TESIS DOCTORAL

INFERENCIA CAUSAL EN EDUCACIÓN CON BASES DE DATOS INTERNACIONALES: APLICACIONES SOBRE EL EFECTO DE LAS ESTRATEGIAS DOCENTES

CAUSAL INFERENCE IN EDUCATION WITH INTERNATIONAL LARGE-SCALE ASSESSMENTS: APPLICATIONS ON THE EFFECT OF TEACHING PRACTICES

VÍCTOR CRISTÓBAL LÓPEZ

PROGRAMA DE DOCTORADO EN ECONOMÍA Y EMPRESA

2018

UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

INFERENCIA CAUSAL EN EDUCACIÓN CON BASES DE DATOS INTERNACIONALES: APLICACIONES SOBRE EL EFECTO DE LAS ESTRATEGIAS DOCENTES

CAUSAL INFERENCE IN EDUCATION WITH INTERNATIONAL LARGE-SCALE ASSESSMENTS: APPLICATIONS ON THE EFFECT OF TEACHING PRACTICES

VÍCTOR CRISTÓBAL LÓPEZ

PROGRAMA DE DOCTORADO EN ECONOMÍA Y EMPRESA

Conformidad de los Directores:

Fdo: José Manuel Cordero Ferrera

Fdo: María Gil Izquierdo

2018

AGRADECIMIENTOS

Parece que fue ayer cuando todo comenzó. En el año 2013 me sentía con la fuerza y motivación necesarias para emprender semejante aventura, mi Tesis Doctoral. Desde entonces, multitud de acontecimientos han acompañado a este proyecto: me he mudado a Alemania, me he casado, he sido padre de un maravilloso hijo, etc. Pero a día de hoy, unos cuantos años después, con sus idas y venidas, altos en el camino y dimes y diretes varios, ha llegado el momento de concluir esta etapa. Ahora, llegado el momento de recoger los frutos de lo sembrado durante estos últimos tres años y medio, es la hora de agradecer a todas y cada una de las personas que me han apoyado y a las que me gustaría dedicar esta Tesis. Sin ellos de ningún modo podría estar en estos momentos escribiendo estas palabras.

En primer lugar, quisiera dar las gracias enormemente a José Manuel Cordero. Como Director de Tesis, me ha apoyado desde el primer minuto. Desde el primer instante he contado con su confianza a ciegas en la consecución de este proyecto. Desde luego, sin su ayuda, todo esto no hubiera sido posible. Le agradezco por su paciencia y dedicación absoluta, por compartir conmigo sus conocimientos sobre economía de la educación, sobre técnicas de inferencia causal, y sobre todo por su profesionalidad y su ilusión en este proyecto. Su constancia ha sido un factor clave durante los muchos momentos de flaqueza que he sufrido a lo largo de este tiempo; su dedicación, motivación e inspiración han sido hasta el bocinazo final la clave del éxito de esta tesis. De él he aprendido multitud de facetas del ámbito académico que me resultaban absolutamente desconocidas hasta la elaboración de esta Tesis. Mil gracias por confiar en mí en todo momento, por enseñarme el oficio de investigador y por tu comprensión durante mis innumerables crisis existenciales/laborales. Me he sentido un auténtico privilegiado por poder compartir contigo esta experiencia.

En segundo lugar quiero dar la gracias a mi Codirectora de Tesis. Si antes he destacado la paciencia que José Manuel ha demostrado durante estos años, qué puedo decir ahora de María Gil. Gracias, gracias de veras por tu aguante, por tu entusiasmo, por tu confianza en mis posibilidades y sobre todo por el nivel de exigencia y compromiso que me has pedido en este periodo. Sin tu ayuda, esta empresa tampoco hubiera llegado a buen puerto. Desde nuestra primera reunión en el campus de la Universidad Autónoma de Madrid muchos han sido los debates, las réplicas, las polémicas y los callejones sin salida que hemos ido experimentando durante este tiempo; muchos altibajos, muchas frustraciones con los datos y las estimaciones, muchos errores y algunos aciertos. Pero todos esos momentos, a día de hoy, cobran sentido y hacen que todo haya merecido la pena. Sin duda, este proceso me ha permitido crecer tanto personal como profesionalmente y mucho, muchísimo de ello te lo debo a ti. Una vez más, gracias.

Especialmente dedico esta Tesis a mis padres, Mº Teresa y Agustín, porque sin ellos no hubiera conseguido realizar de ningún modo este trabajo. Quiero darles las gracias por fomentar desde muy pequeñito mi formación académica en todo momento. Pero sobre todo, quiero agradecerles por la educación que nos han dado, por su amor y su apoyo constante. Constantemente me han animado a emprender nuevos retos, como esta Tesis, me han enseñado a esforzarme para conseguir los objetivos que me he ido proponiendo a lo largo de estos años, y a aprovechar las oportunidades que la vida nos brinda. Gracias por todo. Siempre me habéis servido de ejemplo a seguir. Sin duda alguna, esta Tesis es en gran parte también vuestra. Os agradezco el estar a mi lado en todo momento, incluso a pesar de la distancia que nos separa desde hace unos años. Siempre he sentido vuestro reconocimiento, hiciera lo que hiciese, y en este sentido no encuentro otra manera más digna de mostraros mi gratitud que dándoos las gracias por como sois y por cómo nos habéis criado. Gracias, gracias y mil gracias.

Gracias a mi hermana Patricia por su apoyo durante este viaje. Por estar siempre ahí cuando te he necesitado. Tu ayuda, tu comprensión y tus consejos han sido fundamentales para la realización de esta Tesis. Gracias por tu fe ciega en mis posibilidades. Gracias porque siempre me has aportado cordura y sensatez, sobre todo en las innumerables ocasiones en las que no veía la forma de avanzar en la dirección adecuada. Tu forma de ser me ha servido en todo momento de fuente de inspiración, motivación y de guía para llegar al final de este proyecto. Gracias, Patri.

Por supuesto quiero dar la gracias a mi extraordinaria mujer, Friederike, y a mi maravilloso hijo, Mateo, por su apoyo incondicional en todo momento. Sin ellos, no hubiera sido capaz de acometer el semejante reto que una Tesis doctoral supone. Millones de gracias por vuestro amor, vuestra paciencia, compresión, sacrificio y dedicación durantes estos cuatro años. Sin vuestras constantes palabras de ánimo, en ningún momento hubiera podido mantener la motivación y la perseverancia necesaria para realizar esta Tesis. Mil disculpas por las horas de ausencia, por los innumerables días de biblioteca durante los cuales hemos tenido que sacrificar nuestro tiempo juntos por esa entelequia llamada Tesis doctoral. Hoy por fin, ese concepto abstracto toma forma, llega a su fin y con ello prometo recuperar el tiempo perdido.

De igual manera quiero dar las gracias a mi enorme familia: primos, primas, tíos, tías, a mis abuelas, Eugenia y Tomasa, así como a los que ya no están entre nosotros pero a lo que por supuesto pertenece, sin duda alguna, este éxito: mis abuelos Agustín y Jesús. De diferentes maneras, pero por parte de cada uno de ellos siempre he recibido una muestra de apoyo. Gracias, sois maravillosos.

No me olvido de todos y cada uno de los amigos que me han ayudado en este periodo tan importante de mi vida. No pretendo hacer mención a todos y cada uno de ellos, pero sí me gustaría destacar el papel que desempeñaron en los orígenes de esta Tesis doctoral mis amigas Graciela y Raquel. Gracias porque sin vuestra ayuda nunca hubiera conseguido poner en marcha este proyecto. Asimismo, aprovecho esta oportunidad para dar las gracias a Jose, Elena, Tania, Alejandro, Elena "IET" y Miguel. De manera muy diversas, pero todos sois partícipes de alguna forma en la consecución de este trabajo. Mil gracias chicos, sois fenomenales, ¡qué suerte teneros a mi lado! Del mismo modo no puedo olvidar a mi amigo de toda la vida, Beni. Gracias, amigos.

Finalmente, me gustaría a agradecer a Dieter Dohmen, director del Forschungsinstitut für Bildungs- und Sozialökonomie (FiBS) en Berlín, por darme la oportunidad de colaborar en múltiples proyectos en el ámbito de la economía de la educación durante dos intensos años. Del mismo modo quiero dar las gracias a Daniel Santín por su colaboración en el primer capítulo de esta Tesis. Su aportación supuso una mejora significativa y fundamental en la consecución del texto y la posterior aceptación para su publicación. Muchas gracias, Dani. También quiero agradecer a todas los colegas que en los distintos congresos en los que se han presentado los capítulos que componen esta

tesis han aportado su granito de arena a ella a través de comentarios, sugerencias, críticas, etc. Gracias por vuestra ayuda.

Y por supuesto, también quiero dar la gracias al Instituto Nacional de Educación Educativa (INEE) por poner a nuestra disposición los datos correspondientes a la EGD 2010, sin los cuales la consecución del tercer capítulo de esta Tesis no hubiera sido posible.

A mi Familia, Frie y Mateo

CONTENTS

List of figures xiii
List of tables xv

INTRODUCTION 1

CASUAL INFERENCE ON EDUCATION POLICIES: A SURVEY OF EMPIRICAL STUDIES USING PISA, TIMSS AND PIRLS 9

1.1 Introduction 13
1.2 Literature review and search strategy 15
1.3 Methods 17
1.4 Empirical studies review 23
1.4.1 Instrumental Variables 23
1.4.2 Regression discontinuity designs 28
1.4.3 Difference in differences 30
1.4.4 Propensity Score Matching 37
1.5 Summary of empirical studies 40
1.6 Conclusions 45
ANNEX I 48

TEACHING PRACTICES AND RESULTS IN PISA 2015 61

2.1 Introduction 65
2.2 Literature review 67
2.3 Data and Variables 68
2.4 Results 73
2.5 Conclusions 77

ESPECIALIZACIÓN EN LAS ESTRATEGIAS DOCENTES Y EFECTOS SOBRE LOS RESULTADOS ESCOLARES 79

3.1 Introducción 81
3.2 Revisión de la literatura 84
3.3 Datos y variables 87
3.4 Metodología 95
3.5 Resultados 97
3.6 Conclusiones 100

TEACHING STRATEGIES AND THEIR EFFECT ON STUDENT ACHIEVEMENT: A CROSS-COUNTRY STUDY USING DATA FROM PISA 2015 103

4.1 Introduction 105
4.2 Literature review 106
4.3 Data and variables 108
4.4 Empirical strategy 114
4.5 Results 116
4.6 Conclusions 121

CONCLUDING REMARKS 123

REFERENCES……..…127

List of figures

Figure 1.1 Number of empirical studies (2004-2016)..... 40
Figure 1.2 Datasets used in empirical studies 41
Figure 1.3 Methods used in empirical studies..... 42
Figure 1.4 Topics examined in applications..... 43
Figure 1.5 Subject categories of published papers..... 44
Figure 1.6 Distribution of papers across quartile rankings according to impact factors 45
Figura 3.1. Distribución de los indicadores representativos de las estrategias docentes 90
Figura 3.2 Clasificación de centros entre top-innovador (gris) y el resto (blanco)..... 91
Figura 3.3. Clasificación de centros entre top-clásico (gris) y el resto (blanco).....92
Figure 4.1 Mean values of the teaching indices by countries 113

List of tables

Table 1.1 Causal inference methods applied on international educational databases. 22
Table 1.2 Empirical studies using causal inference with data from international large scale Assessments 48
Table 2.1 Definition of teacher variables and classification according to Criterion 1 (cognitive, active and teacher-led) and Criterion 2 (classical and modern)..... 71
Table 2.2 Descriptive statistics of the variables included in the model 73
Table 2.3 Estimation of student achievement for science depending on teaching strategies..... 74
Table 2.4 Estimation of the relationship between science outcomes and teaching strategies by means of multilevel quantile regressions 76
Tabla 3.1 Distribución de variables entre las estrategias docentes clásicas e innovadoras 89
Tabla 3.2 Estadísticos descriptivos de las variables incluidas en el modelo..... 93
Tabla 3.3 Estadísticos descriptivos de las variables incluidas en el modelo por tipo de centro..... 94
Tabla 3.4 Efectos marginales del modelo probit para los centros “top-innovadores” 98
Tabla 3.5 Efecto del tratamiento promedio de centros “top-innovadores”..... 99
Tabla 3.6 Efectos marginales del modelo probit para los centros top-clásicos..... 99
Tabla 3.7 Efecto de tratamiento promedio de centros “top-clásicos”..... 100
Table 4.1 Dataset composition..... 109
Table 4.2 Classification of variables about teaching practices (traditional and modern) 111
Table 4.3 Distribution of total variance of teaching indices between and within schools..... 112
Table 4.4 Descriptive statistics..... 114
Table 4.5 Estimates using a least squares multi-level regression approach..... 117
Table 4.6 Estimates using a student fixed effects model 119
Table 4.7 Results of interquartile regressions 120
Table 4.8 Estimates using a student fixed effects model: teaching strategies interactions with school variables 121

INTRODUCTION

The identification of factors associated with students´ attainment has been a main concern for researchers in the field of education over the last decades. This interest is driven by the extensive existing evidence demonstrating that there is a high correlation between the quality of education, measured frequently by the skills demonstrated by students in international tests, and the level of development and economic growth of countries (Krueger and Lindahl, 2001; Barro, 2001; Hanushek and Woessman, 2011). The study of the economic effects of education has generated great interest since the development of the theory of human capital (Schultz, 1960; Becker, 1964). According to this theory, education can be considered as a form of investment whose profitability will depend on the effects of education in terms of higher productivity, which will result in a higher probability of employment, better working conditions and higher salaries throughout working life.

The study of the factors that affect educational performance has led to the development of a line of research dedicated to the study of the educational production function, in which the primary objective is to determine which factors have influence on results and the direction (positive or negative) of that influence. This trend was initiated more than five decades ago with the publication of the well-known Coleman Report (Coleman et al., 1966), in which one of the most significant finding was that school resources explained a small proportion of the results obtained by students, while students´ background was found to be the most the most important factor to predict students´ educational success. However, in later studies, many authors have also identified teacher quality as a key determinant of student performance (Hanushek, 1971; Hughes 1999; Darling-Hammond 2000; Hattie, 2003; Rockoff, 2004). This is a multidimensional construct that includes many different aspects, such as their qualifications or skills, abilities to communicate or their beliefs and attitudes towards students and teaching (Wayne & Youngs, 2003; Palardy and Rumberger, 2008; Yeh, 2009).

In this thesis we focus our attention on teaching practices, which involves what teachers actually do in their classroom with their students, including organization of instructional time and educational resources as well as specific activities and strategies proposed and adapted to students´ characteristics. Some previous studies have demonstrated that teaching practices have a significant influence on student's achievement (Schacter and

Thum, 2004), although the evidence is still scarce and not conclusive with regard to the identification of the most effective teaching strategies. This strand of literature has aimed to look inside the "black box" of instructional quality and explore the divergences between the more traditional teaching styles such as teacher-centred instruction and mastery learning, mainly based on lecturing, memorization and repetition, and the emerging modern practices inspired by constructivist approaches involving student-oriented teaching and self-regulated student activity (Seidel and Shavelson 2007; Van de Grift, 2014). Despite the evidence available on the impact of different teaching strategies on academic performance is still contradictory and inconclusive, educational authorities in many countries advocate a greater use of those modern practices in detriment of more traditional methods (Capps et al., 2012). Therefore, the study of the effectiveness of different teaching styles represents a topic of great relevance from a policy perspective.

This line of research has benefited from the development of several international datasets that collect comparable data at student level from educational systems around the world like TIMSS (*Trends in Internacional Mathematics and Science Study*), PIRLS (*Progress in International Reading Literacy Study*) or PISA (*Programme for International Student Assessment*). These databases contain a large volume of information related to the competencies or skills demonstrated by students in different domains (e.g. mathematics, reading and science) as well as multiple factors that can affect their attainment, including data about teachers. Until very recently, most part of research on teacher effects with international datasets had used data from TIMSS (Mullis et al., 2012), since this was the only survey that provided data on students, teachers and schools. For instance, Schwerdt and Wuppermann (2011) and Van Klaveren (2011) used TIMSS 2003 data for US and Netherlands, respectively, to examine the influence of teaching practices on student achievement. House (2009) and Bietenbeck (2014) analyze the effect of different types of instruction using data from TIMSS 2007 for fourth-grade students in Japan and US eight-grade students, respectively. Zuzovsky (2013) and O´Dwyer et al. (2015) also explore the relationship between instructional practices and eighth grade students´ performance using data from TIMSS 2007 in a cross-country approach. Finally, the recent book edited by Nilsen and Gustafsson (2016) is a valuable contribution to this growing body of research, since it contains several empirical studies analysing TIMSS data across different countries.

Those empirical studies are focused on primary education, where one single teacher is responsible for all the curricula, thus the analysis is focused on individual styles of teaching. In contrast, this PhD thesis aims to make a practical contribution by providing solid empirical evidence on the effects of different teaching styles applied by teachers in secondary schools. In this education level, it is frequent that teachers at the same school are prone to use similar teaching strategies and even share the same teaching materials (Le Donné et al. 2016), developing what is known in the literature as a "teaching culture" (Echazarra et al. 2016). Thus, we explore how student achievement can be influenced by this culture, which represents an innovative approach with respect to previous studies focused on activities carried out by each teacher individually.

Most part of the existing research exploring the effects of teaching practices is based on traditional estimation strategies, i.e. ordinary least squares (OLS) or multilevel regressions. However, these approaches present serious limitations that could lead to biased results (Hanushek, 1979; Todd and Wolpin, 2003). Basically, they are related to the well-known endogeneity problem. This occurs when unobserved variables affecting the dependent variable are correlated with explanatory variables included in the regression. For instance, it could be argued that student performance might be determined by teachers´ motivation. However, motivation is very difficult to measure and, in addition, it can depend on the quality of students; thus, a reverse causality problem might arise in this case. Something similar might occur when information about students' prior achievement is missing and therefore it is difficult to disentangle whether a certain teacher is using a particular strategy because she has high-achieving students, or whether that strategy is producing high-achieving students (O´Dwyer et al., 2015). This problem might be even more challenging considering that sorting of students to teachers might not be non-random (Rothstein, 2010), since children from families with greater economic and cultural capital are likely to attend schools with better resources. Therefore, there can be a problem of unobserved heterogeneity in data that complicates the econometric estimation of teacher effects (Gustafson, 2013).

In order to deal with these problematic issues several quasi-experimental methods have been developed in recent literature. Among them, the most common estimation techniques used to estimate causal effects are Difference-in-Differences (DiD), Propensity Score Matching (PSM), Instrumental Variables (IV) and Regression

Discontinuity (RD). The use of this type of techniques in the educational context has experienced a remarkable growth in recent years (Webbink, 2005; Schlotter et al., 2011; Hanushek and Woessmann, 2014), although its implementation requires fulfilling a series of requirements that are often difficult to meet when available data comes from large-scale assessments with a cross-section structure.

