions. The main novelty of their approach is that they implement this estimation strategy within the framework of stochastic parametric frontier analysis. Finally, Gee and Cho (2014) analyze the problem of aggressive behaviors in South Korea comparing single-sex versus coeducational schools. In their empirical study, they use data from TIMSS 2011 and the 2005 Korea Education Longitudinal Study (KELS) and also rely on the PSM approach to reduce the threat of selection bias between the two groups of schools.

Tracking

In a comparative study, Lee (2014) applies the propensity score matching technique to PISA 2009 data to compare the effect of academic and vocational tracks on students' educational expectations and whether the effect varies across different socio-economic

statuses in Austria and Italy. Austria and Italy were selected for comparison because they apply tracking at different stages of the educational system (early stages in Austria and later in Italy). Similarly, Arikan et al. (2016) also use PSM to predict the mathematics achievement of Turkish students compared to Australian students. In particular, they match the Australian and Turkish samples from TIMSS 2007 and 2011 based on relevant background variables (educational resources at home and self-confidence).

Jakubowski (2015) evaluates differences in the magnitude of student progress across two types (vocational and general vocational) of upper secondary education in Poland using data from the PISA 2006 national study that extended the sample to cover 16- and 17-year-olds (enrolled in tenth and eleventh grade in the Polish school system). This dataset provides supplementary information on students´ previous scores in national exams. This makes it possible to control for students' innate abilities using a PSM approach. More specifically, the main contribution of this study is that the proposed model adds a latent variable to propensity score matching. This latent variable should make the treatment estimates more precise than a standard approach, where matching is conducted considering only the set of observable variables.

Other topics

Agasisti and Murtinu (2012) employ propensity score matching to investigate the effects of perceived competition among Italian secondary schools on their performance in mathematics using data from PISA 2006. Specifically, the authors exploit the information provided by school principals regarding whether or not the school is operating in an area where there is competition for students to split the available sample into two groups. Consequently, the presence of competition is considered as a potential endogenous treatment. In another study referred to the case of Italy, Ponzo (2013) examines whether being a victim of school bullying affects educational achievement. Specifically, using data from PIRLS 2006 and TIMSS 2007, Ponzo analyzes the impact on performance in two different subjects (math and science) for students enrolled in the fourth and eighth grade levels, applying PSM to control for a wide number of individual characteristics.

Jiang and McComas (2015) apply the PSM approach to examine the effects of the level of openness of inquiry teaching on student science achievement and attitudes using PISA data from 2006. In the context of their study, the term inquiry teaching includes very different teaching practices, all of which somehow involve student decision-making. In order to evaluate such practices, the authors define five different levels of inquiry teaching considered as five categories of treatments in their causal analysis. Since the treatment is a five-level categorical variable, the generalized propensity scores were estimated using multinomial logistic regression. This generates one set of propensity scores for each treatment level (Imbens, 2000). The empirical analyses were conducted separately for each country participating in PISA. Thus it is possible to examine whether the impact of inquiry teaching is consistent across different countries.

Finally, Hogrebe and Strietholt (2016) use data from PIRLS 2011 to estimate the effect of not attending preschool on grade-four students' reading achievement by implementing propensity score matching. The empirical analysis is performed for nine different countries with well-established early childhood education systems with high enrollment rates. Thus they are well suited for identifying both control and treatment groups. It is noteworthy that their binary treatment variable is defined in such a way that non-attendance is the treatment condition[9], since they consider this effect to be more relevant for policy makers who are considering extending preschool attendance.

5. Summary of empirical studies

After reviewing the four approaches and the contents of all the applications, we now synthesize the main aspects of these papers and provide an overview of the journals in which they were published. From our viewpoint, this should provide sound guidance for researchers interested in combining the use of causal inference techniques with educational data from large-scale assessments. In this manner, they would be able to identify the best outlets for their empirical studies. First of all, we find that the number of studies has increased substantially over the analyzed period, as shown in Figure 1. Thus it is clear that the use of causal inference methods with educational data from large-scale international assessments is gradually becoming a more common practice in the field of education economics, and this trend is very likely to continue to grow in the near future.

Figure 1. Number of empirical studies (2004-2016)
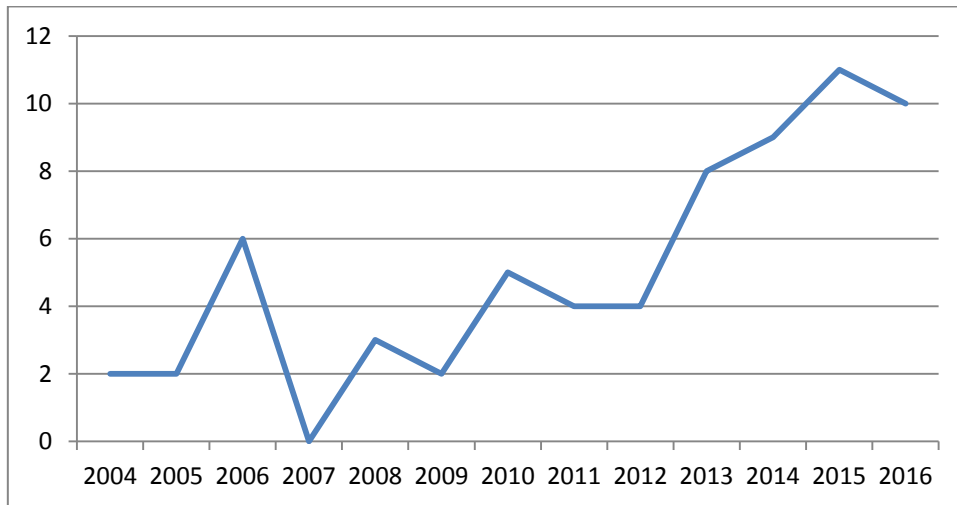
Regarding the data sources, PISA is clearly the most common option used by researchers given that this dataset provides the world's most extensive and rigorous information about the knowledge and skills of secondary school students. As a result, it is employed in two out of every three studies (Figure 2), although it is sometimes combined with other datasets. Then, of the two surveys conducted by the IEA, TIMSS seems to be more popular among researchers, especially in older articles, since it started PIRLS (1995 vs. 2001). Moreover, TIMSS is repeated every four years. This means that there are more available waves of data. It also provides information about student outcomes in two different subjects (mathematics and sciences) or at two different stages of the educational system (fourth and eighth grades). Thanks to this, the difference in differences approach can be implemented. In contrast, PIRLS only assesses one subject (reading) for fourth graders, and there are only three different waves available.