As an initial approximation of the usefulness of these methods, in the first chapter of this thesis we provide a review on how these approaches can be applied to cross-sectional data like those available in international large-scale assessments in education. This complete literature review covers research published works from 2004, when these techniques started to be applied in educational contexts, to 2016. In addition, this chapter provides a description and a critical evaluation of these approaches. The main purpose is to provide an overview about the estimation strategies that can be adopted in order to overcome the frequent problem of endogeneity in these data. The compiled studies have been divided according to the method applied and, subsequently, they have been clustered depending on the issue studied (e.g. type of ownership, teachers, tracking, etc.). The analysis performed allows us to identify that some limitations in terms of data structure and design of a valid counterfactual group for specific topics remain still unaddressed.

Chapter 2 is conceived as a preliminary investigation that serves as a starting point for analyzing the effects of teaching practices. In this chapter we apply standard regression methods as a previous step before using more sophisticated approaches. We exploit recent data from PISA 2015, which for the first time provides information on teachers through a specific teacher's questionnaire. Our interest focuses on examining the influence of teaching strategies in the specific Spanish context. The empirical strategy relies on a multilevel regression model, considering that students are grouped at a higher level represented in this case by schools. Additionally, a quantile regression model (for quartiles) also with a multilevel structure is estimated in order to control for differences across the distribution of results.

Subsequently, in chapters 3 and 4, we apply two of the aforementioned quasi-experimental methods to examine the effects of different styles of teaching implemented by secondary education Specifically, in chapter 3 we use the propensity score matching method, while in chapter 4 we apply student fixed effects, which can be understood as

an extension of the DiD approach. By conducting this research, we highlight that applying these approaches demand careful statistical designs, additional data transformation and numerous checks to enable reliable estimation results in terms of casual inference.

In chapter 3 we focus again on the specific case of Spain using the information available in the General Diagnostic Evaluation (GDE) survey conducted by the National Institute for Educational Evaluation (INEE) in secondary schools in 2010. This survey has similar characteristics to international datasets, but it has hardly been used in empirical studies so far. Some previous studies had used information from Spanish GDE survey like Santín and Sicilia (2018), Hidalgo and Lopez-Mayán (2015) or González-Betancor and López-Puig (2017) for different purposes, but they exploited data from the 2009 edition, which refers to primary education. Therefore, our study constitutes one of first empirical studies exploiting the information available in the GDE 2010 about secondary schools.

According to teachers´ responses about what they usually do in their classes we have built two indicators representing two type of teaching strategies applied by teachers in the same school: modern and traditional. Once we have those indicators, we have divided the sample of available schools in different groups according to their level of specialization in applying each type of teaching style, so that we can apply propensity score matching. Specifically, our classification is based on quartiles, so we can make a clear distinction between schools mainly focused on applying traditional methods, i.e. those placed in the first quartile for this indicator, and schools frequently applying modern practices, i.e. those situated in the first quartile for the corresponding indicator. Using this classification we can have a treated and a control group of schools, which is represented by schools belonging to the remaining quartiles (2, 3 and 4 quartiles).

In chapter 4, we use data from the last wave of the PISA survey (PISA 2015) in a cross-country analysis. Since the participation of teachers is not compulsory, only 16 out of 65 countries participating in PISA 2015 provide this information, so our empirical analysis is referred to those countries. In this case, we also construct indices representing different styles of teaching, but we adopt an estimation strategy based on applying student fixed-effects across different subjects. Moreover, covariates at student and school level are used as control variables along the estimation process. The findings

corroborate some of the previous results about the effect of different teacher strategies on student achievement. In particular, the results suggest that traditional strategies have a positive and significant effect on student performance. In contrast, modern strategies have a smaller effect, mainly concentrated in high-performers students.

Although the issue of endogeneity has been addressed in many previous studies conducted in the educational sector, the evidence concerning teaching practices and student performance is still scarce. In light of that, this research aims to be a significant contribution in order to investigate to what extent diverse teaching styles trigger student performance. Specifically, this research tackles this issue by applying innovative econometric techniques of causal inference to data from large-scale assessments to draw more meaningful and robust conclusions, which might be helpful for the design of educational policies. In the end, the results of this doctoral thesis confirm the complex relationship between teaching practices and student achievement.

Chapter 1

CASUAL INFERENCE ON EDUCATION POLICIES: A SURVEY OF MPIRICAL STUDIES USING PISA, TIMSS AND PIRLS

JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12217

CAUSAL INFERENCE ON EDUCATION POLICIES: A SURVEY OF EMPIRICAL STUDIES USING PISA, TIMSS AND PIRLS

José M. Cordero* and Víctor Cristóbal

University of Extremadura

Daniel Santín

Complutense University of Madrid

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. Causal inference; Education; International assessments; Literature review; Selection-bias

1.1. 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 some potential caveats of these databases and techniques 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.

1.2. 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.

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 1.2 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.

1.3. 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 DiD we account for an indicator variable that takes out mean differences between treated and control units so that the effect of the evaluated program or policy can be identified by the changes experienced by the other variables over time.

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 1.1 summarizes the main characteristics of these four econometric techniques, as well as their main strengths and weaknesses for their use with international databases.

Table 1.1 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 from 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. Most of times we find a fuzzy RDD.
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 individuals and control individuals that have a similar probability of being treated but are in the control group.	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 the same distribution on unobservable variables, PSM mitigates the endogeneity problem.	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 will be equally distributed in the treated and untreated groups.

It is worth highlighting here that the four methods reported in Table 1.1 can be wisely combined to enhance the fulfillment of assumptions before estimations. For example, PSM can be used as a trimming procedure to determine a common support region in the baseline observed characteristic previous to apply DiD in order to make that the parallel trends assumption is more likely to be hold. RDD relies on the assumption that treated and control units around the cut-off points are closely similar. However, if some differences between both groups remain, an alternative is to combine DiD in outcomes with RDD. Moreover, as we mention above in the discussion, fuzzy RDD can be interpreted as an IV problem in which the cut-off point defined in the running variable is used as the instrument.

1.4. 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.

1.4.1. 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 Santín (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.

We conclude this section about IV applications remarking once again that a good instrument obtained from international databases should fulfill three well-known main conditions. First, the instrument must be correlated with receiving the treatment even under the presence of other covariates; second, the instrument should be fully random in the sense that is not related with unobserved characteristics captured by the error term; and third, the exclusion restriction says that there is no other direct or indirect relationship between the instrument and the outcome but the described channel. As it is not possible to directly test whether the instrument is exogenous, a strong theoretical support for this assumption is required instead.

For this reason we think that researchers should be prone to use historic or clear sources of exogenous variations instead of principals', parents' or students' opinions where it is more likely to find alternative channels for explaining relationships leading to question the instrument and the empirical results. In any case the selected instrument must be fully justified from a theoretical point of view. Additionally, in our literature analysis we have not found two sample IV studies from the same population, where the first stage is estimated on one dataset and the reduced form on another dataset. This fact opens a research line taking into account that education international databases are quite specialized samples with many omitted variables about other more general purposes, although these treatment variables might be gathered from other datasets.

1.4.2. 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 repetition 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.

1.4.3. 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.

As we can observe DiD is perhaps the most popular approach to be used with international databases although we always have to bear in mind its two main drawbacks. First, the scarce number of waves makes difficult to test the common trends assumption so researches should justify in depth that the studied intervention was fully exogenous. Second, as it was highlighted in Bertrand et al. (2004) standard errors calculations should be computed assuming that DiD deals with serially correlated data. To avoid misleading conclusions researchers normally resort to calculate clustered standard errors. This method assumes that a large number of groups or period is available in order to have many clusters (see Angrist and Pischke, 2008, 2014 for details), but this requirement is not always possible with international databases. For this reason, DiD should provide more robust results when analysis is performed with enough clusters at regional or state level.

1.4.4. 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.

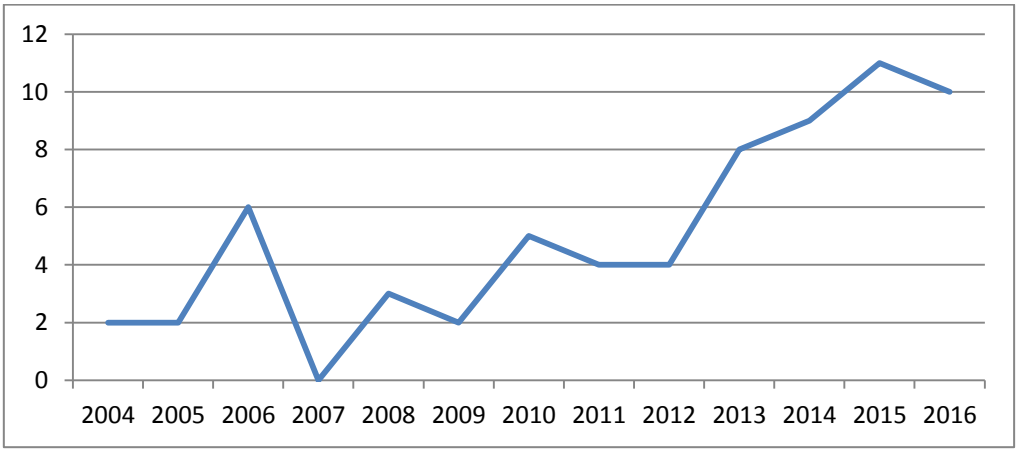
Jiang and McComas (2015) apply the PSM approach to examine the effects of the level of openness of inquiry teaching on student science achievement and attitudes using PISA data from 2006. In the context of their study, the term inquiry teaching includes very different teaching practices, all of which 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.

1.5. Summary of empirical studies

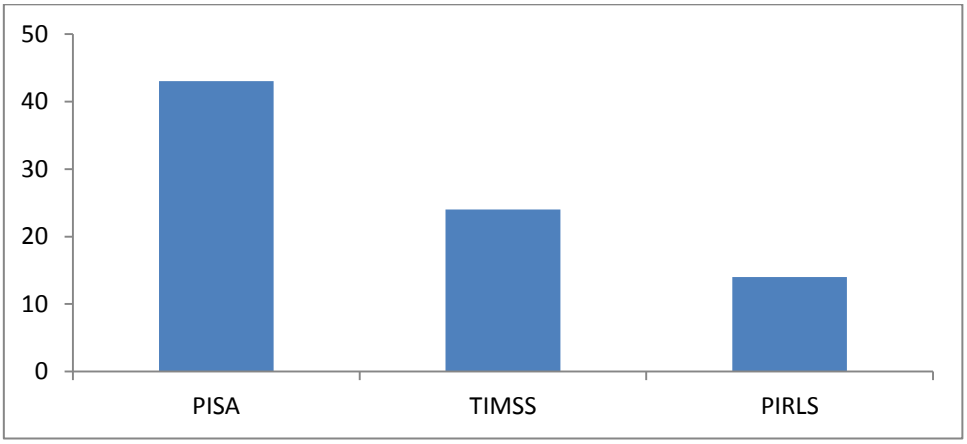
After reviewing the four approaches and the contents of all the applications, we now synthesize the main aspects of these papers and provide an overview of the journals in which they were published. From our viewpoint, this should provide sound guidance for researchers interested in combining the use 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 1.1 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 1.1 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 1.2), 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 earlier than 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.

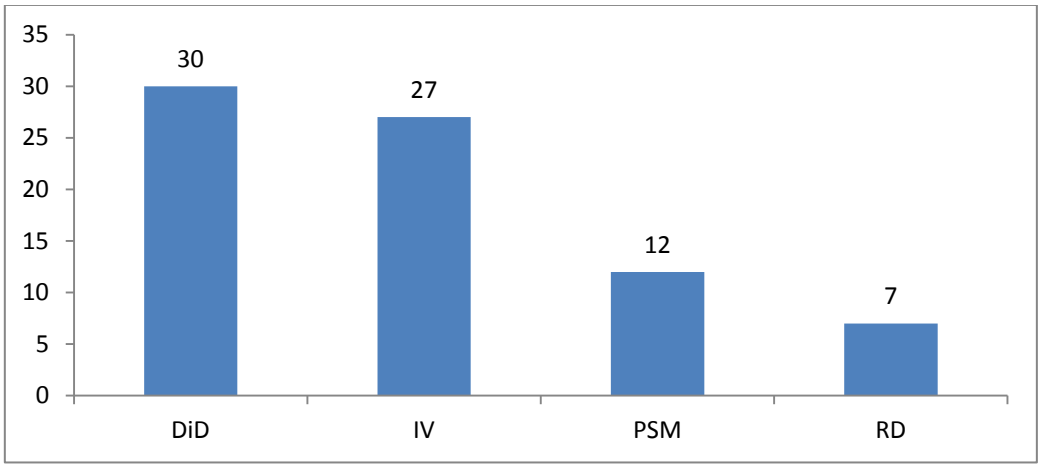
Figure 1.2 Datasets used in empirical studies



In addition, Figure 1.3 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. For this reason, the most evident drawback when using DiD with international studies is satisfying the parallel trends requirement. The fulfillment of this assumption is weak in

empirical applications because, at best, the number of repeated cross sections will be limited although the number of waves continues increasing.

Figure 1.3 Methods used in empirical studies

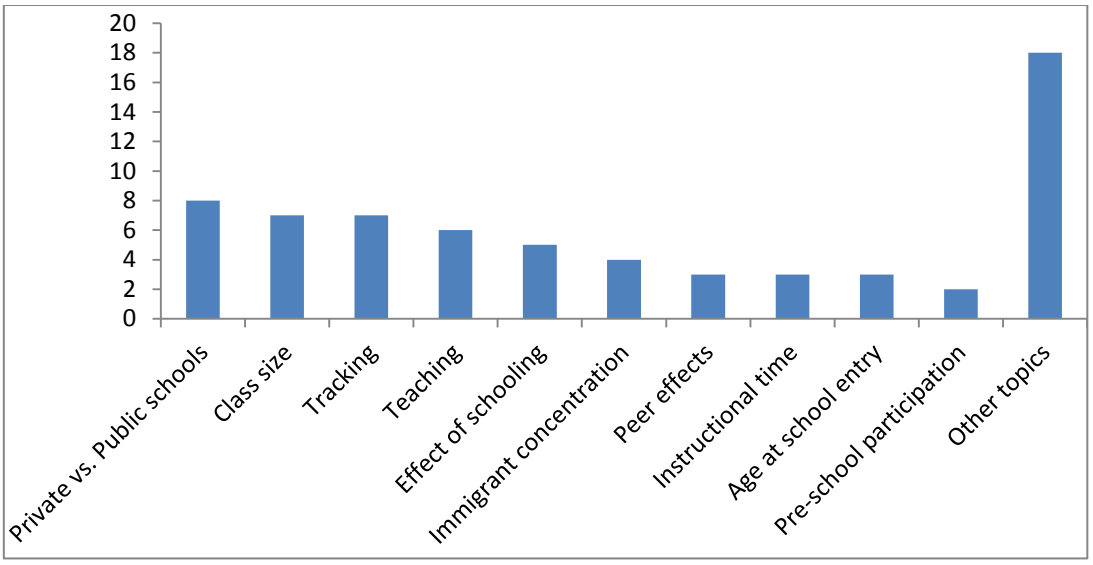


In the case of IV, all that is required for implementation is to find a good instrument that suits a particular problem and meets the required assumptions. Although this also demands some creativity on the part of the researcher, the advantage is the wide range of variables provided by large-scale assessments makes this search more likely10. However, finding a good instrument in practice is a difficult task. The menace of using weak instruments, those that only shows a very low correlation with the treatment, or the presence of non-random measurement errors in the endogenous variable might be problematic. In these two cases IV results may yield inconsistent estimations and unreliable p-values and confidence intervals (Betz, 2013).

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.

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 early tracking, clearly stand out from the rest. Other noteworthy key issues studied in several papers are the effect of different aspects related to teaching, 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 1.4 summarizes the frequency of studies applied to these topics.

Figure 1.4 Topics examined in applications

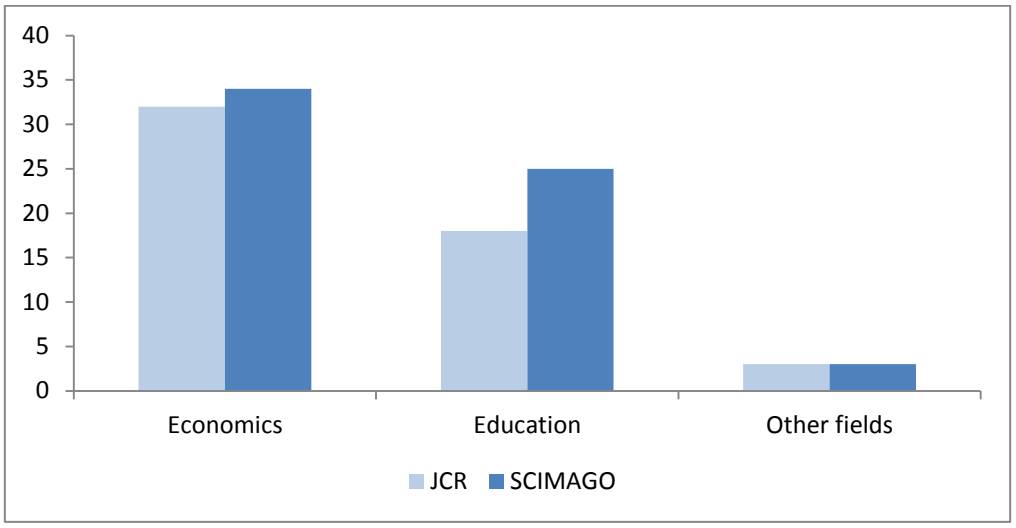


Unfortunately, there is not enough evidence yet to derive general policy implications. For the three topics with more empirical applications (private-public schools comparison, class size and tracking) we do not observe a similar pattern in their conclusions. For example, for the private-public comparison four studies identify better results in private schools, two do not find significant differences and one concludes that public schools have better results than the private ones. For class size, the conclusions are mixed as well, with one study finding beneficial effects of smaller classes on achievement and other for larger classes, although most studies (5 out of 7) do not identify any significant effect. The effect of early tracking on performance is also unclear, although most studies agree that it increases educational inequality. Therefore, we would suggest that authors should be cautious when providing specific policy recommendations about these issues based on the results from a single empirical study.

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 1.5).

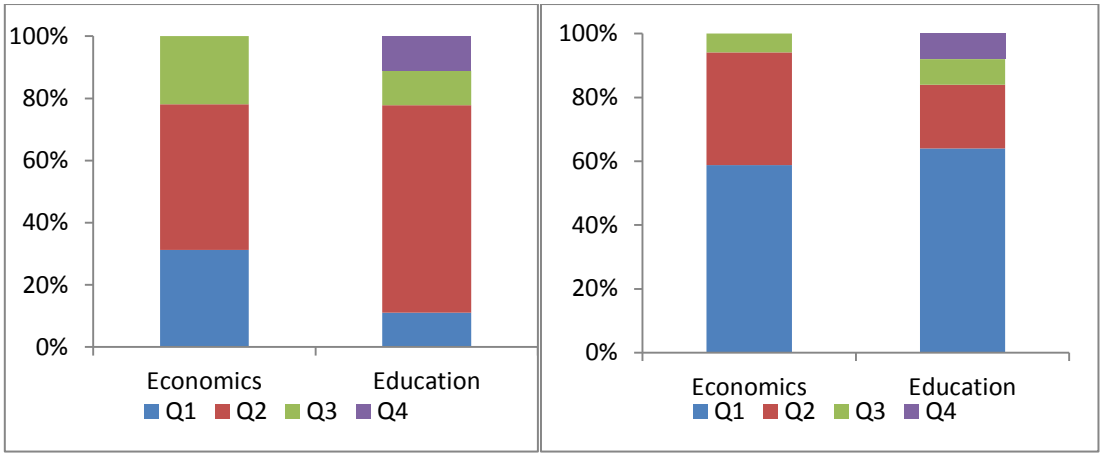
Figure 1.5 Subject categories of published papers



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 1.6 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 1.6 Distribution of papers across quartile rankings according to impact factors

(a) JCR (b) SCIMAGO



1.6. 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. Additionally, sample designs are not straightforward, implying multi-stage selection, stratification, plausible values or student and school different weights to represent population. These features should be more explicitly recognized in empirical applications when reporting standard errors or hypothesis tests.

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.

ANNEX I

Table 1.2 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.	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	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

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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
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

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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
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

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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
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.

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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
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

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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´achivement
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
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

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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. Santín, 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

Year	Authors	Datasets	Type of data	Cross-country vs. single country	Estimation method	Research question
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.
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

Chapter 2

TEACHING PRACTICES AND RESULTS IN PISA 2015[1]

[1] This chapter was originally written in Spanish in order to be submitted to the Spanish journal "*Revista de Educación*". After being accepted for publication, it was translated into English following the rules of publication of the journal. Here we only present the English text, but both versions are available in the website of the journal.

Teaching practices and results in PISA 2015[1]

Las estrategias docentes y los resultados en PISA 2015

DOI: 10.4438/1988-592X-RE-2017-379-368

María Gil Izquierdo
Universidad Autónoma de Madrid
José Manuel Cordero Ferrera
Universidad de Extremadura
Víctor Cristóbal López
Universidad de Extremadura

Abstract

Teachers undoubtedly play an important role in education. However, there is debate about which specific issues have the biggest impact on students' academic performance. This research aims to analyse different teaching strategies applied by teachers working at the same school and the extent to which they contribute to improving student achievement. Apart from the usual separation into classical and modern techniques based on pedagogical criteria, this study also distinguishes between activities that promote active student learning and cognitive activation, as well as teacher-led strategies. Our empirical analysis refers to the Spanish education system. We use data from PISA 2015, which was the first PISA wave to provide information about classroom activities based on a questionnaire completed by teachers. Our estimation procedure is based on multilevel regression techniques, taking into account that students are grouped (or nested) at a higher level represented by schools. The results suggest that the application of traditional teaching strategies, in which teachers adopt a leadership role and manage the activities taking place in the classroom, lead to a significant improvement in educational achievement. In contrast, innovative strategies

[1] This paper was developed with the financial support of the Ramón Areces Foundation and two projects funded by the Spanish Ministry of Economics and Competitiveness (ECO2014-53702-P y EDU2016-76414-R).

2.1. Introduction

Since the famous Coleman Report (Coleman et al., 1966) was published over more than fifty years ago, there has been significant interest in trying to identify the key factors determining educational performance (Creemers and Kyriakides, 2007). After several decades of research, most authors single out the importance of the role played by teachers (Rivkin el al., 2005), although there is debate about which specific issues have a greater impact on academic outcomes (Hattie, 2009; Campbell et al., 2012).

Most of the empirical research that has addressed this issue focuses on analyzing which characteristics make a teacher more successful (Akiba et al., 2007; Godhaber and Anthony, 2007). Apart from being relatively easy to measure, these characteristics are especially appealing from an educational policy viewpoint, as many are used in teacher selection processes. However, more recent literature has shown a growing interest in studying the effectiveness of teaching strategies applied by teachers (Hattie, 2009; Schwerdt and Wuperman, 2011; Lavy, 2011; Bietenbeck, 2014; Hidalgo and López-Mayan, 2015). This is part of a commendable effort to try to offer information about what happens in the classroom. Regarded as a black box that is extremely difficult to explore, this question has traditionally been overlooked by the literature.

Teaching practices or methods refer to a wide range of processes and activities that cover classroom organization and resources, as well as the activities set for students to promote their learning. These activities must be adapted to the particular classroom setting (Razak and Shafaei, 2016). The specialized literature usually makes a distinction between basically two teaching models: classical/traditional and modern (Zemelman et al., 2005). The traditional teaching style advocates activities in which the teacher takes the leading role, explaining the contents or showing how to do exercises or solve problems. These practices lay the emphasis on learning foundational knowledge and basic skills through repetition. On the other hand, more modern teaching methods are student focused involving the completion of group assignments, classroom debates or the extrapolation of knowledge learned to everyday problems (Windschitl, 2002; Opdenakker and van Damme, 2006).

Recent studies conducted by OECD specialists (Echazarra et al., 2016; Le Donné et al., 2016) offer an alternative classification that makes a distinction between three types of strategies: teacher-led learning, active learning and cognitive activation. Teacher-led learning is more or less consistent with the approach taken by traditional methods, whereas the others would be linked to modern teaching styles. Note, however, that active learning takes a more constructivist approach intended to

promote student engagement through group work, the use of new information technologies or self-assessment processes, whereas cognitive activation promotes student autonomy, establishing motivational challenges that help to stimulate higher-level competencies like critical thinking or decision-making skills (Windschitl, 2002).

There is a growing current in many countries nowadays, advocating the increased use of modern student-focused practices in preference to more traditional learning methods (Capps et al., 2012). In this context, the study of the effectiveness of the different teaching styles is a key issue with regard to the development of educational policy strategies. In particular, there is hardly any evidence available in respect of Spain. Therefore, this research is intended to contribute to a new debate based on sound empirical evidence on the role played by teachers in the context of Spanish education. Inquiries into the factors determining educational outcomes do not normally address this issue.

Researchers aiming to analyze the effectiveness of teaching methods face a major problem, namely, the shortage of reliable and relevant information on this type of practices. Until quite recently, the only international assessments that offered information in this regard were TIMSS (Trends in International Mathematics and Science) and PIRLS (Progress in International Reading Literacy Study). This is the reason why most of the papers that have examined this issue use these databases (Cordero et al., 2017). The database setup, including primary or early secondary education students, means that each student can be connected with his or her actual teacher or teachers.

The OECD has recently developed two instruments that are very useful for studying issues related to teaching staff. The first is called the TALIS-PISA link. The TALIS-PISA link provides the statistical instruments required to establish a link between the information on students in PISA 2012 and data on teachers included in TALIS 2013. The countries that decided to participate in this module sampled TALIS teachers at the same schools in which the PISA test had been conducted the year before. Accordingly, it was possible to explore the relationship between teaching strategies implemented by teachers and student achievement at the same school (Méndez, 2015; Le Donné et al. 2016), as well as the inverse relationship, that is, how the school setting affects teaching practices implemented by teachers (Austin et al., 2015). Subsequently, PISA included a questionnaire for teachers for the first time in the 2015 wave. This questionnaire was designed to gather information provided directly by teachers about a host of issues, including the teaching methods that they use in their classrooms, albeit applicable, in this case, to students and teachers at the same school in the same academic year.

The aim of this research is to leverage the information provided by science teachers for the sample of schools participating in PISA 2015 with the aim of examining how related the different teaching strategies implemented by all science teachers at each school are to students' science outcomes. In this respect, we should stress that the proposed empirical analysis assumes that teachers at the same school are inclined to use the same teaching strategies and even share the same teaching materials (Le Donné et al. 2016), developing what is known in the literature as a teaching culture (Echazarra et al. 2016). This is quite an innovative approach with respect to other earlier studies focusing on activities carried out by each teacher individually.

The content of the paper is organized as follows. Section 2 briefly reviews previous research literature concerning the role of teachers, focusing on papers that are concerned with teaching strategies. Section 3 reports the key characteristics of the database used, as well as the procedure used to build the representative indicators of teaching strategies. Section 4 presents and discusses the results of applying different multilevel regressions. As usual, the paper ends with some conclusions.

2.2. Literature review

Some of the issues concerning the influence of teacher quality on achievement that have received most attention in the literature are the characteristics of teaching staff and teaching practices developed in the classroom (Palardy and Rumberger, 2008).

The first issue focuses on analyzing the effect of teachers' cognitive skills, experience level, sex or qualifications (Ehrenberg and Brewer, 1994; Clotfelter et al., 2007; Clotfelter et al., 2010). However, the available evidence about the impact of these characteristics on academic achievement is weak (Hanushek, 1986, 1997), especially for information is sourced from international databases, where the failure to use of common indicators to correctly measure qualifications or experience can lead to rather inconclusive results (Hanushek, 2011). The only definite result is that teachers' performance usually improves during the early years of their teaching experience (Rockhoff, 2004; Croninger et al., 2007).

There is even less evidence with regard to the line of research examining the influence of teaching practice on outcomes, which is the focus of this paper, since these activities are tricky to measure and quantify. However, the studies that do manage to make this type of measurements conclude that they have a very significant effect on student learning (Schacter and Thum, 2004) in both primary

(Santín and Sicilia, 2014) and secondary education (Carbonaro and Gamoran, 2002; Wentzel, 2002).

Furthermore, the empirical evidence on secondary education regarding this type of activities available in the literature highlights the positive impact of traditional practices based on the repetition of contents and knowledge acquisition exercises or problem solving (Brewer and Goldhaber, 1997; Schwerdt and Wuppermann, 2011; Bietenbeck, 2014). The same also applies to teaching styles based on problem solving and homework (De Witte and Van Klaveren, 2014). Likewise, students usually achieve better academic outcomes when the information is presented repeatedly, stressing the key concepts through repetition (Rosenshine and Stevens, 1986). However, there are other studies that do not find any relationship between the time spent on these activities and student achievement (e.g., Van Klaveren, 2011). A possible explanation for this result is that these strategies tend to be used more often with students whose performance is worse (Echazarra et al., 2016).

On the other hand, the more innovative teaching methods based on group work, classroom debates, student coaching or extrapolating learned concepts to real-world problems do not appear to be very successful (Lavy, 2011) and may even have negative effects on achievement in some cases (Murname and Phillips, 1981; Brewer and Goldhaber, 1997). An alternative interpretation of these results is, however, that these types of teaching practices possibly have more to do with students developing other skills, such as reasoning ability (Bietenbeck, 2014), or improving their social capital (Algan et al., 2013) than with the acquisition of the knowledge or competencies that are usually assessed by international knowledge tests.

Most of the above studies refer specifically to education in the United States, where students are accustomed to group work. However, very different teaching methods and teaching materials are used in other countries (Hiebert, 2003). Therefore, this research conducted in a setting like Spain that is worlds apart from America should make an important contribution to the existing literature on what role teachers play in developing students' cognitive skills.

2.3. Data and Variables

PISA (Programme for International Student Assessment) is an international study that assesses the knowledge and skills of 15-year-old students once every three years. PISA's biggest potential is that it provides comparable data for a very wide-ranging set of countries. In fact, 540,000 students from

schools in 72 economies participated in the last wave (43 countries participated in the first PISA wave, 28 of which were OECD members).

In 2015 all tests were computer based, and the key competence assessed was science. In this country, an extended sample from all the Spanish Autonomous Communities participated, that is, a total sample composed of 37,205 students from 980 schools whose results are comparable internationally. We picked a subsample composed of 6738 students from 201 schools from all the Autonomous Communities from the above sample. We allocated the appropriate sampling weights to the Autonomous Communities to assure that the data were representative of the whole country.

As mentioned in the introduction, one of the major PISA 2015 innovations was that, apart from questionnaires completed by students and school principals, teachers from the schools based in several countries that were assessed took a survey including wide-ranging questions about their training, experience and classroom activities[2]. In this respect, note that it is not possible to establish an exact connection between students and teachers on two grounds: the sampling structure used in PISA, where the 35 students from each school may be members of different classes, and the characteristics of secondary school teaching, where teachers rotate around different classes. In particular, we selected a random set of 25 secondary school teachers for each school in the representative subsample (10 science teachers and 15 teachers of other competencies) (OECD, 2016), such that a total of 4,286 Spanish teachers completed the teacher questionnaire.

The selected teachers had 30 minutes to complete a computer-based questionnaire module, including wide-ranging questions about their background, previous experience, professional development and teaching beliefs. Science teachers, in particular, answered a number of questions about science teaching/learning environments existing at the school, as well as the teaching strategies that they applied in their classrooms. These teaching practices are the focus of this paper and are detailed below.

Most of the questions concerning teaching strategies are generally stated as follows: How often does the following happen in your science class? This question is followed by a description of the activities: “Students are asked to draw conclusions”, “I demonstrate an idea”, “Students read materials from a textbook”, etc. Additionally, there is another block of questions directly concerning the application of a series of teaching techniques. The response format is very similar in both blocks: a four-category Likert scale: (1) Never or almost never; (2) Some lessons; (3) Many lessons; (4) Every lesson or almost every lesson.

[2] A total of 18 countries offer this information, nine of which are members of the OECD.

As there is a wide-ranging set of questions related to teaching strategies, one of the first actions that we took was to build indicators to summarize the information provided by teachers. Based on previous literature, we applied two different criteria to set up teaching styles for this purpose. The first is based on the proposal by Le Donné et al. (2016), which make a distinction between active learning, cognitive activation and teacher-led learning. The second criterion groups the teaching strategies into classical or modern (Zemelman et al. 2005; Bietenbeck, 2014; Hidalgo and López-Mayán, 2015). By using both criteria to build indicators, we can use the results to test their robustness. Table 2.1 shows the classification of questions for teachers according to each of the two alternatives[3].

To help with the interpretation of the different responses, we established the following weights for each of the responses: (1) 0%, (2) 33%, (3) 66%, and (4) 100%. Accordingly, the responses stand for the percentage of time spent on each of the teaching activities carried out by each teacher. Note that the teaching methods used are not mutually exclusive, that is, each teacher can apply more than one teaching method in the same class, while spending a different amount of time on each activity. After rescaling the variable values, we built the indicators using the classification proposed in Table 2.1.

After building the indicators representing the activity carried out by each teacher (how often he or she applies each of the teaching strategies), we clustered this information at school level. On the abovementioned grounds (no direct link between the student and teacher), it was necessary to work at school level, that is, we had to summarize the information provided by teachers in a single school-level indicator (for each teaching strategy) by calculating the average value of the indicators for each teacher employed by the school. In other words, our working hypothesis is that each teacher has a teaching style defined by his or her teaching methods. The school teaching style is the sum of the teaching styles of all teachers. Considering the structure of the teaching strategy indicators and the indicator of each strategy at school level, greater indicator values can be taken to mean that the school's teachers apply this type of strategies more often in their science lessons[4].

[3] The other variables for the teaching practices module completed by science teachers do not belong to only one of the created categories and were not therefore used to compute the indicator.
[4] Although teachers may each have a different teaching style, their scores will all be taken into account to calculate the school's mean score.

Table 2.1 Definition of teacher variables and classification according to Criterion 1 (cognitive, active and teacher-led) and Criterion 2 (classical and modern)

Criterion 1		
Cognitive	(a)	Students are asked to draw conclusions.
	(b)	Students are given opportunities to explain their ideas.
	(c)	The teacher discusses questions that students ask.
Active	(d)	A whole class discussion takes place.
	(e)	Current scientific issues are discussed.
	(f)	Students write up laboratory reports.
Teacher-led	(g)	The teacher explains scientific ideas.
	(h)	The teacher demonstrates an idea.
	(i)	Tailored tasks are assigned to the weakest as well as to the best students.
Criterion 2		
Classical	(g)	The teacher explains scientific ideas.
	(c)	The teacher discusses questions that students ask.
	(f)	Students write up laboratory reports.
	(h)	The teacher demonstrates an idea.
Modern	(a)	Students are asked to draw conclusions.
	(b)	Students are given opportunities to explain their ideas.
	(d)	A whole class discussion takes place.
	(e)	Current scientific issues are discussed.

In the process of building the teaching style indicators, the data were debugged according to several methods in order to be able to guarantee data reliability. On one hand, we only accounted for teachers who answered 100% of the questions covered by the teaching strategy indicators. As these variables were the main targets of our study, we opted not to use any missing value imputation procedure, relying exclusively on the information provided by the teachers that provide all the responses (85% of the sample). On the other hand, we found that there was a low percentage response from teachers at some schools. In such cases, we opted to remove any schools where the ratio of the percentage response to the teacher questionnaire (with respect to all schools selected for this purpose) over school size was less than 20%. This procedure was designed to assure that the strategies reported by only one teacher or very few teachers were not attributed to a school[5].

The information provided by science teachers is also used to create a set of control variables that are included in our estimations: average age, average experience, and the percentage of teachers with a higher qualification than required. The model should account for the different types of teachers

[5] The application of the criterion reduces the number of schools to 167 (5,411 students), but there is a guarantee that there is a high enough percentage response at the remaining schools.

existing at a school in order to discount their effect on student achievement in order to analyze the specific influence of the different teaching strategies.

Finally, the database that contains the information on teachers aggregated at school level has been merged with the databases for students and schools, using the common school identifier provided by PISA. This provides access to a data file containing detailed information on student characteristics, the school that they attend and the teachers at that school.

The dependent variable used in our model was the first plausible value of the science outcome (main competence assessed in PISA 2015)[6]. It also includes a set of explanatory variables at individual and school level that are consistent with the ones usually reported in the literature (see, for example, Calero and Escardíbul, 2007 or Cordero et al., 2013) as being the key determinants of educational achievement (sex, grade retention, mother's educational level, number of books in the home, school ownership and location or peer effect, i.e., average ESCS variable for the students at the school[7]). Table 2.2 summarizes all the selected variables, also including representative indicators of the teaching strategies and teacher characteristics with their respective descriptive statistics.

[6] It is acceptable, according to PISA specialists, to use a single plausible value or all five values available in the data base, as this does not lead to an appreciable difference in large samples (see OECD, 2009, p. 44).
[7] ESCS (index of economic, social and cultural status) is prepared by PISA specialists combining information on parent educational level and occupational status with household indicators.

Table 2.2 Descriptive statistics of the variables included in the model

	Mean	D.T	Min.	Max.
Student level				
First plausible value in science	498.48	85.89	210.69	754.33
Sex (Female)	0.50	0.50	0.00	1.00
Has repeated a year	0.27	0.44	0.00	1.00
First-generation immigrant	0.08	0.28	0.00	1.00
Mother's educational level (higher than post-compulsory secondary education)	0.62	0.49	0.00	1.00
Owns computer	0.92	0.26	0.00	1.00
More than 200 books in the home	0.25	0.43	0.00	1.00
Teacher level				
Average age of teachers	45.72	0.31	37.89	55.32
Average work experience	18.09	3.55	8.66	28.33
Higher than required qualification	0.26	0.44	0.00	1.00
Active teaching practices indicator	0.46	0.10	0.18	0.68
Cognitive teaching practices indicator	0.71	0.07	0.48	0.92
Teacher-led teaching practices indicator	0.72	0.08	0.46	0.94
Classical teaching practices indicator	0.67	0.07	0.46	0.82
Innovative teaching practices indicator	0.55	0.09	0.33	0.79
School level				
School ownership (private/grant-aided private)	0.32	0.47	0.00	1.00
Town centre	0.62	0.49	0.00	1.00
ESCS (school mean)	-0.47	0.68	-1.89	1.06

2.4. Results

In this section, we report and discuss the results of the empirical analysis conducted to assess the influence of teaching strategies on student outcomes. We first estimate a hierarchical or multilevel regression model including all the individual, school and teacher variables described above. This methodological approach takes into account that students from the same school have similar values for school variables. As a result, the average correlation between the variables (including variables related to teaching practices) will be greater for students at the same school than for students at different schools (Hox, 2002).