Figure 2. Datasets used in empirical studies
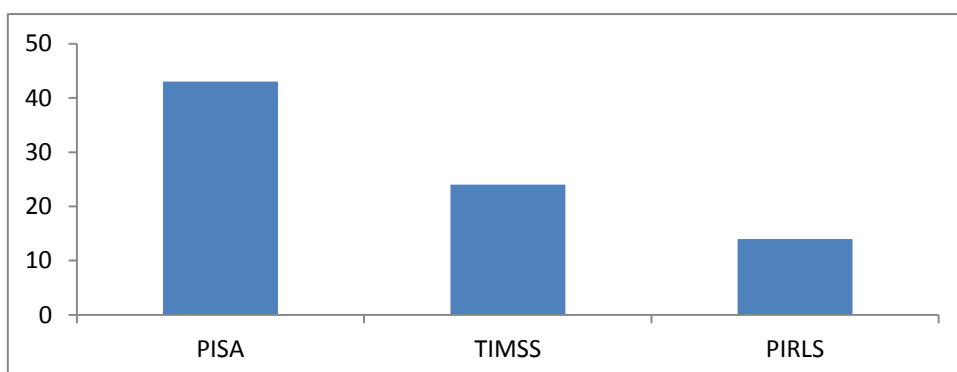
In addition, Figure 3 highlights that the most common strategy employed in the cited studies is DiD closely followed by IV. Although the work with different cross-sectional waves complicates the use of DiD (Rutkowski and Delandshere, 2016), the assumptions required for adopting this strategy are less demanding than for other methods. As a result, we find that a considerable number of papers use this approach. However, DiD requires researchers to be creative, since they have to emulate an ideal situation in which students or schools can be evaluated at two different times (before and after implementing the evaluated intervention) without actually having longitudinal data. In the case of IV, all that is required for implementation is to find an instrument that suits a particular problem and meets some basic assumptions. Although this also requires some creativity on the part of the researcher, the wide range of variables provided by large-scale assessments makes this search more feasible[10]. Other methods such as PSM or RDD require a huge number of observations with similar characteristics. This condition might be difficult to satisfy in many cases, and therefore they are used less frequently.

Figure 3. Methods used in empirical studies
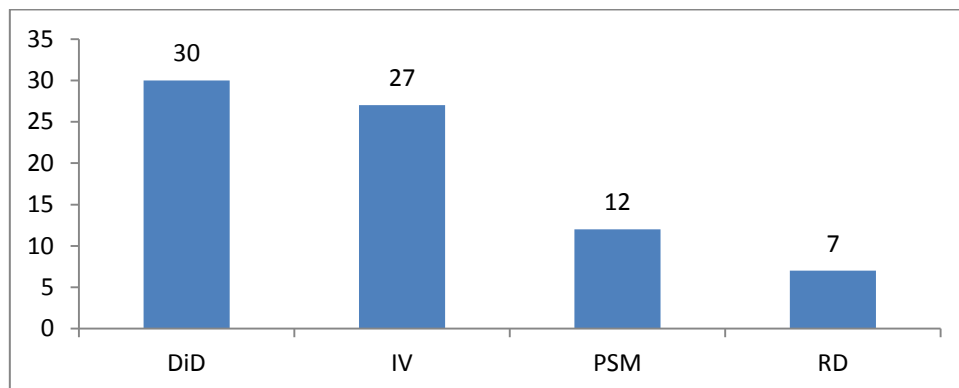
The examined papers cover a wide range of topics. Nevertheless, some, such as the comparison between public and private schools, class size effects and the influence of tracking, clearly stand out from the rest. Other noteworthy key issues studied in several papers are different aspects related to teaching, the effect of additional schooling, the consequences of immigrant concentration, well-known peer effects, the expansion of instructional time, age at school entry, and the long-term effects of pre-school education. Figure 4 summarizes the frequency of studies applied to these topics.

Figure 4. Topics examined in applications
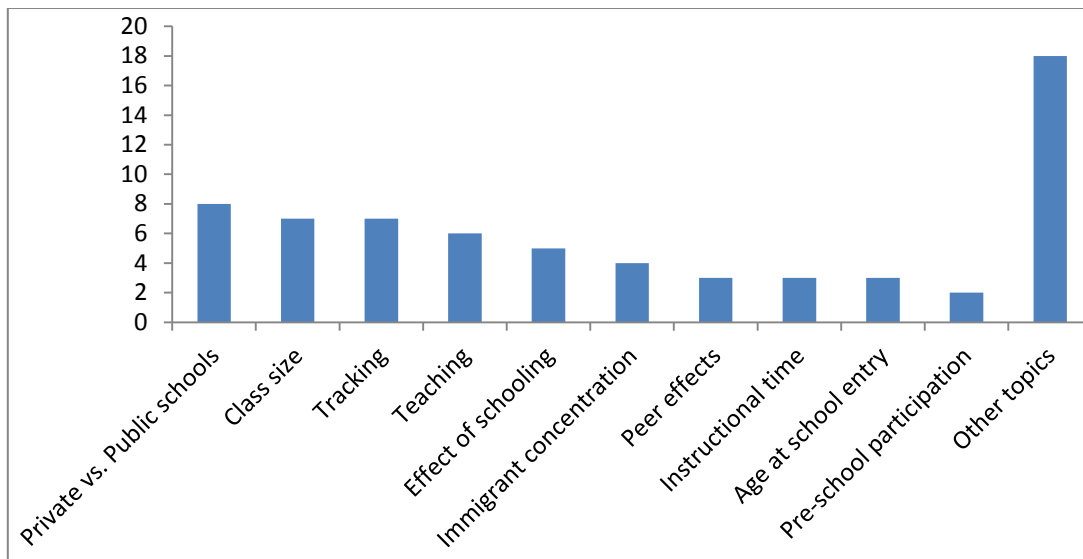
Finally, we aim to provide some advice for researchers interested in identifying where they might publish their empirical research using causal inference methods. For this purpose, we have compiled the name of the journals in which the surveyed papers were published. They are classified according to the subject categories provided by two of the best-known academic journal classifications: the Journal of Citation Reports (JCR) index published by Thomson Reuters and the SCImago rank developed from the Scopus database[11].