The aim of the model is to discover what relationship there is between school teacher teaching strategies (as well as other control variables) and student achievement for science, making a distinction between the above two criteria used to build the indicators. We then estimate a quantile regression model (Koenker and Bassett, 1978), again with a multilevel structure. The approach is designed to study whether the teaching strategies have a different effect for different science score cut-off points. This is a more precise approach, as the design accounts for four segments within the

results distribution (four quartiles) and estimates the effect of the explanatory variables on each segment. Accordingly, the slopes for the dependent variable may vary.

All the estimates were made using bootstrap techniques that cluster the data by schools, using 50 iterations to calculate the approximate standard error, as per OECD instructions (OECD, 2013). The errors are heteroscedasticity-consistent standard or robust errors. Students' original scores for science have been transformed into z-scores, meaning that the mean is 0 and the standard deviation is 1. This makes the results easier to interpret in terms of deviation. Table 2.3 shows the results of the estimations for teaching practices considering the two alternative criteria.

Table 2.3 Estimation of student achievement for science depending on teaching strategies

Variables	**Criterion 1**	**Criterion 2**
Sex	-0.193***	-0.193***
	(0.0212)	(0.0229)
Has repeated a year	-0.937***	-0.936***
	(0.0218)	(0.0235)
First-generation immigrant	-0.161***	-0.162***
	(0.0350)	(0.0464)
Mother's educational level	0.126***	0.124***
	(0.0230)	(0.0257)
Owns computer	0.157***	0.155***
	(0.0392)	(0.0409)
More than 200 books in the home	0.283***	0.282***
	(0.0214)	(0.0253)
ESCS (school mean)	0.207***	0.190***
	(0.0232)	(0.0219)
School ownership	-0.0771**	-0.0108
	(0.0340)	(0.0354)
Town centre	0.0947***	0.0990***
	(0.0255)	(0.0242)
Mean age of teaching staff	0.0336***	(0.158)
	(0.00806)	0.0400***
Mean job experience	-0.0196***	(0.00830)
	(0.00677)	-0.0216***
Higher teaching qualification than prescribed	0.0871***	(0.00695)
	(0.0254)	0.0768***
Cognitive teaching practices indicator	-0.390	
	(0.261)	
Teacher-led teaching practices indicator	0.838***	
	(0.197)	
Active teaching practices indicator	-0.206	
	(0.179)	
Classical teaching practices indicator		0.997***
		(0.241)
Innovative teaching practices indicator		-0.747***
		(0.158)
Constant	-1.210***	-1.522***
	(0.326)	(0.271)
Observations	5,411	5,411
Number of groups	167	167

Standard error in parentheses *** p<0.01, ** p<0.05, * p<0.1

Focusing on the key variables of interest, the results show that there is a statistically significant and positive relation between the indicator representing teacher-led teaching practices and academic achievement. This means that students get better scores for science at schools where the teacher uses this type of strategies more often (standard deviation of 0.838). On the other hand, the sign is negative for the other two strategies (cognitive activation and active learning), that is, marks tend to be lower, although the relation between the variables is not statistically significant. If instead of using three techniques, we separate by classical and innovative methods only, the results are even more conclusive, as the classical techniques have greater positive impact on science achievement (standard deviation of 0.997), whereas the relationship is again negative for the modern methods and is, in this case, also statistically significant.

Generally speaking, the positive and significant relationship detected for the more classical (traditional and teacher-led) activities is consistent with some of the evidence available in the previous literature (Schwerdt and Wuppermann, 2011; Bietenbeck, 2014), as is the negative impact of more innovative practices (Brewer and Goldhaber, 1997). However, some recent studies based on PISA data have discovered a positive relationship between cognitive activation strategies and student achievement in some participant countries (Le Donné et al. 2016), as well as for specific activities like small group work or the use of new technologies to carry out projects or complete classroom exercises (Méndez, 2015). Note, however, that these studies refer to mathematics teachers and focus on their relationship with mathematics achievement.

It is also noteworthy that the results for the parameter estimates of the control variables included in the model are generally what we expected. Thus, outcomes are worse for girls than for boys, student grade retention or immigrant status lowers science scores, and outcomes are better for students who own a computer, have more books at home and whose mother has a higher educational level. With respect to school variables, we find that outcomes are significantly and positively related to schools in urban areas and with a higher socioeconomic level and negatively correlated with private and grant-aided private schools[8]. Finally, with regard to school teachers, there is a significant and positive correlation of outcomes with age and qualifications, whereas experience has an unexpected, albeit small negative value.

The results of applying quantile regressions for the main segments of the science score distribution (25%, 50% and 75%) corroborate and, in some cases, fine-tune the results reported above. For simplicity's sake, Table 2.4 only reports the values of the parameter estimates for the key analysed

[8] The inclusion of the mean school ESCS variable in the model accounts for this result.

variables, as the sign and significance of the values of the other variables are very similar to the above. The results suggest that teacher-led teaching practices have a positive impact on only the top segments (quartile 50 and 75). On the other hand, the use of cognitive activation strategies is detrimental for students with better marks, as are active practices for students with poorer grades.

Table 2.4 Estimation of the relationship between science outcomes and teaching strategies by means of multilevel quantile regressions

	Three strategies			Two strategies		
VARIABLES	Q25	Q50	Q75	Q25	Q50	Q75
Cognitive practices indicator	0.142	-0.784**	-0.644*			
	(0.337)	(0.327)	(0.383)			
Teacher-led practices indicator	0.305	1.039***	0.909***			
	(0.264)	(0.233)	(0.306)			
Active practices indicator	-0.442*	0.0387	0.0755			
	(0.242)	(0.227)	(0.233)			
Innovative practices indicator				-0.791***	-0.725***	-0.706***
				(0.223)	(0.243)	(0.201)
Classical practices indicator				0.884***	1.002***	1.014***
				(0.263)	(0.267)	(0.255)
Student controls	X			X		
School controls	Y			Y		
Teacher controls	Z			Z		
Constant	-1.553**	-1.065**	-0.832	-1.963***	-1.681***	-1.020**
	(0.780)	(0.525)	(0.530)	(0.614)	(0.519)	(0.419)

Standard errors in parentheses *** p<0.01. ** p<0.05. * p<0.1

This evidence is not consistent with the results of some previous studies, which found that students that perform worse gain most from teacher-led strategies, whereas the better students benefit from the use of cognitive activities (Lavy, 2011, Le Donné et al. 2016)[9]. Although these are studies addressing different competencies and contexts outside Spain, this inconsistency between the results leads us to raise the need to conduct future research exploring other possible factors that may be affecting the relationship between teaching methods and academic outcomes.

Finally, we find that, if grouped into only two categories (innovative vs. classical), the estimates have significant, albeit disparate, values for all the segments. In fact, the coefficients of traditional strategies are positive (and upward), whereas they are negative (and downward) for innovative practices.

[9] These studies deal with students from a favourable or unfavourable socioeconomic background, although this condition is known to be clearly correlated with academic achievement.

2.5. Conclusions

This paper examined the relation between teaching strategies used by teachers from the same school and the results achieved by their students. The empirical analysis conducted for the specific context of secondary education in Spain is based on information available in the PISA 2015 database. This is the first PISA database to include information supplied by teachers about the teaching practices that they apply within the classroom. Based on the responses by teachers to a broad spectrum of questions, we built diverse indicators representing different teaching styles according to several criteria used in previous studies related to this issue.

The results of our estimations suggest that traditional methods, which are usually correlated with practices where the teacher plays a leading role as a conveyor of knowledge, are the ones that contribute most to improving the educational achievement of Spanish students in the science competence field, which was the focus of the PISA 2015 test. This result is constant across all the segments of the results distribution, as we infer from the estimates based on quantile regressions. On the other hand, the use of more innovative practices, which aim to promote student engagement through group work and the use of new technologies, as well as by setting challenges to stimulate critical thinking, does not appear to contribute to improving achievement and, in some cases, can even turn out to be detrimental. The competence assessment method used in the PISA test surely accounts for this result, as, despite the OECD's aim is to assess students' capacities to apply knowledge and skills in everyday life, it is very much linked to the knowledge acquired.

The reported results must be carefully analyzed, as they cannot be interpreted in terms of causality. This would require optimal experimental conditions where the teaching practices applied by teachers were independent of student achievement. Some previous studies have tried to emulate this ideal scenario by applying fixed effects at student level and discovering the effects of varying teaching practices across two different subjects for the same student (Cordero et al., 2017). However, this type of strategy can only be implemented if students can be linked to the teacher who actually taught the evaluated subjects. Unfortunately, the PISA database does not meet this criterion, as teacher information cannot be matched to student data. This led us to use aggregate, school-level data. This procedure is applicable considering that, as several previous studies have suggested (e.g. Méndez. 2015 or Le Donné et al. 2016), the teaching practices of the teachers from the same school are correlated with each other. If this is the case, the results derived from the fixed effects analysis based on the variation between teaching practices applied by teachers teaching different subjects can, in fact, be biased if there is such a correlation.

The results reported in this research aim to contribute to the debate on a very hot topic, namely, the role of teachers in the Spanish education system. On this issue, one of the points made by the recent White Paper on the Teaching Profession and the Educational Environment (Marina et al., 2015, p.11) was the need to "study and evaluate more efficient international educational innovations and advise teachers on the best teaching procedures and techniques". In this respect, this study provides empirical evidence on an issue that is as yet an open field in Spain because there are hardly any reliable data on the classroom activities performed by teachers. However, more and more databases are providing information in this regard (TIMSS, PISA –as of 2015 and previously by means of the TALIS-PISA link– or the different diagnostic tests developed by the Spanish Ministry of Education). On this ground, it is to be expected that other studies analyzing this question and offering further potentially useful evidence for decision making on this key issue of practices that can help to improve our students' learning process will be published in the near future.

Chapter 3

ESPECIALIZACIÓN EN LAS ESTRATEGIAS DOCENTES Y EFECTOS SOBRE LOS RESULTADOS ESCOLARES[10]

[10] This chapter has been written in Spanish because it has also been submitted to the Spanish journal "*Revista de Educación*", which requires that original submissions are written in this language.

3.1. Introducción

La relevancia del papel del profesorado es un hecho ampliamente aceptado por la comunidad educativa a nivel internacional (Hanushek, 2011). Sin embargo, tras más de medio siglo de investigación sobre esta cuestión, sigue sin estar claro qué deben hacer los profesores para mejorar el proceso de aprendizaje y, por ende, el resultado de sus estudiantes. El presente estudio se centra en analizar los efectos de los distintos estilos docentes que pueden adoptar los profesores sobre el rendimiento educativo de sus estudiantes, cuestión que ha generado un enorme interés en la literatura reciente tanto a nivel internacional (Hattie, 2009; Schwerdt y Wuperman, 2011; Bietenbeck, 2014) como en el contexto específico español (Hidalgo y López-Mayan, 2015, Méndez, 2015, Álvarez-Morán et al., 2018; Gil et al., 2018).

Cuando nos referimos a los estilos docentes, habitualmente se suele distinguir entre dos enfoques: las técnicas de enseñanza clásicas, en las que el profesor asume un rol protagonista exponiendo los contenidos o resolviendo ejercicios y problemas, y las estrategias más innovadoras en las que se adopta un enfoque más constructivista, tratando de promover el aprendizaje activo por parte de los estudiantes para conseguir desarrollar competencias de orden superior, como el pensamiento crítico o la toma de decisiones. Aunque normalmente los profesores tienen a combinar prácticas docentes que podrían catalogarse dentro de estas dos estrategias, es muchos casos resulta posible identificar que los profesores que pertenecen a u mismo centro educativo tienden a utilizar estrategias similares, desarrollando lo que se conoce en la literatura como una "cultura de enseñanza" (Le Donné et al., 2016), siendo por tanto posible que exista una especialización de los centros en el uso de uno u otro estilo docente. Basándonos en esta premisa, en esta investigación pretendemos examinar si los centros en los que los profesores declaran hacer un uso más intensivo de estas estrategias docentes, esto es, los centros que pueden ser catalogados como especializados en enseñanza clásica o innovadora, consiguen mejoras significativas en el rendimiento académico de sus alumnos. Este enfoque está en sintonía con otras iniciativas recientes, como por ejemplo *PISA for schools* (OCDE, 2017), en las que el foco se pone en lo que pueden hacer los centros educativos para mejorar sus resultados en lugar de analizar las iniciativas concretas llevadas a cabo por los profesores de manera individualizada.

Nuestro análisis empírico está basado en los datos proporcionados por la Evaluación General de Diagnóstico (EGD, en adelante) realizada en el año 2010 por el Instituto Nacional de

Evaluación Educativa (INEE), en la que participaron los alumnos de segundo curso de Educación Secundaria Obligatoria en España (INEE, 2011). Se trata de un programa nacional de evaluación del sistema educativo que evalúa el rendimiento de los estudiantes de segundo curso de ESO en cuatro competencias distintas: (a) Comunicación lingüística; (b) Matemáticas; (c) Conocimiento e interacción con el mundo físico y (d) Social y ciudadana. Hasta donde conocemos, esta fuente de información no se ha utilizado hasta el momento en estudios empíricos en nuestro país centrados en las prácticas docentes del profesorado. Por tanto, consideramos de gran interés comprobar si los resultados obtenidos al explotar esta información son similares a los de otros estudios previos basados en otras fuentes de información alternativas. Además, al disponer de información relativa a los estudiantes en cuatro competencias distintas, resulta posible examinar la posible existencia de divergencias respecto al impacto de las distintas estrategias docentes dependiendo de la competencia analizada.

Para poder evaluar correctamente el efecto de los diferentes estilos docentes en términos de causalidad, lo ideal sería disponer de información procedente de un experimento aleatorio controlado en el que los estudiantes se asignasen a los centros de manera aleatoria en función de la variable de análisis. En este contexto, para analizar los efectos de la aplicación de un estilo docente determinado, como puede ser el uso de estrategias de enseñanza de tipo constructivista o innovador (grupo tratado), habría que comparar los resultados de las escuelas que han aplicado este tipo de prácticas docentes de forma intensiva con los obtenidos por otras escuelas con estudiantes de similares características en las que este tipo de prácticas docentes no se apliquen de manera habitual, según lo declarado por sus profesores, que serviría como grupo de control (Rubin, 1974; Cook, 2002).

La utilización de este tipo de experimentos en el contexto educativo resulta cada vez más habitual en los países anglosajones (principalmente en Estados Unidos y Reino Unido). Sin embargo, en España prácticamente no se han utilizado hasta el momento, por lo que un estudio como el que aquí se plantea debe basarse en la información disponible en estudios de corte transversal. En este tipo de estudios la distribución de los alumnos entre los centros no es exógena (Rothstein, 2010), como tampoco lo son las estrategias implementadas por los profesores, que tenderán a adaptarse al nivel de rendimiento de sus alumnos (O´Dwyer et al., 2015). Por tanto, las características de los centros educativos que implementan unas

determinadas estrategias de enseñanza suelen ser distintas a las que no la ponen en práctica, tanto en lo referente a las variables para las que disponemos de información como de otros aspectos no observables, como puede ser la organización interna de los centros o la motivación intrínseca de los profesores. Esto puede generar un problema de endogeneidad en los datos, lo que complica en gran medida la estimación del efecto de las prácticas docentes (Gustafson, 2013), ya que la simple comparación entre los resultados obtenidos por los estudiantes que pertenecen a centros que utilizan un enfoque más innovador o más tradicional respecto al resto daría lugar a unos resultados sesgados. Por lo tanto, resulta fundamental asegurar que los grupos que se van a comparar (tratado y control) sean tan similares entre sí como sea posible.

Con este propósito, utilizamos como herramienta de análisis el propensity score matching (PSM), con el que se pretende replicar el diseño de un experimento aleatorio utilizando datos observados. De esta manera resulta posible evaluar el impacto de una medida aislando el efecto causal de la misma sobre la variable dependiente utilizando información relativa a escuelas que presentan características similares entre ellas antes de la intervención analizada (Stuart, 2007). Para ello, será necesario controlar nuestros resultados considerando un conjunto de variables representativas de la escuela y de su profesorado que están relacionados con la probabilidad de que un centro pertenezca o no al grupo tratado, como pueden ser el acceso a las nuevas tecnologías, la participación en proyectos de innovación docente o la asistencia a cursos de formación.

Al utilizar el PSM como método de evaluación, una de las principales dificultades que plantea el análisis propuesto es la fijación del criterio de división de la muestra de centros para poder configurar los grupos de tratamiento y control. Esta decisión resulta relativamente sencilla cuando la distinción entre los centros viene determinada por el fenómeno analizado, como por ejemplo la comparativa entre centros públicos y privados (Vandenbergue y Robin, 2004; Dronkers y Avram, 2010) o los que están sujetos (o no) a competencia en el entorno cercano (Agasisti y Murtinu, 2012). Pero en nuestro caso esta distinción no resulta tan evidente, puesto que la catalogación de una escuela como especializada en el uso de un determinado estilo docente no obedece a un criterio previamente establecido. Así, nuestra propuesta se basa en segmentar la muestra en función de los valores de dos indicadores continuos que reflejan la frecuencia con la que desempeñan determinadas prácticas docentes los profesores

del centro (innovadoras y clásicas). Concretamente, distinguiremos entre las escuelas que se sitúan en los extremos de la distribución de dichos indicadores, tal y como se explica en la sección tercera, de modo que el análisis propuesto se basa en comparar el desempeño de los alumnos pertenecientes a las escuelas especializadas en el uso de estrategias innovadoras (25% superior de la distribución) con respecto al resto y, de manera análoga, las escuelas con un uso más intensivo de las prácticas docentes clásicas (nuevamente el 25% superior de la distribución) en comparación con el resto.

El resto del artículo se organiza de la siguiente manera. La sección segunda contiene una exhaustiva revisión de los trabajos previos que han analizado el efecto de las prácticas docentes, prestando especial atención a la evidencia disponible para el caso español. En la sección tercera se exponen las principales características de la base de datos utilizada, así como el procedimiento utilizado para construir los indicadores representativos de las estrategias docentes. En la sección cuarta se explica el enfoque metodológico utilizado, así como los supuestos e hipótesis en los que se basa. Posteriormente, en la sección quinta se presentan y discuten los principales resultados obtenidos. El trabajo concluye con el habitual apartado de conclusiones.

3.2. Revisión de la literatura

Existe un amplio número de trabajos que ha destacado el papel fundamental que juegan los profesores en el proceso de aprendizaje de los alumnos (Hattie, 2009; Campbell et al., 2012; Chetty et al., 2014). En sus orígenes, esta línea de investigación se centraba en analizar características observables de los profesores, como su nivel de experiencia o su cualificación (Ehrenberg y Brewer, 1994, Rowan et al., 2002). No obstante, estos factores tienen una influencia indirecta sobre el aprendizaje, ya que en realidad estos factores ofrecen una aproximación a la manera en la que los profesores imparten sus clases (Pallardy y Rumberger, 2008). De hecho, cuando se dispone de información directa sobre las prácticas docentes los efectos detectados sobre el aprendizaje suelen ser mucho más relevantes que las características de los profesores (Xue y Meisels, 2004; Schacter y Thum, 2004).

La posibilidad de disponer de información acerca de lo que los profesores hacen en el aula tradicionalmente ha resultado muy compleja al considerarse el aula como una caja negra cuya exploración resulta extremadamente compleja. Sin embargo, la proliferación de las bases de

datos educativas internacionales, en las que cada vez resulta más común que los profesores rellenen un cuestionario sobre multitud de aspectos relacionados con su actividad docente, ha facilitado en gran medida la obtención de información acerca de este tipo de actividades, lo que a su vez ha dado lugar al desarrollo de una amplia literatura dedicada principalmente a comparar el efecto de las prácticas docentes clásicas con las más innovadoras.

La conclusión más habitual de los trabajos que han abordado esta cuestión es que las estrategias docentes clásicas tienen un efecto positivo sobre el rendimiento educativo de los estudiantes (Brewer y Goldhaber, 1997; Schwerdt y Wuppermann, 2011; Lavy, 2015; Bietenbeck, 2014; Caro et al., 2016). Así, diversos estudios han demostrado la influencia positiva de repetir determinados ejercicios en clase e incluso memorizar la manera de resolverlos (House, 2009; Camburn & Han, 2011; De Witte & Van Klaveren, 2014) o el hecho de exponer los conceptos fundamentales de manera redundante (Rosenshine y Stevens, 1986). Sin embargo, no todos los estudios coinciden. Así, por ejemplo, Zuzovsky (2013) identifica una influencia positiva de las lecciones magistrales, mientras que Van Klaveren (2011) no identifica ningún efecto significativo para este tipo de prácticas.

Por el contrario, la evidencia empírica disponible acerca del efecto de las prácticas docentes más innovadoras resulta contradictoria. Así, podemos encontrar trabajos en los que su influencia sobre los resultados es mínima (Lavy, 2011) o incluso negativa (Murname y Phillips, 1981; Brewer y Goldhaber, 1997), pero también otros en los que la utilización de este tipo de estrategias está asociada con un mejor rendimiento (Papanastasiou, 2008; Baumert et al., 2010; Le Donné et al., 2016).

Otros trabajos identifican que este tipo de prácticas pueden ser efectivas en unos países, pero no en otros. Así, en un estudio que analiza información relativa a 49 países participantes en TIMSS 2007, Zuzovsky (2013) identifica una influencia positiva de este tipo de prácticas en el caso de los estudiantes pertenecientes a países con mejores resultados, pero negativa para los alumnos de países con resultados más mediocres. De manera similar, resulta posible encontrar resultados contradictorios dependiendo de la tipología del alumnado. Por ejemplo, en otro estudio referido a un amplio conjunto de países (62) participantes en PISA 2012, Caro et al. (2016) llegan a la conclusión de que la influencia positiva de las estrategias innovadoras sobre los resultados académicos se concentra en los alumnos de mayor nivel socioeconómico.

Un planteamiento alternativo que permite explicar en cierta medida estos resultados tan contradictorios en relación con las estrategias innovadoras es que su implementación está encaminada a desarrollar de una serie de habilidades que difícilmente pueden medirse mediante los test estándar de conocimientos. Así, por ejemplo, Bietenbeck (2014) identifica un efecto positivo de este estilo de enseñanza sobre la capacidad de razonamiento, mientras que Algan et al. (2013) señalan que están vinculadas con un comportamiento más colaborativo en clase e incluso con una mayor participación ciudadana.

En el contexto específico del sistema educativo español, al igual que ocurre en el contexto internacional, la evidencia disponible sobre la influencia de los diferentes estilos de enseñanza en el rendimiento educativo ofrece resultados muy dispares. En este sentido, cabe señalar que los estudios que han analizado esta cuestión pueden clasificarse en dos grandes grupos. Por un lado, están los que, al disponer únicamente de información agregada a nivel de centro educativo, estudian cómo influyen las estrategias de enseñanza utilizadas por el conjunto de los profesores de la escuela sobre el rendimiento de los estudiantes, utilizando normalmente un enfoque econométrico tradicional basado en correlaciones entre variables. Por otro lado, encontramos trabajos que examinan los efectos de las prácticas docentes sobre los resultados de sus alumnos a partir de información individualizada sobre cada profesor.

Dentro del primer grupo se encuadran los trabajos de Méndez (2015) y Gil y Cordero (2018a, 2018b), en los que se utiliza la base de datos resultante de casar la información sobre el alumnado participante en PISA 2012 con el profesorado de los mismos centros que fue encuestado en TALIS 2013 mediante el denominado TALIS-PISA link. Como principal conclusión, Méndez (2015) identifica una relación positiva entre un uso más intensivo de determinadas estrategias innovadoras como el trabajo en grupo reducido o el uso de las nuevas tecnologías y el rendimiento académico de los alumnos. Sin embargo, en los trabajos de Gil y Cordero se identifica el efecto contrario, es decir, una ligera influencia negativa para las prácticas innovadoras.

Los trabajos de Gil et al. (2018) y Álvarez-Morán (2018) analizan la información contenida en PISA 2015, en la que por primera vez los profesores informaban acerca de las actividades que realizan en el aula. En el primero de estos trabajos los autores señalan que en los centros en los que los profesores utilizan mayoritariamente estrategias docentes clásicas o tradicionales los alumnos obtienen mejores resultados, mientras que el uso mayoritario de

estrategias innovadoras no resulta tener una incidencia significativa e, incluso en algunos casos, puede hacer que los resultados sean peores. Los autores del segundo estudio cuestionan la evidencia anterior relativa a la influencia de las prácticas clásicas, al considerar que no existe suficiente evidencia para justificar que las estrategias de enseñanza centradas en el uso de este tipo de actividades conduzcan a un rendimiento educativo superior.

En el segundo grupo se sitúa el trabajo de Hidalgo y López-Mayán (2015), en el que se explota la información ofrecida por la Evaluación General de Diagnóstico realizada por el Ministerio de Educación en el año 2009 a los alumnos de cuarto curso de Educación Primaria. A partir de la información facilitada por los profesores sobre sus actividades junto con la de los alumnos a los que éstos imparten clase, se estiman diferentes modelos de efectos fijos que permiten identificar una influencia positiva sobre el rendimiento académico de las prácticas docentes de estilo más moderno.

En el presente estudio se utiliza la información disponible en la EGD 2010, en la que únicamente se dispone de información agregada sobre los profesores de cada centro educativo, por lo que nos incluiríamos dentro del primer grupo. No obstante, en nuestro caso, hemos tratado de identificar un efecto causal mediante la aplicación de técnicas de propensity score matching, lo que, sumado al hecho de que hasta el momento esta base de datos no ha sido utilizada ningún estudio previo, otorga al trabajo un marcado carácter innovador.

3.3. Datos y variables

Los datos utilizados en este trabajo proceden de la EGD 2010. La base de datos tiene un diseño muestral basado en un sistema de muestreo aleatorio simple por conglomerados en dos etapas, en el que los estratos son las CC.AA. y los conglomerados son las escuelas. En la evaluación participaron los alumnos, sus profesores y los directores de los centros escolares, que proporcionan información sobre diversos factores que pueden tener incidencia en los resultados educativos. La muestra está compuesta por un total 29.154 alumnos, 4.488 profesores y 843 directores pertenecientes a 870 centros.

La EGD 2010 proporciona una valiosa información captada a través de tres cuestionarios. El primero de ellos está dirigido a los directores de centros escolares, que son los encargados de proporcionar información del contexto escolar y los recursos educativos disponibles. El segundo es el cuestionario dirigido a los profesores de secundaria del centro, entre los cuales

debe estar el orientador, el jefe de estudios y el tutor de cada uno de los grupos evaluados, en el que incluyen diversas cuestiones relativas tanto al contexto del aula como a sus opiniones sobre la actividad docente. En tercer lugar, el cuestionario dirigido al alumnado incluye multitud de variables relativas a su contexto personal y familiar, junto con los resultados obtenidos en las pruebas de conocimientos de cada una de las cuatro competencias evaluadas, representadas a través de cinco valores plausibles obtenidos mediante la aplicación de los modelos de respuesta al ítem (Rasch, 1960/1980)[11], habituales en todas las pruebas de conocimientos internacionales (por ejemplo, PISA, TIMSS o PIRLS), que reflejan el conjunto de habilidades, destrezas y conocimientos demostrados por los estudiantes.

Dado el propósito de nuestro estudio, nuestro principal interés se centra en el cuestionario de los profesores. Además de incluir preguntas relativas a sus antecedentes, su formación y sus creencias sobre diversos aspectos de la docencia, hay varias cuestiones específicas acerca de las estrategias docentes que llevan a cabo en sus clases. La mayoría de ellas tiene la siguiente formulación: "En general, ¿cuál es la frecuencia con la que ha usado durante este curso con los alumnos de segundo curso de Educación Secundaria Obligatoria los siguientes procedimientos didácticos?". Tras la pregunta, se detallan una serie de ítems relacionados con la actividad en la clase, por ejemplo: "Explico durante la mayor parte de la clase", "Los alumnos exponen temas o trabajos", "Promuevo los debates en clase", o "Los alumnos trabajan en pequeños grupos", entre otras. Las respuestas se basan en una escala Likert, con los siguientes valores: Nunca o casi nunca (1); Algunas veces (2); Casi siempre (3) y Siempre (4).

Una vez detectadas las variables relativas a las prácticas docentes, hemos utilizado un procedimiento de síntesis para poder resumir esta información en dos índices representativos de los distintos estilos de enseñanza utilizados por cada profesor que completa el cuestionario. Basándonos en los criterios utilizados en otros trabajos previos (por ejemplo, Hidalgo y López-Mayán, 2015 o Gil et al., 2018), se ha seguido el esquema propuesto por Zemelman et al. (2005), que agrupa las estrategias docentes en clásicas o innovadoras. Este planteamiento nos permite considerar el carácter no excluyente de ambas estrategias (Caro et al., 2016). Se asume, por tanto, que un docente puede combinar el uso de ambos tipos de prácticas tanto en la misma actividad como a lo largo de un curso escolar, en función de sus necesidades o

[11] Para un análisis detallado de esta cuestión véase Wu (2005) o Martínez-Arias (2006).

programación docente. Con el objetivo de facilitar la interpretación de los resultados, se establecen las siguientes ponderaciones para cada una de las respuestas: (1) 0%; (2) 33%; (3) 66% y (4) 100%. Los porcentajes obtenidos podrían asimilarse a la proporción de tiempo empleado en cada una de las actividades de enseñanza llevadas a cabo por el docente. Una vez re-escalados los valores de las variables, se construyen los indicadores utilizando la clasificación de los ítems propuesta en la Tabla 3.1.

Tabla 3.1. Distribución de variables entre las estrategias docentes clásicas e innovadoras

Estrategias docentes		
Clásica	(a)	Explico durante la mayor parte de la clase
	(f)	Los alumnos hacen los ejercicios o actividades que propongo
	(g)	Los alumnos trabajan individualmente
	(i)	Los alumnos toman apuntes
Innovadora	(b)	Los alumnos exponen temas o trabajos
	(e)	Promuevo los debates en clase
	(h)	Los alumnos trabajan en pequeños grupos

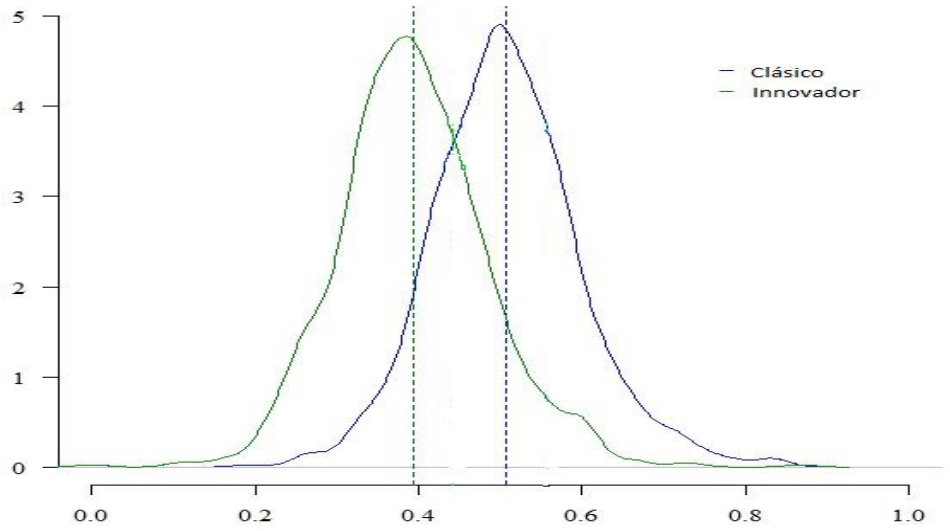
Posteriormente, se procedió a agrupar los diferentes indicadores obtenidos para cada profesor a nivel de su centro, calculando el valor medio de cada estrategia docente por escuela. La necesidad de trabajar a nivel de centro viene determinada por la falta de enlace directo entre profesor y alumno en la EGD[12.] Por tanto, se asume que el promedio de los índices sintéticos de los profesores que pertenecen a un mismo centro representa una aproximación de la cultura docente del centro. Tanto a nivel individual como agregado los indicadores representativos de las estrategias docentes clásicas e innovadoras se interpretan de la misma forma, esto es, valores más elevados suponen que una utilización más intensiva de un determinado estilo docente en sus clases.

El siguiente paso sería dividir la muestra de centros según su grado de especialización en el uso de una u otra estrategia. En este sentido, tras realizar un análisis exploratorio de la distribución de estas dos variables (Figura 3.1), se observó una elevada concentración de los valores alrededor de la media en ambos casos, lo que nos llevó a pensar que una simple

[12] Este supuesto resulta lógico al trabajar en el nivel de educación secundaria, en el que los alumnos tienen profesores diferentes para cada materia, a diferencia de lo que ocurre en primaria.

división de la muestra entre los que se situaban a la izquierda y la derecha de este valor nos podría conducir a una identificación incorrecta de los centros que realmente apuestan de un modo decidido por el uso intensivo de alguna de estas estrategias.

Figura 3.1. Distribución de los indicadores representativos de las estrategias docentes



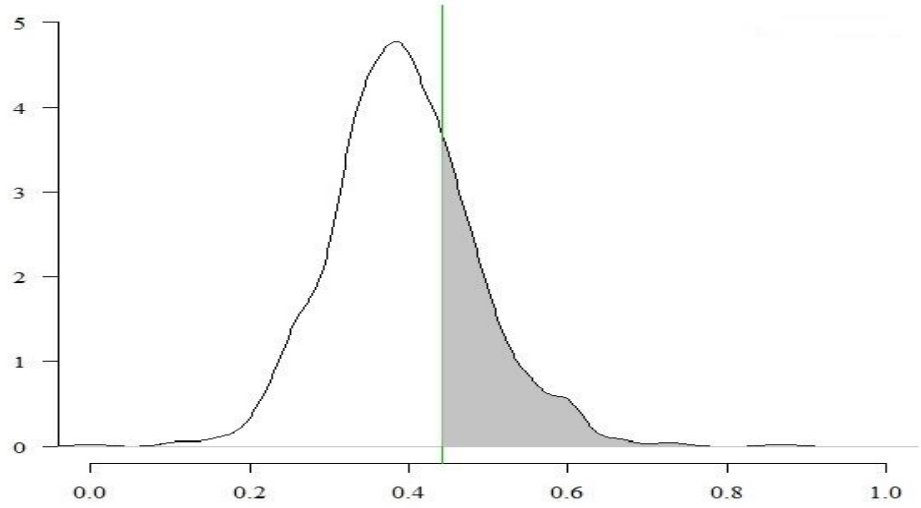
Por lo tanto, decidimos establecer como criterio para considerar a una escuela como especializada en la enseñanza innovadora o clásica debía pertenecer al 25% superior de la distribución de esta variable, es decir, el percentil 75[13]. Así, podemos clasificar del siguiente modo. Por un lado, seleccionamos una submuestra formada por los colegios en los que predomina la aplicación de estrategias docentes innovadoras (colegios top-innovadores), es decir, los que se sitúan en el área sombreada (gris) de la Figura 3.2 y los comparamos con el resto de los centros (en blanco). Por otro lado, seleccionamos una segunda submuestra formada por los centros que utilizan más intensivamente técnicas de enseñanza clásicas (colegios top-clásicos), esto es, los situados en el área sombreada (gris) de la Figura 3.3 y los comparamos con el resto de centros (en blanco).

Una vez que hemos identificado a los centros, nuestra estrategia de evaluación exige disponer de información adicional sobre las características de los centros educativos que puede tener incidencia sobre la probabilidad de que sea incluido dentro de la categoría de top-innovador o top-clásico y sobre los resultados de los estudiantes. Para ello, hemos recopilado información

[13] La fijación de un porcentaje del 25% en lugar de otro más bajo (20, 10 o 5 por ciento) obedece a la necesidad de tener un número suficiente de centros representativos de un estilo docente para poder implementar posteriormente el *propensity score matching*.

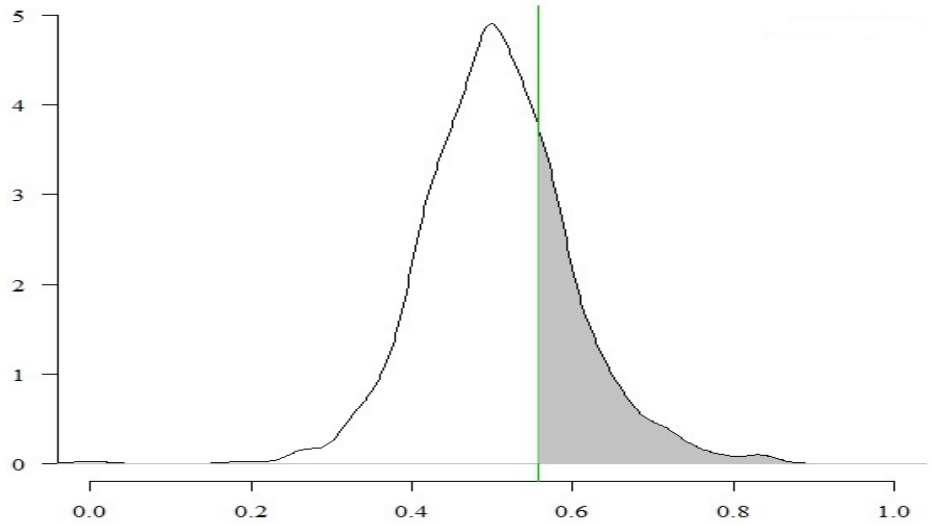
sobre multitud de aspectos relacionados con la actividad laboral de los profesores tomando como referencia la información proporcionada por éstos en su cuestionario específico y, posteriormente, agregando dicha información a nivel de centro mediante el cálculo de los valores medios de los docentes que pertenecen a la misma escuela. En concreto, las variables construidas han sido las siguientes: antigüedad en el centro, cualificación superior a la exigida, grado de satisfacción con su labor docente y con el resultado académico de los alumnos[14], reuniones colectivas en las que han participado a lo largo del curso, cursos de menos de 50 horas a los que han asistido, jornadas de formación en las que han participado, horas dedicadas a proyectos de innovación docentes, horas dedicadas al análisis de la práctica docente y frecuencia de uso de varios recursos docentes (materiales elaborados por los propios docentes; prensa o revistas especializadas y ordenadores e internet). Finalmente, tomando como referencia la información proporcionada por los directores de los centros en su cuestionario específico, se han configurado dos variables representativas de la titularidad del centro (público o privado) y el número de profesores que participan en el plan de formación del centro.