The first conclusion of this analysis is that the huge majority of the surveyed empirical papers (55 out of 66) were published in journals ranked in the above classifications (55 in SCImago and 46 in JCR). The exceptions are two chapters in books, six working papers and three journals not included in either the SCImago or JCR classifications. Another interesting conclusion derived from this exercise is that significantly more papers are published in economics journals than in education journals (Figure 5).

Nevertheless, we consider that the quality of the journals should also be taken into consideration. To do this, we explore the quartile rankings of the journals using the impact factor data estimated in each classification[12]. In this respect, the information reported in Figure 6 indicates that most papers using these estimation strategies were published in the two highest quartiles for both categories. Therefore, we take the view that adopting a causal inference approach to deal with large-scale data facilitates access

to publication in top-ranking journals, irrespective of the subject category in which those journals are included.

Figure 5. Subject categories of published papers
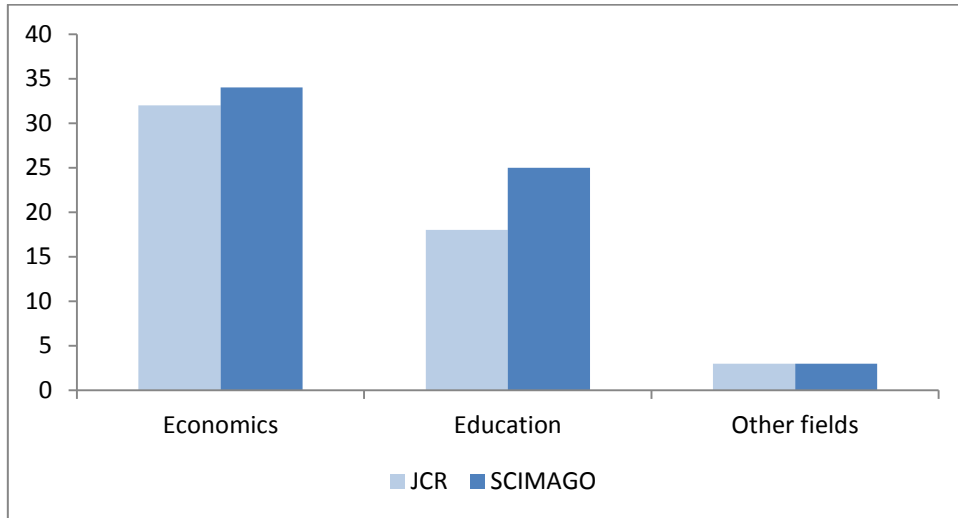
Figure 6. Distribution of papers across quartile rankings according to impact factors

(a) JCR   (b) SCIMAGO
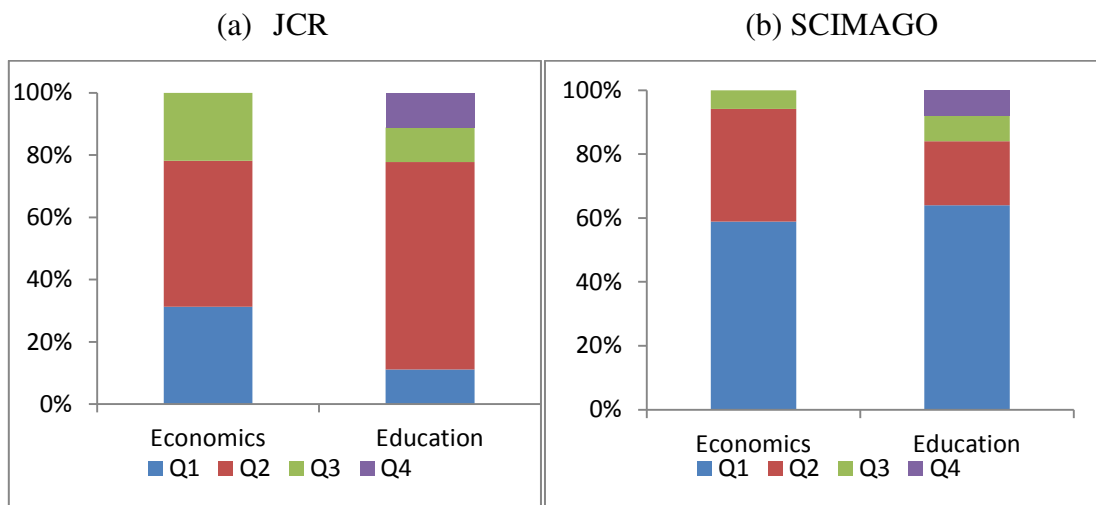
6. Conclusions

Grounded on a systematic literature review, this paper provides a detailed and comprehensive description of four estimation strategies (IV, DiD, RDD and PSM) employed in multiple empirical studies using data from the best-known large-scale educational assessments (PISA, TIMSS and PIRLS) to evaluate the effectiveness of different interventions through causal inference techniques. We believe that this

research is potentially of use for policy makers, professionals, researchers and practitioners interested in implementing rigorous evaluations of the available databases based on quasi-experimental designs. Thus we focus essentially on the methodological issues related to the econometric approach employed and not on the significance of the investigated effects.

Our literature review reveals a wide range of alternative estimation strategies that can be adopted to avoid the recurrent problem of endogeneity. Endogeneity frequently biases the results of traditional econometric methods based on associations between variables, especially when only cross-sectional data are available. Actually, the shortage of reliable data and/or the low quality of the available information are the main problems that researchers wishing to conduct causal inference analysis in the field of education economics have to face in most countries. Thus, their only option for performing an empirical analysis in many cases is to fall back on data provided by international comparative surveys.

The main weaknesses of such datasets are that they do not provide information about a previous measure of achievement and their cross-sectional and pseudo-panel structure. Nevertheless, many authors have demonstrated that it is possible to draw causal inference from these datasets, even if there is no clear exogenous variation in the observed data. In particular, some authors exploit existing information about different classes within the same school (this is only possible with TIMSS), having students enrolled in different courses and being evaluated in different subjects (this applies for PISA and also for TIMSS 1995) or, alternatively, the use of institutional rules as an instrumental variable or cut-off point to apply a regression discontinuity approach. On the other hand, others make a greater effort to emulate the existence of longitudinal data by matching data retrieved from different datasets implemented at different times of the educational track (e.g. TIMSS for fourth or eight graders and PISA for 15-year-old pupils) or build pseudo-panels using data from different waves of the same dataset.