Figura 3.2. Clasificación de centros entre top-innovador (gris) y el resto (blanco)



[14] Estas dos variables se definen tomando como referencia lo que indican los profesores en dos cuestiones en las que las respuestas adoptan una escala de frecuencia tipo Likert con cuatro opciones.

Figura 3.3. Clasificación de centros entre top-clásico (gris) y el resto (blanco)



La fusión de la base de datos que contiene la información sobre los profesores agregada a nivel de centro con la base de datos con las puntuaciones de los estudiantes se realiza utilizando el identificador común de la escuela que proporciona la EGD 2010. Como variable representativa de las puntuaciones obtenidas por los estudiantes en las cuatro competencias evaluadas se utilizará únicamente el primer valor plausible[15]. Las Tablas 3.2 y Tabla 3.3 resumen las variables seleccionadas con sus correspondientes estadísticos descriptivos, tanto para el total de la muestra como para los centros agrupados atendiendo a los diferentes grupos utilizados en las estimaciones.

[15] Aunque la EGD 2010 presenta cinco valores plausibles por competencia evaluada, la utilización de un único valor plausible no representa una diferencia apreciable cuando se trata de muestras suficientemente grandes, como es nuestro caso, tal y como se ha puesto de manifiesto en diversos informes técnicos de bases educativas internacionales (por ejemplo, véase OCDE, 2009, p. 44).

Tabla 3.2. Estadísticos descriptivos de las variables incluidas en el modelo

	Media	**D.T**	**Min.**	**Max.**
Variables dependientes				
Comunicación lingüística	504,4	98,3	145,9	848,6
Matemáticas	503,6	100,0	138,3	854,8
Conocimiento e interacción con el mundo físico	504,2	98,7	147,2	848,4
Social y ciudadana	503,2	99,2	192,2	840,0
Variables explicativas				
Nivel profesor				
Indicador prácticas docentes clásicas	***0,51***	***0,09***	***0,00***	***0,84***
Indicador prácticas docentes innovadoras	***0,39***	***0,09***	***0,00***	***0,87***
Cualificación superior a la exigida	0,03	0,07	0,00	0,50
Antigüedad media en la docencia	3,89	1,14	1,00	7,00
Satisfacción media del profesorado con su función	2,70	0,62	1,00	4,00
Satisfacción con el resultado académico de los alumnos	2,03	0,47	1,00	4,00
Nº medio de reuniones colectivas	2,67	0,83	1,00	5,00
Nº medio de horas que se atiende a cursos < 50 horas	45,47	32,53	0,00	285,00
Nº medio de horas que se atiende a proyectos de innovación	6,40	12,75	0,00	108,00
Nº medio de veces que se dedican al análisis de la práctica docente	1,07	4,18	0,00	62,00
Frecuencia media del uso de materiales elaborados	0,57	0,14	0,17	1,00
Frecuencia media del uso de prensa o revistas especializadas	0,54	0,16	0,00	1,00
Frecuencia media del uso de ordenadores e internet	0,49	0,13	0,00	1,00
Frecuencia del uso de medios audiovisuales	0,56	0,14	0,16	1,00
Necesidad de aspectos formativos relacionados con la enseñanza por competencias	0,54	0,16	0,00	1,00
Nivel escuela				
Titularidad del centro (Privada)	0,33	0,47	0,00	1,00
Nº de profesores de secundaria que participan en el plan de formación	41,79	36,24	1,00	88,00
Número de centros = 870				

Fuente: Elaboración propia a partir de la información disponible en la EGD 2010.

Tabla 3.3. Estadísticos descriptivos de las variables incluidas en el modelo por tipo de centro

	Top Innova		Grupo de control Top innovador		Top Clásico		Grupo de control Top Clásico	
	N=215		N=655		N=210		N=660	
	Media	**D.T**	**Media**	**D.T**	**Media**	**D.T**	**Media**	**D.T**
Variables dependientes								
Comunicación lingüística	499,2	97,5	506,1	98,48	505,7	98,6	504,0	98,1
Matemáticas	502,5	99,1	504,0	100,3	503,5	100,5	503,6	99,8
Conocimiento e interacción con el mundo físico	500,8	98,6	505,3	98,7	506,5	98,7	503,4	98,6
Social y ciudadana	500,7	99,4	504,0	99,1	506,8	99,8	502,0	99,1
Variables explicativas								
Nivel profesor								
Indicador prácticas docentes clásicas	***0,50***	***0,09***	***0,50***	***0,09***	***0,62***	***0,05***	***0,46***	***0,06***
Indicador prácticas docentes innovadoras	***0,51***	***0,05***	***0,35***	***0,06***	***0,39***	***0,10***	***0,39***	***0,08***
Cualificación superior a la exigida	0,02	0,07	0,02	0,07	0,02	0,07	0,02	0,07
Antigüedad media en la docencia	3,82	1,22	3,91	1,10	3,85	1,09	3,90	1,14
Satisfacción media del profesorado con su función	2,82	0,62	2,66	0,61	2,80	0,65	2,66	0,60
Satisfacción con el resultado académico de los alumnos	2,06	0,45	2,02	0,47	2,13	0,48	2,00	0,45
Nº medio de reuniones colectivas	2,76	0,86	2,64	0,81	2,60	0,82	2,69	0,83
Nº medio de horas que se atiende a cursos < 50 horas	51,53	38,73	43,46	29,96	46,47	38,98	45,14	30,20
Nº medio de horas que se atiende a proyectos de innovación	8,88	16,79	5,58	10,99	6,56	14,27	6,35	12,24
Nº medio de veces que se dedican al análisis de la práctica docente	1,72	7,12	0,84	2,50	0,93	2,83	1,10	4,52
Frecuencia media del uso de materiales elaborados	0,51	0,14	0,48	0,12	0,50	0,14	0,48	0,12
Frecuencia media del uso de prensa o revistas especializadas	0,27	0,11	0,20	0,10	0,22	0,11	0,22	0,11
Frecuencia media del uso de ordenadores e internet	0,37	0,13	0,32	0,13	0,32	0,15	0,34	0,13
Frecuencia del uso de medios audiovisuales	0,57	0,15	0,56	0,14	0,57	0,15	0,56	0,14
Necesidad de aspectos formativos relacionados con la enseñanza por competencias	0,55	0,15	0,53	0,16	0,53	0,17	0,54	0,15
Nivel escuela								
Titularidad del centro (Privada)	0,30	0,46	0,33	0,47	0,31	0,46	0,32	0,47
Nº de profesores de secundaria que participan en el plan de formación	43,61	36,57	41,18	36,13	38,63	36,21	42,76	36,22

Fuente: Elaboración propia a partir de la información disponible en la EGD 2010.

3.4. Metodología

El método de *propensity score matching* que proponemos utilizar para analizar el efecto de los estilos docentes es un método cuasi-experimental desarrollado originalmente por Rosenbaum y Rubin (1983), con el que se pretenden solventar algunas de las principales limitaciones a las que se enfrentan los investigadores cuando únicamente disponen de información de tipo observacional, en las que la composición de los grupos que se pretenden comparar no es exógena, como ocurre en los experimentos aleatorios. De hecho, es una técnica que se ha utilizado en diversos estudios empíricos para tratar de identificar efectos causales a partir de datos procedentes de bases de datos internacionales como TIMSS o PISA (Cordero et al, 2017).

Esta metodología es una extensión de los métodos no paramétricos de *matching* que tiene por objeto realizar emparejamientos (en nuestro caso, entre centros especializados en un determinado estilo docente –tratados- y los que no lo están – control–), para evitar los posibles sesgos derivados de sus diferentes características, tanto observables como no observables. Se trata de buscar la diferencia entre los resultados obtenidos por los estudiantes que pertenecen a una tipología de escuela (factual) y los que esos mismos estudiantes habrían conseguido, en ausencia del tratamiento asumiendo que todo lo demás sería igual (situación contrafactual). El promedio de estas diferencias para la totalidad de los individuos de la muestra permite aproximar el efecto medio del tratamiento (ATE, *average treatment effect*).

Una vez que hemos dividido la muestra de acuerdo a los criterios expuestos en la sección anterior, el primer paso para la implementación del PSM consiste en el cálculo de la ecuación de selección, esto es, el modelo que nos va a permitir determinar cuál es la probabilidad de que un centro pertenezca a una tipología u otra (uso intensivo de una estrategia docente vs uso poco intensivo). De esta manera se pretende reducir el problema de emparejamiento a una única dimensión, lo que simplifica en gran medida el procedimiento (Wilde y Hollister, 2007). La idea subyacente es que si dos escuelas tienen el mismo *propensity score*, pero pertenecen a diferentes grupos (tratado o control), puede considerarse que la asignación es aleatoria, aunque en realidad no lo sea.

De las diferentes especificaciones que se pueden realizar (Guo y Fraser, 2010), utilizaremos un modelo *probit* para calcular la probabilidad de pertenecer al grupo tratamiento. En dicha ecuación, tratando de reducir el posible sesgo de selección entre los centros evaluados, incluiremos una amplia batería de variables de control relativas a la escuela que puedan estar relacionadas con la asignación del tratamiento (uso de diferentes estrategias de enseñanza) y simultáneamente con la

variable de interés, que en nuestro caso serán los resultados de los estudiantes (Heckman y Navarro, 2004; Stuart y Rubin, 2008). Estas variables se han seleccionado teniendo en cuenta la literatura previa y no las posibles correlaciones entre los grupos que componen la base de datos (Caliendo y Kopeinig, 2008). A continuación, se procede a balancear las muestras de ambos grupos utilizando el indicador de *propensity* obtenido en el paso anterior. Este indicador representa la probabilidad condicional de pertenecer al centro en el caso de cada individuo de la muestra, dadas sus características observables X, es decir:

$$ps = P(W - 1|X) \tag{3.1}$$

Por último, se procede a realizar el procedimiento de *matching*, ligando a una unidad de grupo de control con otra del grupo tratamiento (ATT es el ATE para los tratados), en el caso de que tengan valores similares del indicador de propensity.

$$\widehat{ATT_{match}} = E\left(\widehat{Y_{match.1}} \middle| W_{match} = 1\right) - E\left(\widehat{Y_{match,0}} \middle| W_{match} = 0\right) \tag{3.2}$$

donde (W) es el impacto de la intervención evaluada sobre el resultado obtenido (Y), siendo W=1 un centro especializado (top innovador o top clásico) y 0 el caso contrario.

Tras realizar pruebas con los diferentes algoritmos existentes para realizar el matching (por ejemplo, el vecino más cercano -*nearest neighbor*- o diferentes tipos de *kernel*[16]), finalmente nos hemos decantado por el uso del *kernel Epanechnikov* con un ancho de banda (*bandwidth*) de 0.06, al ser el que proporciona una mayor reducción en los sesgos. Tras el cálculo del PSM, se obtiene una muestra emparejada en la que se produce una reducción de los sesgos que resultan ser inferiores al 5%. Esto indica que la técnica utilizada permite solucionar en gran medida el problema de la falta de comparabilidad entre el grupo de tratamiento y el grupo de control. Por tanto, una vez aplicado el PSM, los resultados de grupo tratamiento (los top-innovadores o top-clásicos) son comparables con los del grupo de control (no pertenecientes a estos grupos).

[16] Los métodos kernel pueden definirse como estimadores de *matching* no paramétricos que comparan el resultado de cada unidad tratada con una media ponderada de los resultados de todas las unidades del grupo de comparación, utilizando las mayores ponderaciones para las unidades con un valor de *propensity* más parecido al que se compara (Rodríguez-Coma, 2012).

3.5 Resultados

En este apartado presentamos los resultados obtenidos en las estimaciones realizadas mediante el PSM a partir de los datos disponibles en la EGD 2010. En primer lugar, se presentan los resultados de la estimación de un modelo *probit* en el que se utiliza como tratamiento el criterio de especialización en un estilo docente innovador (top-innovador), incluyendo las variables a nivel de centro escolar descritas en la sección tercera (Tabla 3.4). A continuación, en la Tabla 3.5 se muestran los resultados obtenidos al estimar los modelos PSM propiamente dichos, con los que resulta identificar el efecto del uso intensivo de este tipo de prácticas docentes sobre los resultados de los alumnos para las cuatro competencias evaluadas en la EGD. A la hora de interpretar estos resultados debe tenerse en cuenta que los parámetros de este tipo de modelos reflejan la influencia de cada predictor sobre el *propensity score*, es decir, la contribución de cada covariable a la explicación de la probabilidad condicionada de asistir a un colegio intensivo en prácticas docentes innovadoras. De manera análoga, en las Tablas 3.6 y 3.7 se presentan los resultados obtenidos al considerar el caso de los colegios especializados en el uso de estrategias de enseñanza clásica (top-clásico). En las siguientes líneas se resumen las principales conclusiones que se derivan de ambas estimaciones.

Criterio "top-innovador"

En la Tabla 3.4 se observa que la mayoría de las variables de control incluidas en la regresión tienen una incidencia significativa como determinantes de la probabilidad de que un centro se incluya dentro de esta categoría. Entre ellas son mayoría las que influyen de manera positiva: las dos variables representativas del grado de satisfacción de los profesores (con sus funciones y con los resultados de los alumnos), el número de profesores que participan en el plan de formación del centro, la participación en reuniones colectivas, la participación en cursos de formación, las horas dedicadas a proyectos de innovación, la participación en grupos de análisis de prácticas docentes, el uso frecuente de prensa y revistas especializadas, el número de profesores de secundaria que participan en el plan de formación y la necesidad de aspectos formativos relacionados con la enseñanza por competencias. Las dos únicas variables que parecen incidir de manera negativa y significativa son la titularidad privada y la antigüedad media de los profesores. Por último, las variables de uso de materiales elaborados por el docente, el uso frecuente en el aula tanto de ordenadores como de medios audiovisuales y la posesión de una cualificación superior a la exigida no resultaron ser significativa.

Tabla 3.4 Efectos marginales del modelo *probit* para los centros "top-innovadores"

	Coef.	Std.Err.	Z	P>|z|
Cualificación superior a la exigida	-0,126	0,132	-0,96	0,34
Antigüedad media en la docencia	-0,041	0,009	-4,48	0,00
Satisfacción media del profesorado con su función	0,187	0,017	10,94	0,00
Satisfacción con el resultado académico de los alumnos	0,078	0,022	3,51	0,00
Nº medio de reuniones colectivas	0,044	0,012	3,75	0,00
Nº medio de horas que se atiende a cursos < 50 horas	0,004	0,000	13,34	0,00
Nº medio de horas que se atiende a proyectos de innovación	0,007	0,001	8,97	0,00
Nº medio de veces que se dedican al análisis de la práctica docente	0,015	0,003	4,87	0,00
Frecuencia media del uso de materiales elaborados	0,029	0,079	0,37	0,71
Frecuencia media del uso de prensa o revistas especializadas	3,221	0,094	34,18	0,00
Frecuencia media del uso de ordenadores e internet	-0,030	0,072	-0,42	0,67
Frecuencia del uso de medios audiovisuales	0,132	0,070	1,87	0,06
Necesidad de aspectos formativos relacionados con la enseñanza por competencias	0,667	0,064	10,36	0,00
Titularidad del centro (Privada)	-0,059	0,021	-2,85	0,00
Nº de profesores de secundaria que participan en el plan de formación	0,004	0,000	13,65	0,00
Constante	-2,921	0,097	-30,12	0,00

Fuente: Elaboración propia a partir de la información disponible en EGD 2010.

Una vez realizadas las comprobaciones para confirmar que el emparejamiento ha sido correctamente realizado (se reduce el sesgo por debajo del 5% tras el emparejamiento de forma significativa en prácticamente todas las variables, lo que permite utilizar con garantías la muestra emparejada para el análisis y que todas las observaciones se encuentran en la zona de soporte), nos centraremos ahora en los resultados del efecto que tiene el hecho de pertenecer a un colegio top innovador sobre los resultados de las cuatro competencias evaluadas. En la Tabla 3.5 se presentan los resultados tras realizar el PSM, donde podemos observar la existencia de diferencias significativas en favor del grupo de control en las competencias de comunicación lingüística y conocimiento e interacción con el mundo físico (-4,198 y -4,062, respectivamente), mientras que para las otras dos (matemáticas y social y ciudadana) no se registran diferencias significativas entre el grupo formado por las escuelas catalogadas como "top-innovadoras" y el resto. Estos resultados están en la misma línea de los obtenidos en algunos trabajos previos referidos al sistema educativo español (Gil et al., 2018, Gil y Cordero, 2018a, 2018b), en los que parece evidenciarse que el uso frecuente de prácticas docentes de tipo innovador por parte de los centros educativos no contribuye a mejorar los conocimientos de los alumnos, al menos los que pueden evaluarse mediante test estandarizados como los de PISA o la EGD.

Tabla 3.5. Efecto del tratamiento promedio de centros "top-innovadores"

Competencia	Muestra	Tratado	Control	Diferencia	D.E.	T-stat
Comunicación lingüística	ATT	502,898	507,096	-4,198	1,68	-2,49
Matemáticas	ATT	504,657	505,866	-1,208	1,73	-0,70
Mundo físico	ATT	504,111	508,174	-4,062	1,69	-2,39
Social y ciudadana	ATT	503,925	505,469	-1,544	1,71	-0,90

Criterio "top-clásico"

Según se desprende de la información mostrada en la Tabla 3.6, nuevamente la mayoría de las variables de control incluidas en el modelo de regresión *probit* resultan tener una influencia significativa sobre la probabilidad de que un centro sea catalogado como "top-clásico". Como cabía esperar, las variables que tenían una incidencia negativa y significativa en el caso anterior (titularidad privada y antigüedad media del profesorado), aquí no la tienen. Para la primera de ellas, la influencia sigue siendo significativa, pero ahora el signo es positivo, mientras que para la segunda el efecto deja de ser significativo. No obstante, sí hay otras variables que influyen negativamente, como el número de reuniones colectivas o la frecuencia de uso de determinados recursos, como prensa o revistas especializadas, ordenadores e internet, el número de profesores de secundaria que participan en el plan de formación y la necesidad de aspectos formativos relacionados con la enseñanza por competencias.

Tabla 3.6. Efectos marginales del modelo *probit* para los centros top-clásicos

	Coef.	Std. Err.	Z	P>|z|
Cualificación superior a la exigida	-0,233	0,144	-1,62	0,10
Antigüedad media en la docencia	-0,009	0,009	-1,06	0,29
Satisfacción media del profesorado con su función	0,152	0,017	9,15	0,00
Satisfacción con el resultado académico de los alumnos	0,332	0,023	14,67	0,00
Nº medio de reuniones colectivas	-0,134	0,012	-11,46	0,00
Nº medio de horas que se atiende a cursos < 50 horas	0,001	0,000	3,96	0,00
Nº medio de horas que se atiende a proyectos de innovación	0,001	0,001	1,71	0,09
Nº medio de veces que se dedican al análisis de la práctica docente	0,001	0,003	0,47	0,64
Frecuencia media del uso de materiales elaborados	0,856	0,078	11,02	0,00
Frecuencia media del uso de prensa o revistas especializadas	-0,329	0,091	-3,61	0,00
Frecuencia media del uso de ordenadores e internet	-0,642	0,069	-9,33	0,00
Frecuencia del uso de medios audiovisuales	0,229	0,069	3,32	0,00
Necesidad de aspectos formativos relacionados con la enseñanza por competencias	-0,180	0,060	-3,00	0,00
Titularidad del centro (Privada)	0,047	0,020	2,38	0,02
Nº de profesores de secundaria que participan en el plan de formación	-0,002	0,000	-6,26	0,00
Constante	-1,574	0,091	-17,31	0,00

Fuente: Elaboración propia a partir de la información disponible en EGD 2010

Por último, cabe señalar que hay variables que repiten como factores positivamente asociados con el fenómeno estudiado, como son las dos variables representativas del grado de satisfacción de los profesores, la participación en cursos de formación, el uso de materiales elaborados por el propio docente y el empleo de medios audiovisuales. Por último, los coeficientes correspondientes a la participación en proyectos de innovación e investigación y en grupos de análisis de la práctica docente no resultan concluyentes debido a su no significatividad.

Del mismo modo que para el primer análisis, tras realizar los controles necesarios para comprobar que el emparejamiento ha sido correctamente realizado, analizamos los resultados obtenidos correspondientes al efecto que tiene el tratamiento de pertenecer a un colegio top clásico sobre los resultados de las competencias evaluadas en la EGD 2010. En la Tabla 3.7 se presentan los resultados obtenidos tras la estimación del modelo PSM que en este caso sólo refleja diferencias significativas entre el grupo tratado y el de control para la competencia social y ciudadana a favor de los centros con un uso intensivo de estrategias clásicas, seguramente porque la adquisición de los conocimientos evaluados en esta competencia esté más ligada al uso de esquemas más tradicionales. Para el resto de competencias, las diferencias no resultan significativas.

Tabla 3.7. Efecto de tratamiento promedio de centros "top-clásicos"

Competencia	Muestra	Tratado	Control	Diferencia	D.E.	T-stat
Comunicación lingüística	ATT	508,435	508,694	-0,258	1,61	-0,16
Matemática	ATT	505,239	506,806	-1,567	1,66	-0,94
Mundo físico	ATT	509,016	508,175	0,840	1,63	0,51
Social y ciudadana	ATT	509,557	504,969	4,588	1,65	2,78

3.6. Conclusiones

Con el presente trabajo se ha tratado de profundizar en el estudio de la relación existente entre los estilos docentes utilizados por los profesores de educación secundaria españoles y el rendimiento académico de sus alumnos. Para ello se ha utilizado una fuente de información que apenas se había explotado previamente, como son los datos correspondientes a la Evaluación General de Diagnóstico realizada por el Instituto Nacional de Evaluación Educativa en el año 2010. Una de las ventajas que ofrece esta base de datos es que recopiló una amplia información relativa a los profesores respecto a las actividades que realizan en el aula, así como otras muchas actividades que componen su rutina diaria que ayudan a conocer el perfil del profesorado con el que cuenta cada centro educativo. Además, en esta prueba se evaluaron los conocimientos de los alumnos en cuatro

competencias distintas, lo que nos ha permitido obtener unos resultados más robustos, además de ser capaces de registrar algunas diferencias entre ellas en cuanto al efecto de las estrategias docentes.

El principal objetivo del artículo ha sido analizar si la especialización del profesorado de los centros educativos en un determinado estilo docente, ya sea clásico o innovador, contribuye a mejorar los resultados obtenidos por los alumnos. El criterio utilizado para considerar a una escuela como especializada en una u otra alternativa ha consistido en segmentar la muestra disponible según el valor que presenta en dos indicadores (clásico e innovador) construidos a partir de las respuestas de los profesores sobre la intensidad con la que realizan determinadas prácticas docentes y que pueden asociarse con cada uno de los estilos considerados. En concreto, se ha catalogado a un centro como especializado en un tipo de estrategia si se situaba dentro del 25% de centros con los mayores valores en alguno de los dos indicadores, de modo que podemos distinguir entre los centros "top-clásicos" y el resto y entre los centros "top-innovadores" y el resto.

A grandes rasgos, los resultados obtenidos en ambos análisis ponen de manifiesto que la especialización del profesorado de alguno de los dos estilos docentes evaluados no parece conducir a unos mejores resultados. En este sentido, compartimos con Zabalza (2011) la idea de que quizás la opción más apropiada sea una metodología mixta que combine prácticas de tipo tradicional con elementos más innovadores. Asimismo, hemos detectado que el uso intensivo de estrategias innovadoras, como pueden ser el trabajo en grupo, la exposición de trabajos o la realización de debates en clase, pueden llegar incluso a ser contraproducentes, un resultado que resulta especialmente relevante si se tiene en cuenta que las autoridades educativas de muchos países están fomentando el uso de este tipo de prácticas en detrimento de los métodos más tradicionales (Capps et al., 2012).

En todo caso, cabe señalar que estos resultados deben ser interpretarlos con cautela, puesto que nos referimos únicamente a la influencia que tienen las prácticas docentes sobre los resultados del proceso educativo que podemos medir, esto es, sobre las competencias demostradas por los alumnos en un test estandarizado de conocimientos. Por lo tanto, quedaría pendiente explorar los posibles efectos sobre determinadas actitudes y comportamientos como el pensamiento crítico o el trabajo colaborativo, sobre los que cabe pensar que los métodos innovadores de enseñanza deberían influir en mayor medida, aunque desafortunadamente no disponemos de una medida válida que nos permita realizar este tipo de evaluación.

Chapter 4

TEACHING STRATEGIES AND THEIR EFFECT ON STUDENT ACHIEVEMENT: A CROSS-COUNTRY STUDY USING DATA FROM PISA 2015

4.1. Introduction

One of the key findings from decades of educational effectiveness research is that teachers play the most relevant role in the student learning process (Rivkin et al., 2005; Hanushek, 2011). As a result, researchers and policymakers worldwide are increasingly interested in analyzing which makes a teacher more successful, including many different aspects such as their background characteristics, their abilities to communicate or their beliefs and attitudes towards teaching and students (Wayne and Youngs, 2003; Palardy and Rumberger, 2008; Boonen et al., 2014). This type of analysis has blossomed in recent years with the availability of large-scale datasets that link teachers to students´ test scores.

In this paper we focus our attention on examining the effectiveness of teaching strategies applied by secondary education teachers in their classes. Specifically, we refer to a wide range of processes and activities that cover classroom organization and resources, as well as the activities implemented in order to promote students´ acquisition of knowledge. Previous literature on this topic has been mainly focused on exploring the divergences between the traditional teaching practices based on lecturing and memorization and the emerging modern practices inspired by constructivist approaches involving student-oriented teaching and the extrapolation of knowledge learned to everyday problems (Opdenakker and van Damme, 2006; Seidel and Shavelson 2007; Van de Grift, 2014).

Despite the evidence available on the impact of different teaching styles on academic performance is stilly contradictory and inconclusive, educational authorities in many countries advocate a greater use of those modern practices in detriment of more traditional methods (Capps et al., 2012). Therefore, the study of the effectiveness of different teaching styles represents a topic of great relevance from a policy perspective. In particular, we intend to contribute to the existing debate about this topic by providing solid empirical evidence on the effects of the aforementioned teaching styles using an extensive dataset of more than 90.000 observations from 14 different countries participating in the most well-known educational survey worldwide: the Programme for International Student Assessment (PISA).

Specifically, we exploit the information contained in its most recent wave of this study, PISA 2015, which includes for the first time a teacher questionnaire completed by teachers about multiple aspects related to their background, attitudes and beliefs as well as the instructional activities applied in the classrooms, although this information is only available for science teachers, since science is the core domain in PISA 2015. In this respect, we should stress that the design of the

database does not allow us to establish a link between individual student information and the teachers who taught them. Therefore, when we refer to teaching strategies we rely on aggregated data built from the responses given by all the teachers in the same school, thus we implicitly assume that teachers working in the same school tend to adopt similar teaching approaches and even share teaching materials, developing what is known in the literature as a *teaching culture* (Echazarra et al. 2016). This is quite an innovative approach with respect to other earlier studies focusing on activities carried out by each teacher individually.

Demonstrating relationships between specific teaching practices and student achievement has proven difficult when using data from cross-sectional educational research designs because they do not provide information about students' prior achievement, thus we cannot know whether some teachers are applying certain strategies to adapt the characteristics of their students or whether students´ achievement depend on the implementation of those strategies (O´Dwyer et al., 2015). In addition, it is common that children from families with greater economic and cultural capital are likely to attend schools with better resources, thus there can be sorting of students across schools (Rothstein, 2010). Therefore, there can be a problem of unobserved heterogeneity or endogeneity in data that complicates the econometric estimation of teacher effects (Gustafson, 2013). In order to deal with this issue of selection bias and obtain consistent estimations, we apply student fixed effects models exploiting variation across different subjects assessed in PISA. In particular, we focus on exploring the influence of teaching practices on science achievement, since only science teachers provide information about activities carried out in their classes.

The rest of the paper is structured as follows. Section 2 provides a brief literature review about previous studies analyzing the effect of different styles of teaching. Section 3 describes the main characteristics of the dataset and the variables considered in our empirical analysis, while Section 4 presents our estimation strategy. The results from the empirical analysis are reported and discussed in Section 5. Finally, Section 6 outlines the main conclusions.

4.2. Literature review

Since the publication of the pioneer work by Woessmann (2003) using micro data from an international large-scale assessment, multiple studies have exploited information contained in those studies to explore the determinants of educational achievement from different perspectives (Hanushek and Woessmann, 2011; Strietholt et al., 2014; Johansson, 2016). These studies are mainly focused on the identification of significant causal relationships between students´

background and school-related variables and educational outcomes (typically represented by test scores) using econometric techniques[17]. Some examples are studies analyzing the effect of class size (Woessmann and West, 2006; West and Woessmann, 2006), instructional time (Lavy, 2015; Rivkin and Schiman, 2015) or the divergences in performance between public and private schools (Vandenberghe and Robin, 2004; West and Woessmann, 2010).

The literature devoted to analyze the effect of teaching practices on learning outcomes using data from international surveys is more limited, since the collection of information about these activities is complex. Actually, until recently, only two waves of the TIMSS (Trends in Mathematics and Science Study) database (2003 and 2007) provided data about instructional activities applied by teachers in the classroom (Mullis et al., 2012) and establish a link with students taught by those teachers. Therefore, it is not surprising that practically all of the internationally available evidence on the effect of teaching practices is based on data from those datasets.

The most frequent conclusion derived from those studies is the positive effect of traditional teacher-centred instruction. For instance, Schwerdt and Wuppermann (2011) analyze data from the sample of US students participating in TIMSS 2003 and conclude that traditional lecture style teaching has a positive impact on student achievement. De Witte and Van Klaveren (2014) found similar evidence for teaching styles based on problem solving and homework using data about Dutch students in TIMSS 2003, as well as House (2009) and Bietenbeck (2014) for Japanese fourth-grade students and US eight-grade students, respectively, using data from TIMSS 2007. Zuzovsky (2013) also found a positive relationship with achievement in a cross-country analysis using data about the 49 countries participating in TIMSS 2007. Nevertheless, Van Klaveren (2011) found that lecturing in front of the class has no significant effect on student outcomes in an empirical analysis based on TIMSS 2003 data from the Netherlands.

The evidence about the effect of modern teaching practices based on TIMSS data is more mixed. In this sense, Zuzovsky (2013) found that the implementation of constructive modes of instruction focused on students was positively associated with learning outcomes in high- and medium-achieving countries, but negatively associated in low-achieving countries. An alternative interpretation of these results is that these types of teaching practices possibly have more to do with students developing other skills, such as reasoning ability (Bietenbeck, 2014), or improving their social capital (Algan et al., 2013) than with the acquisition of the knowledge or competencies that are usually assessed by international knowledge tests.

[17] See Cordero et al. (2017) for a detailed review of this literature.

Nevertheless, we can also find some insights about teacher effects in some recent studies using data from other well-known international studies conducted by the OECD like PISA, TALIS (Teaching and Learning International Survey) or PIAAC (Programme for the International Assessment of Adult Competencies). However, in those cases the link between teachers and students is not straightforward, so researchers need to develop some alternative ways of exploiting the available data. For example, Hanushek et al. (2014) and Meroni et al. (2015) identify that part of the variation in student performance measured in PISA can be explained by teacher cognitive skills measured by PIAAC aggregated at country level. Likewise, Le Donné et al. (2016) have exploited the information contained in the so-called TALIS-PISA link database, which connects the available data about students in PISA 2012 with data about teachers aggregated at school level from TALIS 2013, to explore the effects of teaching strategies implemented by teachers belonging to the same school. Their results show that modern teaching strategies based on cognitive activation and active learning strategies present a strong association with students' achievement in mathematics. In contrast, Gil-Izquierdo and Cordero (2018) do not find any significant relationships for teaching strategies in their empirical analysis based on Spanish data from the TALIS-PISA link database.

An alternative option would be using data provided by students about the strategies implemented by their teachers as suggested by Caro et al. (2016). Those authors found that modern cognitive activation strategies were positively and consistently related to mathematics performance across education systems, especially in schools with a positive disciplinary climate and for students from advantaged socio-economic backgrounds. The traditional teacher-directed strategies are also found to be positively related to mathematics performance, although this association becomes negative for higher levels of teacher-directed instruction. The main limitation of this approach is that student responses might be subject to bias, since they may be influenced by personality characteristics of the teacher or by their grades (Goe et al., 2008).

4.3. Data and variables

Our empirical analysis is based on data provided by OECD´s PISA (Programme for International Student Assessment) 2015 survey. This international study assesses the knowledge and skills of 15-year-old students in three main areas (reading, mathematics and sciences). The survey takes place every three years, starting in 2000, thus PISA 2015 represents the sixth wave of this study. For each assessment, one competence is chosen as the major domain and given greater emphasis. In 2015, the key competence was science (as well as in 2006). PISA's biggest potential is that it provides comparable data for a very wide-ranging set of countries (72 in 2015) about students´ background

and the characteristics of their schools collected from questionnaires completed by students and school principals. In addition, one of the major PISA 2015 innovations was the existence of an optional teacher questionnaire that gathers information about teachers´ training, experience and instructional activities[18]. These data were only provided by 18 countries, although some of them presented very few observations or many missing variables in relevant questions, thus in our empirical analysis we analyze data about 14 countries[19]. The total number of available observations available for each country in our dataset is reported in Table 4.1.

Table 4.1 Dataset composition

	Observations	%	Cum. %
AUSTRALIA	10,833	11.98	11.98
BRAZIL	12,028	13.30	25.29
CHILE	4,853	5.37	30.66
COLOMBIA	7,759	8.58	39.24
CZECH REPUBLIC	6,304	6.97	46.21
GERMANY	4.204	4.65	50.86
SPAIN	6,218	6.88	57.74
HONG KONG	5,172	5.72	63.46
ITALY	8,169	9.04	72.50
KOREA	3,999	4.42	76.92
PERU	3,912	4.33	81.25
PORTUGAL	5,681	6.28	87.53
CHINA-TAIPEI	6,378	7.05	94.59
UNITED STATES	4,895	5.41	100
TOTAL	**90,405**	**100.00**	

Source: Own elaboration from PISA 2015 database

PISA collects its data in a two-stage clustered sampling design (Willms and Smith, 2005). In the first stage of sampling, schools having age-elegible students are sampled systematically with probabilities proportional to the school size. A minimum of 150 schools is selected in each country. Subsequently, 35 15 year-old students are randomly selected from each school to participate in the survey. The sampling design thus implies that all students are observed three times in the data (once in science, once in maths and once in reading). In order to account for this complex sampling design, student sampling weights provided with the PISA database are used throughout the analysis (Rutkowski *et al*., 2010).

[18] This questionnaire is partly based on instruments previously established in the OECD Teaching and Learning International Survey (TALIS).

[19] Countries excluded were United Arab Emirates, Dominican Republic and two Chinese areas (Macao and the group of cities and regions named Beijing-Shanghai-Jiangsu-Guangdong).

One of the main advantages of using PISA data is that this study does not evaluate cognitive abilities or skills through using one single score but each student receives different scores (plausible values) that represent the range of abilities that a student might reasonably have (see OECD, 2014 for details). Specifically, the dataset provides measures on student achievement based upon pupils' responses to different test booklets, each of which includes only a limited number of test questions. Thus, it is difficult to make claims about individual performance with great accuracy. Using a complex process based on item response theory model, the survey organizers produce test scores for each domain taking into account the difficulty of each test question[20]. Plausible values can therefore be defined as random values drawn from this distribution of proficiency estimates (Mislevy et al., 1992; Wu, 2005). In our analysis we use these plausible values as our dependent variables. Although the dataset provides ten plausible values for each competence, we only consider the first one, since on large samples using one plausible value or more does not really make a substantial difference (OECD 2009, p. 44). The original values have a mean of 500 and standard deviation of 100, but we have converted them into international z-scores (meaning that the mean is 0 and the standard deviation is 1) by subtracting the mean score achieved amongst all pupils in the 14 countries and dividing by the standard deviation. This facilitates the interpretation of the estimated parameters in our regressions[21].

Given that the main focus of this paper is exploring the effect of teaching practices, it is worth mentioning that only science teachers report this information (10 in each school), thus we only rely on data about teachers to build our indicators. These teachers had 30 minutes to complete a computer-based questionnaire module, including wide-ranging questions about their background, previous experience and teaching strategies that they applied in their classrooms. For the latter, teachers are required to focus on a class attended by 15-year-old students that they teach and answer the question "*How often does this happen in your -school science- lessons?*". This question is followed by a description of the activities: "Students are asked to draw conclusions", "I demonstrate an idea", "Students read materials from a textbook", etc. The responses adopt the form of a Likert scale with four possible alternatives: (1) Never or almost never; (2) Some lessons; (3) Many lessons; (4) Every lesson or almost every lesson. For the sake of simplicity, we have rescaled answers to each item by assigning a proportional value as follows: (1) 0%, (2) 33%, (3) 66%, and (4) 100%. Accordingly, the responses stand for the percentage of time spent on each of the teaching activities carried out by each teacher. Note that the teaching methods used are not mutually exclusive, that is,

[20] See Von Davier and Sinharay (2013) for further details.
[21] See Brown et al. (2007) for details.

each teacher can apply more than one teaching method in the same class, while spending a different amount of time on each activity.