According to our systematic review, the most common strategy employed in empirical studies is to use difference in differences and instrumental variables. The difference in differences method has weaker assumptions, and the only requirement for the instrumental variables technique is to find an instrument suitable for a particular

problem. Both methods require some level of creativity on the part of the researcher, but the wide range of variables provided by large-scale assessments makes this search easier. Likewise, researchers might also gather information from other external sources of data. Other methods such as propensity score matching or regression discontinuity design require a lot of observations with similar characteristics. This condition might be difficult to satisfy in many cases. Thus they are less often used in empirical studies.

Even though educational researchers have demonstrated that it is possible to evaluate interventions based on the data available in the analyzed international datasets, we would like to alert policy makers about the need to improve the volume and quality of data in national and international datasets. This would help researchers to apply an appropriate evaluation procedure for the process of evaluating interventions or practices. For example, several such enhancements have already been implemented as national options for the PISA studies in Germany or Poland (Klieme, 2013; Jakubowski, 2015). In view of the importance of assessing the impacts of educational policies in particular, we would like to draw attention to the need to build longitudinal datasets at student or school level. In this manner, it would be possible to follow up the assessed units of analysis over a long period. This is the type of data that is required to evaluate the effectiveness of particular interventions in the long run.

Notes

[1] See Angrist and Pischke (2008), Khandker et al., (2010) or Gertler et al., (2016) for a more comprehensive discussion of these methods and their practical implementation.

[2] See Angrist and Pischke (2008, 2014) for details.

[3] This approach is also known as a "cutting-point design" (Rossi et al., 2004, p. 289).

[4] See Imbens and Lemieux (2008) for details.

[5] The straightforward solution would be to widen the margins around the threshold. However, this option also has its limitations, since the probability of the units placed above and below the cut-off value being similar with regard to their treatment status is lower with a wider bandwidth.

[6] This estimation strategy was only possible using data from TIMSS 1995. In the TIMSS study conducted in 1999, data was collected for students from only one grade (eighth, but not seventh), making the between-grade comparison impossible.

[7] For example, a student born in March 1985 received a score of 85.25, and a student born in April received a score of 85.33.

[8] In repeated cross-sectional surveys, the composition of the groups with respect to the fixed effects term must be unchanged to ensure before-after comparability (Blundell and Dias, 2009).

[9] Another possible alternative would be to model different preschool doses (See Imai and van Dyck, 2004 for details).

[10] Researchers might also gather information from other external data sources.

[11] In some cases, the journal can be classified in more than one category (e.g. *Economics of Education Review* is included in both categories -Economics and Education-).

[12] We use the impact factor (IF) of the journal in 2015. Q1 denotes the top 25% of the IF distribution, Q2 signifies a middle-high range (between top 50% and top 25%), Q3 indicates middle-low range (top 75% to top 50%), and Q4 refers to the bottom 25% of the IF distribution.

References

Agasisti, T., Murtinu, S. (2012). 'Perceived' competition and performance in Italian secondary schools: new evidence from OECD–PISA 2006. British Educational Research Journal, 38(5), 841-858.

Ammermüller, A. (2012). Institutional features of schooling systems and educational inequality: Cross-Country evidence from PIRLS and PISA. *German Economic Review*, *14*(2), 190-213.

Ammermüller, A., Dolton, P. (2006). Pupil-teacher gender interaction effects on scholastic outcomes in England and the USA, *ZEW Discussion Paper 06-060. Mannheim, Germany: Centre for European Economic Research.*

Ammermueller, A., Pischke, J. S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study, *Journal of Labor Economics*, 27(3), 315-348.

Anghel, B., Cabrales, A., Sainz, J., Sanz, I. (2015). Publicizing the results of standardized external tests: does it have an effect on school outcomes?. *IZA Journal of European Labor Studies*, 4(1), 1.

Angrist J.D., Lavy V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement, *Quarterly Journal of Economics*, 114(2), 533-575.

Angrist, J. D., Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D., Pischke, J. S. (2014). *Mastering metrics: the path from cause to effect*. Princeton University Press.

Angrist, J.D., Pischke, J.S. (2015). *Mastering metrics*, Princeton University Press, NJ.

Arikan, S., van de Vijver, F., Yagmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian Students in TIMSS 2007 and 2011. *Eurasia Journal of Mathematics, Science and Technology Education*, 12, 2039-2059.

Bedard, K., Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects, *The Quarterly Journal of Economics*, 1437-1472.

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics*, 30, 143-153.

Blundell, R., Dias, M. C. (2009). Alternative approaches to evaluation in empirical microeconomics. Journal of Human Resources, 44(3), 565-640.

Cattaneo, M.A., Oggenfuss, C., Wolter, S.C. (2016). The more, the better? The impact of instructional time on student performance, *CESIFO Working Paper, 5813*.

Cho, I. (2012). The effect of teacher–student gender matching: Evidence from OECD countries. *Economics of Education Review*, 31(3), 54-67.

Choi, Á., Calero, J., Escardíbul, J. O. (2012). Private tutoring and academic achievement in Korea: An approach through PISA-2006, *KEDI Journal of Educational Policy*, 9(2), 299-302.

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects, *Economics of Education Review*, 26(6), 673–82.

Cornelisz, I. (2013). Relative Private School Effectiveness in the Netherlands: A Reexamination of PISA 2006 and 2009 data, *Procedia Economics and Finance*, 5, 192-201.

Creemers, B. P. M., Kyriakides, L. (2008). *The dynamics of educational effectiveness*, London, Rouledge.

Crespo-Cebada, E., Pedraja-Chaparro, F., Santín, D. (2014). Does school ownership matter? An unbiased efficiency comparison for regions of Spain. Journal of Productivity Analysis, 41(1), 153-172.

De Witte, K., López-Torres, L. (2015). Efficiency in education: a review of literature and a way forward. *Journal of the Operational Research Society*, *forthcoming.*

Denny, K., Oppedisano, V. (2013). The surprising effect of larger class sizes: Evidence using two identification strategies, *Labour Economics,* 23, 57-65.

Dronkers, J., Avram, S. (2010). A cross-national analysis of the relations of school choice and effectiveness differences between private-dependent and public schools. *Educational Research and Evaluation*, 16(2), 151-175.

Edwards, S., Marin, A. G. (2015). Constitutional rights and education: An international comparative study, *Journal of Comparative Economics*, 43(4), 938-955.

Falck, O., Woessmann, L. (2013). School competition and students' entrepreneurial intentions: International evidence using historical Catholic roots of private schooling, *Small Business Economics*, 40(2), 459-478.

Felfe, C., Nollenberger, N., Rodríguez-Planas, N. (2015). Can't buy mommy's love? Universal childcare and children's long-term cognitive development, *Journal of Population Economics*, 28(2), 393-422.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.

Gamboa, L., Rodríguez, M., García, A. (2013). Differences in motivations and academic achievement. *Lecturas de Economía*, 78, 9-44.

García-Pérez, J. I., Hidalgo-Hidalgo, M., Robles-Zurita, J. A. (2014). Does grade retention affect students' achievement? Some evidence from Spain. *Applied Economics*, 46(12), 1373-1392.

Gee, K., Cho, R.M. (2014). The effects of single-sex versus coeducational schools on adolescent peer victimization and perpetration. *Journal of Adolescence*, 3, 1237-1251.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., Vermeersch, C. M. (2016). *Impact evaluation in practice. 2ⁿᵈ edition*, World Bank Publications.

Green, A. D., Pensiero, N. (2016). The effects of upper secondary education and training systems on skills inequality: A quasi-cohort analysis using PISA 2000 and the OECD survey of adult skills. *British Educational Research Journal*, 42(5), 756-779.

Gustafsson, J. E. (2007). Understanding causal influences on educational achievement through analysis of differences over time within countries. In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement*, Washington, DC: Brookings, pp. 37–63.

Gustafsson, J. E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, 7(1), 1-17.

Gustafsson, J.E. (2013). Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement, *School effectiveness and School Improvement*, 24(3), 275-295.

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

Hanushek, E.A., Link, S., Woessman, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA, *Journal of Development Economics*, 104, 212-232.

Hanushek, E. A., Piopiunik, M., Wiederhold, S. (2014). The value of smarter teachers: International evidence on teacher cognitive skills and student performance, *National Bureau of Economic Research*, Working Paper nº 20727.

Hanushek, E. A., Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal*, 116, 63–76.

Hanushek, E. A., Woessman, L. (2011). The economics of international differences in educational achievement. In Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), *Handbook of the economics of education (vol. 3)*, North Holland, Amsterdam, pp. 89–200.

Hanushek, E. A., Woessmann, L. (2014). Institutional Structures of the Education System and Student Achievement: A Review of Cross-country Economic Research. *Educational Policy Evaluation through International Comparative Assessments*, 145.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. (1979). Sample selection bias as an especification error, *Econometrica*, 47, 153-161.

Heckman, J., Navarro, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics*, 86(1), 30-57.

Hogrebe, N., Strietholt, R. (2016). Does non-participation in preschool affect children's reading achievement? International evidence from propensity score analyses. *Large-scale Assessments in Education*, 4(8), *in press*.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81, 945–960.

Imai, K., Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score, *Journal of the American Statistical Association*, 99(467), 854–866.

Imbens, G.W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3), 706–710.

Imbens, G. W., Lemieux, T. (2008). Regression discontinuity designs: A guide to practice, *Journal of Econometrics*, 142(2), 615-635.

Isphording, I. E., Piopiunik, M., Rodríguez-Planas, N. (2016). Speaking in numbers: The effect of reading performance on math performance among immigrants, *Economics Letters*, 139, 52-56.

Jakubowski, M. (2010). Institutional Tracking and Achievement Growth: Exploring Difference-in-Differences Approach to PIRLS, TIMSS, and PISA Data. In Dronkers (ed.), *Quality and Inequality of Education*, Springer Netherlands, pp. 41-81.

Jakubowski, M. (2015). Latent variables and propensity score matching: a simulation study with application to data from the Programme for International Student Assessment in Poland, *Empirical Economics*, 48(3), 1287-1325.

Jensen, P., Rasmussen, A. W. (2011). The effect of immigrant concentration in schools on native and immigrant children's reading and math skills, *Economics of Education Review*, 30(6), 1503-1515.

Jiang, F., McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. International Journal of Science Education, 37(3), 554-576.

Johansson, S. (2016). International large-scale assessments: what uses, what consequences?, *Educational Research*, 1-10. *forthcoming*.

Jürges, H., Schneider, K. (2004). International differences in student achievement: An economic perspective, *German Economic Review*, 5(3), 357-380.

Jürges, H., Schneider, K., Büchel, F. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. Journal of the European Economic Association, 3(5), 1134-1155.

Kamens, D.H. (2009). Globalization and the growth of international educational testing and national assessment, *Comparative Education Review*, 54(1), 5-25.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: a review and synthesis. *Large-scale Assessments in Education*, 4(1), 1-24.

Khandker, S. R., Koolwal, G. B., Samad, H. A. (2010). *Handbook on impact evaluation: quantitative methods and practices*. World Bank Publications.

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics*, 21(5), 447-463.

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development, in Von Davier, M., Gonzalez, E., Kirsch, I. (eds.). *The role of international large-scale assessments: perspectives from technology, economy, and educational research,* Springer Netherlands, 115-147.

Konstantopoulos, S., Shen, T. (2016) Class size effects on mathematics achievement in Cyprus: evidence from TIMSS. *Education Research and Evaluation,* 22, 86–109.

Konstantopoulos, S., Traynor, A. (2014). Class Size Effects on Reading Achievement Using PIRLS Data: Evidence from Greece. *Teachers College Record*, 116(2), n2.

Kuzmina, J., Carnoy, M. (2016). The effectiveness of vocational versus general secondary education: Evidence from the PISA 2012 for countries with early tracking, *International Journal of Manpower*, 37(1), 2-24.

Lavrijsen, J., Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal*, 14(3-4), 206-221.

Lavrijsen, J., Nicaise, I. (2016). Educational tracking, inequality and performance: New evidence from a differences-in-differences technique. *Research in Comparative and International Education*, 11(3), 334-349.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, 125(588), 397-424.

Lee, B. (2014). The influence of school tracking systems on educational expectations: a comparative study of Austria and Italy, *Comparative Education*, 50(2), 206-228.

Lee, J., Fish, R. (2010). International and interstate gaps in value-added math achievement: Multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education*, 117(1), 109-137.

Li, W., Konstantopoulos, S. (2016). Class Size Effects on Fourth Grade Mathematics Achievement: Evidence From TIMSS 2011. *Journal of Research on Educational Effectiveness*, in press.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education*, 32(3), 397–429.

Luyten, H., Peschar, J., Coe, R. (2008). Effects of Schooling on Reading Performance, Reading Engagement, and Reading Activities of 15-Year-Olds in England, *American Educational Research Journal*, 45(2), 319-342.

Luyten, H., Veldkamp, B. (2011). Assessing effects of schooling with cross-sectional data: Between-grades differences addressed as a selection-bias problem. *Journal of Research on Educational Effectiveness*, 4(3), 264-288.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. Santa Mónica, CA: The RAND Corporation.

Morgan, S. L., Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. Cambridge: University Press.

Lavrijsen, J., Nicaise, I. (2015). New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. *European Educational Research Journal*, 14(3-4), 206-221.

Neyman, J. (1923). Statistical problems in agriculture experiments. *Journal of the Royal Statistical Society*, Series B, 2, 107–180.

Pedraja-Chaparro, F., Santín, D., Simancas, R. (2016). The impact of immigrant concentration in schools on grade retention in Spain: a difference-in-differences approach. *Applied Economics*, 48(21), 1978-1990.

Perelman, S., Santin, D. (2011). Measuring educational efficiency at student level with parametric stochastic distance functions: an application to Spanish PISA results. *Education Economics*, 19(1), 29-49.

Pfeffermann, D., Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics*, 5(3), 1726.

Piopiunik, M. (2014). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. *Economics of Education Review,* 42, 12–33.

Pokropek, A. (2016). Introduction to instrumental variables and their application to large-scale assessment data. *Large-scale Assessments in Education*, 4(1), 1.

Ponzo, M. (2013). Does bullying reduce educational achievement? An evaluation using matching estimators. *Journal of Policy Modeling*, 35, 1057–1078.

Ponzo, M., Scoppa, V. (2014). The long-lasting effects of school entry age: Evidence from Italian students, *Journal of Policy Modeling*, 36(3), 578-599.

Puhani, P. A., Weber, A. M. (2008). Does the early bird catch the worm?. In Dutsman, C., Fitzenberg, B., Machin, S. (Eds.). *The economics of education and training*, Physica-Verlag HD, 105-132.

Rivkin, S.G., E.A. Hanushek, Kain, J.F. (2005). Teachers, schools, and academic achievement, *Econometrica*, 73 (2), 417–58.

Rivkin, S. G., Schiman, J. C. (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal*, 125(588), F425-F448.

Robinson, J.P. (2014). Causal Inference and Comparative Analysis with Large-Scale Assessments. In Rutkowski, L., von Davier, M., Rutkowski, D. (eds.). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, Boca Raton.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, 11(3), 207–224.

Rosen, M., Gustafsson, J. E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-scale Assessments in Education*, 4(1), 1.

Rossi, P. H., Freeman, H. E. Lipsey, M. W. (2004). *Evaluation, a systematic approach*, Thousand Oaks/London/New Delhi, SAGE Publications.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (2008). Objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840.

Ruhose, J., Schwerdt, G. (2015). Does Early Educational Tracking Increase Migrant-Native Achievement Gaps? Differences-In-Differences Evidence Across Countries, *CESIFO Working Paper No. 5248*.

Rutkowski, L., von Davier, M., Rutkowski, D. (2014). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, Boca Raton.

Rutkowski, D., Delandshere, G. (2016). Causal Inferences with Large Scale Assessment Data: Using a Validity Framework. *Large-Scale Assessments in Education*, 4(1), 1–18. doi:10.1186/s40536-016-0019-1.

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

Schneeweis, N., Winter-Ebmer, R. (2008). Peer effects in Austrian schools, *Empirical Economics*, 32, 387-409.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Schütz, G. (2009). Does the quality of pre-primary education pay off in secondary school? An international comparison using PISA 2003. Ifo Institute for Economic Research, Ifo Working Paper, 68.

Schwerdt, G., Wuppermann, A. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review,* 30, 365–379.

Shadish, W.R., Cook, T.D., Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Strietholt, R., Gustafsson, J. E., Rosen, M., & Bos, W. (2014). Outcomes and Causal Inference in International Comparative Assessments. In Stietholt, R., Bos, W., Gustafsson, J.E., Rosen, M. (eds.), *Educational Policy Evaluation through International Comparative Assessments*, Waxman, Münster, New York.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, 36(4), 187-198.

Tiumeneva, Y. A., Kuzmina, J. V. (2015). The Difference That One Year of Schooling Makes for Russian Schoolchildren: Based on PISA 2009: Reading. *Russian Education & Society*, 57(4), 214-253.

Todd, P. E., Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, 113 (485), 3-33.

Vardardottir, A. (2015). The impact of classroom peers in a streaming system. *Economics of Education Review*, 49, 110-128.

Vandenberghe, V., Robin, S. (2004). Evaluating the effectiveness of private education across countries: a comparison of methods. *Labour Economics*,*11*(4), 487-506.

Wilde, E. T., Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment, *Journal of Policy Analysis and Management*, 26(3), 455-477.

Webbink, D. (2005). Causal effects in education. *Journal of Economic Surveys*,*19*(4), 535-560.

West, M.R., Woessmann, L. (2006). Which School Systems Sort Weaker Students into Smaller Classes? International Evidence", *European Journal of Political Economy*, 22 (4), 944-968.

West, M. R., Woessmann, L. (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), 229-255.

Woessmann, L. (2005). Educational production in Europe, *Economic Policy,* 20(43), 446-504.

Woessmann, L., West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3), 695-736.

Woessmann, L. (2007). International evidence on school competition, autonomy, and accountability: A review. *Peabody Journal of Education*, 82(2-3), 473-497.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

APPENDIX

Table A1. Empirical studies using causal inference with data from international large scale assessments

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2004	Vandenberghe, V. Robin, S	PISA 2000	Cross-sectional	Cross-country (9 countries)	IV PSM	Evaluate the effect of private education on educational outcomes across countries
2004	Jürges, H. Schneider, K.	TIMSS 1995	Cross-sectional	Cross-country (23 countries)	IV DiD	Explain what causes between-country gaps in mathematics test score distributions
2005	Jürges, H. Schneider, K. Büchel, F.	TIMSS 1995	Cross-sectional	Germany	DiD	Estimate the causal effect of central examinations on student performance in Germany
2005	Woessmann, L.	TIMSS 1995	Cross-sectional	Cross-country (17 countries)	RD	Evaluate class-size effects on student performance
2006	Hanushek, E. A. Woessmann, L.	TIMSS 1995 TIMSS 1999 PISA 2003 PIRLS 2001	Pooled data	Cross-country (18-26 countries)	DiD	Examine how educational tracking can affect mean performance and inequality across students
2006	Woessmann, L. West, M.R.	TIMSS 1995	Cross-sectional	Cross-country (18 countries)	IV DiD	Evaluate the effect of class size on student performance.
2006	Luyten, H.	TIMSS 1995	Cross-sectional	Cross-country (8 countries)	RD	Analyze the effect of having received an extra year of schooling on student performance

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2006	Bedard, K. Dhuey, E.	TIMSS 1995 TIMSS 1999	Pooled data	Cross-country (10 countries)	IV	Examine the impact of maturity differences on student performance pooling data from different datasets
2006	West, M.R. Woessmann, L.	TIMSS 1995	Cross-sectional	Cross-country (18 countries)	IV DiD	Examine whether the sorting of differently achieving students into differently sized classes results in a different pattern of class sizes
2006	Ammermüller, A. Dolton, P.	TIMSS 1995 TIMSS 1999 TIMSS 2003 PIRLS 2001	Pooled data	England United States	DiD	Investigate the potential existence of pupil-teacher gender interaction effects on performance
2008	Schneeweis, N. Winter-Ebmer, R.	PISA 2000 PISA 2003	Cross-sectional	Austria	DiD	Evaluate the impact of schoolmates (peer effects) on students' academic outcomes
2008	Luyten, H. Peschar, J. Coe, R.	PISA 2000	Cross-sectional	England	RD	Assess the effects of one year of schooling on reading performance, reading engagement, and reading activities
2008	Puhani, P.A. Weber, A.M.	PIRLS 2001	Cross-sectional	Germany	IV	Assess the effect of age of school entry on educational outcomes
2009	Ammermüller, A. Pischke, J.S.	PIRLS 2001	Cross-sectional	Cross-country (6 countries)	DiD	Estimate peer effects for students exploiting variation across classes within schools

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2009	Schütz, G.	PISA 2003	Cross-sectional	Cross-country (41 countries)	DiD	Analyze the impact of the attendance of pre-primary institutions on student performance at age 15
2010	Perelman, S. Santín, D.	PISA 2003	Cross-sectional	Spain	IV	Analyze the effect of the attendance to private and public schools on the level of efficiency estimated for students using parametric stochastic distance functions
2010	Jakubowski, M.	PIRLS 2001 TIMSS 2003 PISA 2000 PISA 2003	Pooled data	Cross-country (23 countries)	DiD	Assess the effects of tracking on students´ performance
2010	West, M.R. Woessmann, L.	PISA 2003	Cross-sectional	Cross-country (29 countries)	IV	Study the relationship between private school competition and student performance historical pattern as a natural experiment
2010	Dronkers, J. Avram, S.	PISA 2000 PISA 2003 PISA 2006	Pooled data	Cross-country (26 countries)	PSM	Estimate the effectiveness of private schools on reading achievement
2010	Lee, J. Fish, R.M.	TIMSS 1995 TIMSS 1999 NAEP 1996 NAEP 2000	Cross-sectional	United States Canada Cyprus Czech Republic Japan Korea Singapore	IV	Examine the value-added school effects considering the sources of variations in nation- and state-level growth of average math achievement

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2011	Schwerdt, G. Wuppermann, A	TIMSS 2003	Cross-sectional	United States	DiD	Investigate the impact of different teaching strategies on student achievement
2011	Pfeffermann, D. Landsman, V.	PISA 2000	Cross-sectional	Ireland	IV PSM	Assess whether private schools offer better quality of education than public schools
2011	Luyten, H. Veldkamp, B.	TIMSS 1995	Cross-sectional	Cross-country (15 countries)	IV	Assess the effect of schooling with cross-sectional data in order to identify different achievements between grades
2011	Jensen, P. Rasmussen, A.W.	PISA 2000 PISA-ethnic 2005	Matched data	Denmark	IV	Study the effect of immigrant concentration in schools on the educational outcomes
2012	Cho, I.	TIMSS 1995 TIMSS 1999 TIMSS 2003 TIMSS 2007	Pooled data	Cross-country (15 countries)	DiD	Assess the impact of teacher–student gender matching on academic achievement
2012	Ammermuller, A.	PIRLS 2001 PISA 2000	Pooled data	Cross-country (14 countries)	DiD	Investigate the relationship between cross-country differences in educational opportunities and institutional features of schooling systems

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2012	Agasisti, T. Murtinu, S.	PISA 2006	Cross-sectional	Italy	PSM	Investigate the effects of perceived competition among schools on their performance in mathematics
2012	Choi, A. Calero, J. Escardíbul, O.	PISA 2006	Cross-sectional	Korea	IV	Evaluate the impact of time spent on private tutoring on the performance of students
2013	Hanushek, E.A. Link, S. Woessman, L.	PISA 2000 PISA 2003 PISA 2006 PISA 2009	Pooled data	Cross-country (42 countries)	DiD	Analyze the effect of school autonomy on student achievement using a cross-country panel dataset
2013	Denny, K. Oppedisano, V.	PISA 2003	Cross-sectional	United States United Kingdom	IV	Estimate the marginal effect of class size on educational attainment of students
2013	Cornelisz, I.	PISA 2006 PISA 2009	Cross-sectional	Netherlands	PSM IV	Assess the causal effects of private- and public school attendance on student achievement
2013	Falck, O. Woessmann, L.	PISA 2006	Cross-sectional	Cross-country (27 countries)	IV	Estimate the effect of private-school competition on students´ occupational intentions
2013	Gustafsson, J.E.	TIMSS 2003 TIMSS 2007	Cross-sectional	Cross-country (22 countries)	IV DiD	Investigate the effects of time spent on homework on mathematics achievement

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2013	Kiss, D.	PIRLS 2001 PISA 2003	Cross-sectional	Germany	DiD	Examine grade discrimination in primary and secondary schools for immigrants and girls
2013	Gamboa, L., Rodríguez, M. García, A.	PISA 2006	Cross-sectional	Cross-country	IV	Analyze the effect of pupils' self-motivation on academic achievement in science across countries.
2013	Ponzo, M.	PIRLS 2006 TIMSS 2007	Cross-sectional	Italy	PSM	Examine the effect of being a victim of school bullying on educational achievement
2014	Bietenbeck, J.	TIMSS 2007	Cross-sectional	United States	DiD	Evaluate the effects of traditional and modern teaching practices on different cognitive skills
2014	Piopiunik, M.	PISA 2000 PISA 2003 PISA 2006	Pooled data	Germany	DiD	Analyze the effects of early tracking on student performance
2014	García-Perez, J.I. Hidalgo-Hidalgo, M. Robles-Zurita, J.A.	PISA 2009	Cross-sectional	Spain	IV	Examine the effect of grade retention on academic performance

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2014	Crespo-Cebada, E. Pedraja-Chaparo, F. Santín, D.	PISA 2006	Cross-sectional	Spain	PSM	Evaluate the impact of school ownership on the technical efficiency of Spanish schools
2014	Ponzo, M. Scoppa, V.	PIRLS 2006 TIMSS 2007 PISA 2009	Pooled data	Italy	IV	Investigate whether the age at school entry affects students´ performance
2014	Hanushek, E. A., Piopiunik, M. Wiederhold, S	PIAAC 2011/12 PISA 2009 PISA 2012	Matched and pooled data	Cross-country (23 countries)	OLS IV DiD	Exploring the role of teachers´ cognitive skills in explaining students´achievement
2014	Lee, B.	PISA 2009	Cross-sectional	Austria Italy	PSM	Compare the effect of academic and vocational tracks on students' educational expectations
2014	Gee, K. Cho, R.M.	TIMSS 2011 KELS 2005	Cross-sectional	Korea	PSM	Identify the effects of single-sex versus coeducational schools on adolescent aggressive behaviors
2014	Konstantopoulos,S. Traynor, A.	PIRLS 2001	Cross-sectional	Greece	IV	Assess the class size effects on student performance in reading
2015	Rivkin, S.G. Schiman, J.C.	PISA 2009	Cross-sectional	Cross-country (72 countries)	DiD	Analyze the link between achievement and instructional time taking into account as well the quality instruction as well as the classroom environment

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2015	Lavy, V.	PISA 2006	Cross-sectional	Cross-country (50 countries)	DiD	Estimate the effects of instructional time on students´ achievement
2015	Vardardottir, A.	PISA 2003	Cross-sectional	Switzerland	DiD	Assess the influence that socio-economic status of class peers has on academic outcomes of students
2015	Anghel, B. Cabrales, A. Sainz, J. Sanz, I.	PISA 2000 PISA 2003 PISA 2006 PISA 2009	Pooled data	Spain	DiD	Analyze the impact of high-quality public childcare on children's cognitive performance
2015	Tiumeneva, Y. A. Kuzmina, J. V.	PISA 2009	Cross-sectional	Russia Czech Republic Hungary Slovakia Germany Canada Brazil	RD	Evaluate the effectiveness of one year of schooling on student achievement in reading
2015	Jiang, F. McComas, W.F.	PISA 2006	Cross-sectional	46 countries (separately)	PSM	Examine the effects of the level of openness of inquiry teaching on student science achievement and attitudes
2015	Edwards, S. García-Marín, A.	PISA 2012	Cross-sectional	Cross-country (61 countries)	IV	Investigate whether the inclusion of educational rights in political constitutions affects the quality of education

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2015	Lavrijsen, J. Nicaise, I.	PIRLS 2006 PISA 2012	Pooled data	Cross-country (33 countries)	DiD	Study how postponing the age of tracking in some countries may reduce the strength of the association between social background and achievement
2015	Ruhose, J. Schwerdt, G.	PIRLS 2001, 2006 TIMSS 1995, 1999, 2003, 2007, 2011 PISA 2000, 2003, 2006, 2009, 2012	Pooled data	Cross-country (45 countries)	DiD	Analyze the effect of tracking controlling for unobserved differences in the characteristics of the migrant and native students
2015	Jakubowski, M.	PISA 2006	Cross-sectional	Poland	PSM	Analyze differences in the magnitude of student progress across two types of upper secondary education
2015	Felfe, C. Nollenberger, N. Rodríguez-Planas, N.	PISA 2003 PISA 2006 PISA 2009	Pooled data	Spain	DiD	Estimate children's long-run cognitive development when introducing universal high quality childcare for 3-year olds
2016	Hogrebe, N. Strietholt, R.	PIRLS 2011	Cross-sectional	Cross-country (9 countries)	PSM	Assess the effect of preschool non-participation on reading literacy at the end of primary school

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2016	Lavrijsen, J. Nicaise, I.	PIRLS 2001 PIRLS 2006 PIRLS 2011 TIMSS 2007 TIMSS 2011 PISA 2006 PISA 2009	Pooled data	Cross-country (23-35 countries)	DiD	Examine the effects of the age at which tracking occurs on student achievement.
2016	Green, A. Pensiero, N.	PISA 2000 SAS 2011-12	Pooled data	Cross-country (21 countries)	DiD	Assess the contribution of upper-secondary education and training to inequalities in skills opportunities and outcomes
2016	Pedraja-Chaparro, F. Santin, D. Simancas, R.	PISA 2003 PISA 2009	Pooled data	Spain	DiD	Evaluate the impact of immigrant concentration in schools on student performance
2016	Kuzmina, J. Carnoy, M.	PISA 2012	Cross-sectional	Austria Croatia Hungary	IV RD	Examine the relative labor market value of vocational and academic education on educational outcomes
2016	Isphording, I. E. Piopiunik, M. Rodríguez-Planas, N.	PISA 2003 PISA 2006 PISA 2009 PISA 2012	Pooled data	Cross-country (16 countries)	IV	Evaluate the effect of immigrant students on reading performance on their math performance
2016	Arikan, S. van de Vijver, F. Yagmur, K.	TIMSS 2007 TIMSS 2011	Cross-sectional	Cross-country (Turkey and Australia)	PSM	Identify factors to predict mathematics achievement of Turkish students in comparison to Australian students.

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
2016	Konstantopoulos,S. Shen, T.	TIMSS 2003 TIMSS 2007	Cross-sectional	Cyprus	IV RD	Examine the association between class size and mathematics achievement in public schools
2016	Li, W. Konstantopoulos,S.	TIMSS 2011	Cross-sectional	Cross-country (14 countries)	IV RD	Examine class size effects on fourth-grade mathematics achievement
2016	Cattaneo, M.A. Oggenfuss, C. Wolter, S.C.	PISA 2009	Cross-sectional	Switzerland	DiD	Examine the causal impact of instruction time on student test scores