As there is a wide-ranging set of questions related to classroom practices, we need to build indicators to summarize the information provided by teachers. Indeed, studies typically assess teaching practices on scales and levels and therefore rule out the existence of a pure or absolute approach (Caro et al., 2016). Following Bietenbeck (2014) we rely on the on the taxonomy by Zemelman et al. (2005) to classify teaching strategies as reflecting either a traditional or a modern style. Table 4.2 displays the specific questions associated with each alternative.

Table 4.2 Classification of variables about teaching practices (traditional and modern)

Teaching style	**Code in PISA**	**Activities**
Traditional	TC037Q03NA	The teacher explains scientific ideas
	TC037Q10NA	The teacher discusses questions that students ask
	TC037Q12NA	Students write up laboratory reports
	TC037Q13NA	The teacher demonstrates an idea
Modern	TC037Q01NA	Students are asked to draw conclusions
	TC037Q02NA	Students are given opportunities to explain their ideas
	TC037Q05NA	A whole class discussion takes place
	TC037Q06NA	Current scientific issues are discussed

Source: Own elaboration from PISA 2015 database

After building the indicators representing the activity carried out by each teacher (how often he or she applies each of the teaching strategies), we clustered this information at school level, since this is the level of analysis recommended by PISA (see OECD, 2016, p. 15)[22]. This implies assuming that we assess the performance of teachers from the same school as a whole, we implicitly assume that teachers from the same school tend to adopt more similar teaching approaches and even share teaching materials, thus fostering a school "*teaching culture*" (Echazarra et al., 2016). In order to illustrate this, Table 4.3 shows that the part of the variance due to the between school-variance is smaller than the within-school variance for both teaching strategies in all countries. This implies that teachers from the same schools tend use teaching strategies (modern and traditional) with more similar levels than teachers from other schools.

[22] Note that it is not possible to establish an exact connection between students and teachers because the sampled students from each school may be members of different classes. In addition, in secondary schools it is frequent that teachers rotate around different classes.

Table 4.3 Distribution of total variance of teaching indices between and within schools

	Traditional style			Modern style		
	Total variance	Between-school variance	Within-school variance	Total variance	Between-school variance	Within-school variance
AUSTRALIA	70.20	14.20	56.00	89.74	18.63	71.11
BRAZIL	69.19	17.95	51.24	70.97	16.12	54.85
CHILE	22.25	5.32	16.92	24.61	5.60	19.01
COLOMBIA	29.28	8.22	21.06	36.01	10.61	25.40
CZECH REPUBLIC	50.64	11.03	39.61	51.34	9.74	41.60
GERMANY	43.42	8.14	35.28	43.46	5.68	37.78
SPAIN	30.49	5.63	24.86	47.17	8.50	38.67
HONG KONG	18.93	3.17	15.76	23.76	3.93	19.83
ITALY	55.43	12.52	42.91	65.91	14.34	51.57
KOREA	20.46	3.47	16.99	23.47	4.29	19.18
PERU	21.47	4.93	16.54	20.88	5.22	15.66
PORTUGAL	27.90	4.10	23.80	37.45	5.40	32.05
CHINA-TAIPEI	33.19	5.39	27.80	40.43	6.30	34.13
UNITED STATES	30.18	6.12	24.06	34.17	6.65	27.52

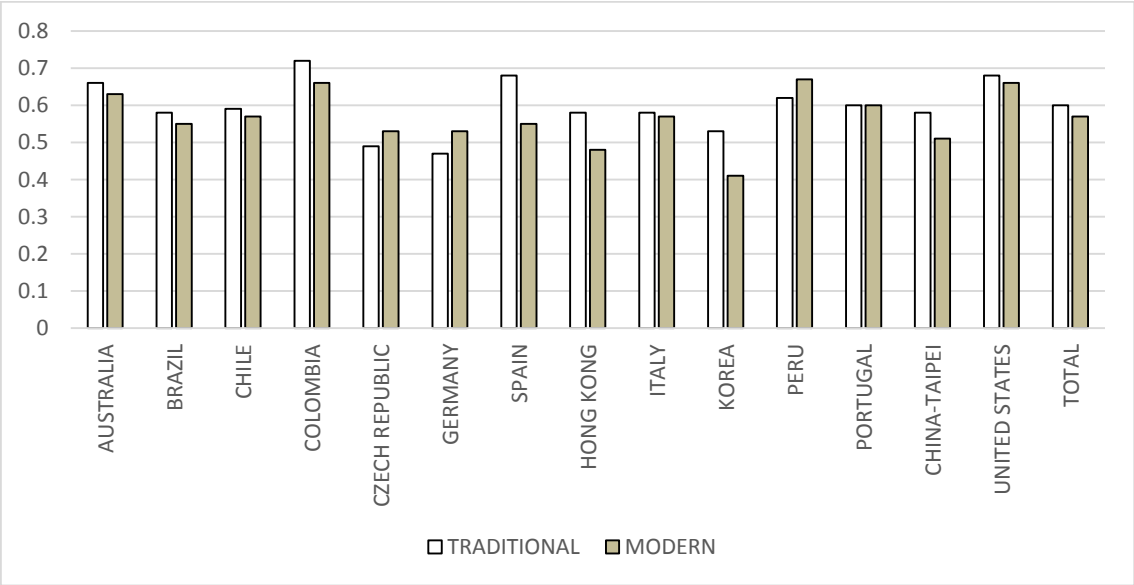
Source: Own elaboration from PISA 2015 database

In terms of interpretation, greater values indicator values indicate that teachers from a school apply this type of strategies more often in their science lessons. However, it is important to keep in mind that these teaching strategies are not mutually exclusive, since teachers can combine different styles of teaching during their classes. Deepening into the values of the indicators reflecting the time devoted to each teaching styles, Figure 4.1 displays the distribution of the mean values by country. In general terms, we cannot observe large differences in the use of different teaching stiles across different educational systems since in most cases the indices present mean values in a short range (between 0.5 and 0.7). Likewise, we do not observe relevant divergences between the mean values of modern or traditional indices, although it is noteworthy that traditional strategies are more commonly applied in most countries.

In the process of building the teaching style indicators, the data were debugged according to several methods in order to be able to guarantee data reliability. On one hand, we only accounted for teachers who answered 100% of the questions covered by the teaching strategy indicators. As these variables were the main targets of our study, we decided not to use any missing value imputation procedure, relying exclusively on the information provided by the teachers that provide all the responses. On the other hand, when we detect a low response percentage in a school, observations from that school are deleted. Specifically, we dropped schools where the response percentage was

less than 20%. This procedure was designed to assure that the strategies reported by only one teacher or very few teachers were not attributed to all teachers in the school[23].

Figure 4.1. Mean values of the teaching indices by countries



Source: Own elaboration from PISA 2015 database

The information provided by science teachers is also used to create a set of control variables that are also included in our estimations to account for the different types of teachers existing at each school. Those variables are their average age, their mean years of experience and the percentage that has higher professional qualifications than required. The consideration of these variables allows us to discount their effect on student achievement when analyzing the influence of the different teaching strategies.

Finally, the database that contains the information on teachers aggregated at school level has been merged with the databases including data about students and schools using the school identifier provided by PISA. As a result, we have a data file containing detailed information on student characteristics, the school that they attend and the teachers belonging to that school. In order to account for potential divergences in student background, in our regressions we also include a set of control variables at student level that have been frequently identified as influential factors in previous literature, such as gender, being a repeater, being an immigrant, having a computer, mother´s level of education or the number of books at home. Likewise, we consider several covariates at school level such as the type of ownership, the location and the average socio-

[23] Bietenbeck (2014) also followed a similar criterion to guarantee a minimum level of precision in the measurement of teaching practices.

economic status of students in the school (ESCS) as a proxy of the peer effect[24]. Table 4.4 shows the descriptive statistics of all the variables classified in three blocks: those related to the student, the ones referred to teachers (aggregated at the school level) and the school variables.

Table 4.4 Descriptive statistics

	Mean	S.D.	Min.	Max.
Individual level				
Science z-scores (PV1SCI)	0.00	1.00	-3.87	3.85
Mathematics z-scores (PV1MATH)	0.00	1.00	-4.44	4.04
Reading z-scores (PV1READ)	0.00	1.00	-4.35	3.67
Gender (Female)	0.51	0.50	0	1
Repeater (1 if the student has repeated a grade)	0.17	0.38	0	1
Immigrant (1 if the student was born in other country)	0.04	0.19	0	1
OwnCPU (1 if the student has an own computer)	0.87	0.34	0	1
Mothedu (Mother has a post-lower secondary education)	0.70	0.46	0	1
Books200 (There are more than 200 books at home)	0.18	0.39	0	1
Teachers at school level				
Teachage (Teachers´ mean age)	43.75	5.01	27.6	59.8
Teachexp (Teachers´ years of experience)	17.25	4.77	3.5	33.74
Teachqualif (Proportion of teachers having a qualification higher	0.37	0.49	0	1
Traditional style index	0.60	0.11	0.22	1
Modern style index	0.57	0.10	0.23	1
School level				
Private school (1 if the school is private)	0.28	0.45	0	1
Rural school (1 if the school is located in a town or village)	0.19	0.39	0	1
ESCSmean (Average index of economic, social and cultural	-0.31	0.72	-2.83	1.53

Source: Own elaboration from PISA 2015 database

4.4. Empirical strategy

We start with a simple model of the relationship between student achievement and several potential determinants including teachers´ instructional practices:

$$Y_{ijk} = \alpha_{ijk} + \delta_k TS_k + \beta_{ik} X_{ik} + \lambda_k Z_k + \varepsilon_{ijk} \qquad I = 1, \dots, N; J = 1, \dots, j \qquad (4.1)$$

[24] ESCS (index of economic, social and cultural status) is prepared by PISA specialists combining information on parent educational level and occupational status with household indicators.

where N is the number of pupils; J is the number of subjects (in our case S = 3); Y_{ij} denotes the test score of student i in subject j in school k, TS represents an aggregate indicator of teaching strategies reported by teachers in the same school, X is a vector of student-level background characteristics that vary only across students but not across subjects, Z is a vector of controls at school level, including average characteristics of teachers in the school, and ε_{ijk} is the usual error term that contains all unobservable influences on student test scores.

In our empirical estimation is important to bear in mind that the PISA sample has a hierarchical structure (students are 'nested' in schools), which makes the average correlation among the variables for students within the same school higher than that between students from different schools (Hox, 2002). Therefore, we need to use multilevel models in our simultaneous estimation of the effects of variables belonging to different levels (Bryk and Raudenbush, 1992; Goldstein, 1995). Moreover, as we are interested in exploring whether the influence of teaching strategies might be different across the distribution of results, we also estimate quantile regression models (Koenker and Bassett, 1978), also with a multilevel structure. This approach estimates parameters for different science score cut-off points (e.g. quartiles or deciles). Accordingly, the slopes for the dependent variable may vary. Moreover, we also estimate interquartile regressions in order to check whether the differences between the estimates for different cut-off points are significantly different.

The aim of the baseline model is to establish a relationship between different teaching strategies adopted by science teachers from the same school and student achievement in this domain, thus the parameter of interest is δ. However, estimating equation (1) using multilevel least squares or quantile regressions produces biased estimates if unobserved school characteristics and the teaching style are correlated. This can be the case if there exists sorting of high ability students or effective teachers into schools. For instance, it is frequent that children from families with greater economic and cultural capital are more likely to attend schools with better resources. Likewise, more experienced teachers tend to be concentrated on schools with students coming from more favorable socioeconomic conditions.

To tackle this issue, we apply a cross-subject student fixed effects model that utilizes variation within the same student but across different subjects following a similar strategy employed in some previous works (e.g. Dee, 2007; Schwerdt and Wuppermann, 2011; Bietenbeck, 2014; Zhakarov et al., 2014). Basically, the identification strategy establishes that student and school characteristics are the same for the three subjects under analysis, so we can triplicate the observations in our

datasets and apply student fixed effects in order to control for any subject-invariant student-level factors or any school characteristics that might influence test scores.

Although the availability of data about teachers' strategies at school level could be interpreted as a potential limitation of data, actually this may contribute to mitigate potential bias in the estimation due to the fact that students are generally not randomly assigned to teachers, who typically adapt their teaching style to the characteristics (or abilities) of the students in the class. As we are assessing the influence of the teaching culture of the school, our estimation should not be affected by this potential bias caused by within-school sorting across classes. In addition, using aggregate data representing the teaching culture of the school allows us to avoid potential errors of measurement that are more common in individual data (Rosen and Gustafson, 2016).

Finally, it is worth mentioning that in our model we are implicitly assuming that teaching practices indices are uncorrelated with the error term conditional on the other regressors. Nevertheless, this assumption could be violated if teachers who make a more intensive use of certain teaching practices had particular other unobserved characteristics (e.g. more motivation) that may be related to students' cognitive skills. This omitted-variable problem is a challenge faced by most researchers attempting to estimate any potential effect of teaching activities or behaviors. In our case, we have tried to address this potential concern by controlling by several teacher characteristics.

4.5. Results

In this section, we present the results of the proposed models. As a starting point, Table 4.5 presents the results of estimating equation (1) using a least squares multilevel approach and considering test scores in sciences as the dependent variable, since this is the main domain assessed in PISA 2015. Regressions have been estimated including both traditional and modern teaching strategies together, since in-class work frequently encompass, and likely require, different types of instructional activities. Therefore, since teaching strategies are complementary, their effectiveness cannot be judged in isolation.

As explained above, we include a set of covariates representing students´ background, as well as control variables representing the main characteristics of teachers from the same school and school variables. All the analyses were conducted using the sampling weights included in PISA to correct for non-response bias and scale the sample up to the size of the national population. Likewise, we make the appropriate adjustment to the estimated standard errors (bootstrapping standard errors by

cluster) to account for this clustering of students within schools[25]. Finally, in the context of a cross-country study, we are also interested in accounting for unobserved heterogeneity across different education systems, thus we incorporate country fixed effects in our estimations.

Table 4.5 Estimates using a least squares multi-level regression approach

Variables	
Traditional style index	0.582***
	(0.119)
Modern style index	-0.409***
	(0.113)
Gender	-0.182***
	(0.00637)
Repeater	-0.591***
	(0.00904)
Immigrant	-0.0820***
	(0.0168)
OwnCPU	0.226***
	(0.0102)
Mothedu	0.137***
	(0.00761)
Book200	0.311***
	(0.00849)
Teachage	-0.0255***
	(0.00404)
Teachexp	0.0250***
	(0.00399)
Teachqualif	0.0123
	(0.0185)
Private	-0.0677***
	(0.0215)
Rural	0.0211
	(0.0199)
ESCSmean	0.544***
	(0.0153)
Constant	0.593***
	(0.160)
Observations	94,609
Country fixed effects	Yes
Number of groups	14

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

[25]. Estimates are bootstrapped by cluster (schools) using 50 replications to calculate approximate standard errors (see OECD, 2013 for details).

The results indicate that a more intensive use of traditional teaching practices is positively and significantly associated with student achievement in sciences. This is in line with some of the evidence available in the previous literature for a single country (Schwerdt and Wuppermann, 2011; Bietenbeck, 2014). In contrast, we find that the relationship with devoting more time to modern practices is negative and also significant, which is also consistent with some previous studies (Brewer and Goldhaber, 1997). For the individual control variables we find the expected values of the coefficients, i.e. a negative correlation with gender and the conditions of repeater and immigrant and positive for mothers´ level of education and home possessions (books and computer). With regard to teachers´ variables, there is a negative relationship with the age and positive with experience, while only the type of school attended seems to have a significant influence (negative) among school variables.

However, this model does not account for non-random assignment of students to schools, thus there might have some bias in the estimations. In order to address this issue we implement the student fixed effect model described in the previous section. The results reported in Table 4.6 confirm the positive and significant effect of using traditional teaching practices more intensively. This evidence is consistent with previous results obtained in some works using data from international large-scale surveys (e.g. Lee and Huh, 2014; Bietenbeck, 2014) and also others focused on science (von Secker, 2002; von Secker and Lissitz, 1999). On the other hand, for modern teaching practices we find now a positive impact on student performance, although it is much smaller than the estimated coefficient for the traditional teaching practices. Some recent studies based on PISA data also found a positive relationship between modern cognitive activation strategies and student achievement in mathematics (Echazarra et al., 2016; Le Donné et al. 2016), although it is concentrated in some specific countries, especially those with higher levels of achievement (Zuzovsky, 2013). The rest of the control variables remain similar to the previous OLS estimation.

Table 4.6 Estimates using a student fixed effects model

Variables	
Traditional style index	0.739***
	(0.027)
Modern style index	0.0617**
	(0.024)
Gender	-0.0925***
	(0.00413)
Repeater	-0.597***
	(0.00502)
Immigrant	-0.0793***
	(0.00982)
OwnCPU	0.211***
	(0.00597)
Mothedu	0.137***
	(0.00442)
Book200	0.281***
	(0.00495)
Teachage	-0.0434***
	(0.00391)
Teachexp	0.0452***
	(0.00395)
Teachqualif	-0.0346*
	(0.0180)
Private	0.0192
	(0.0212)
Rural	-0.0523**
	(0.0219)
ESCSmean	0.258***
	(0.0133)
Constant	0.543***
	(0.131)
Observations	283,827
Country fixed effects	Yes
Number of groups	14

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

With the aim of studying more in depth the effects of each teaching strategy, we have also estimated our fixed effects model adopting a quantile regression approach considering different segments of the science score distribution. Specifically, we have estimated the interquartile regressions that test

for differences between the quartiles (0.75 and 0.25) and the extreme deciles (0.9 and 0.1). The results reported in Table 4.7 indicate that there are non-significant differences for both teaching styles between quartiles 75% and 25%, but the divergences are significant if we compare the tails of the distribution of results, i.e. the 90th and the 10th percentiles[26]. Thus, traditional strategies appear to have more relevant impact on students with lower results, while the effect of modern practices is more intense on high achievers. These results are consistent with previous evidence obtained in other cross-country studies by Zuzovsky (2013) and Caro et al. (2016).

Table 4.7 Results of interquartile regressions

	0.25-0.75	**0.1-0.9**
Traditional style index	-0.00055	0.128**
	(0.0374)	(0.0513)
Modern style index	0.0127	-0.132***
	(0.0430)	(0.0439)

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Finally, we propose an alternative specification for our regressions in which we evaluate interactions of our variables of interest (traditional and modern indices) with several variables representing the school context. Our aim here is to detect whether there are differences between the intensity of use of both teaching strategies if they are combined with some variables related to the characteristics of the school such as having students with higher or lower socioeconomic status (determined by the 50% of the mean ESCS), the location (in a rural or urban area) or the type of ownership (public or private). The results of the coefficients estimated using those interactions are summarized in Table 4.8[27].

Regarding ESCS at school level, results are significant and positive for traditional strategies for both levels of socioeconomic status, but we notice relevant divergences for modern strategies. Thus, the coefficient has a negative value for students form a more disadvantaged background, but positive for those with the highest level of ESCS. This result corroborates the evidence showed above about the concentration of the effect on modern teaching practices on better students, since socioeconomic status is frequently correlated with better academic achievement. Turning now to the effect of location and ownership, we observe many similarities. In both cases, we cannot detect relevant divergences in the effect for classical strategies, which is positive for all schools. Nevertheless, the effect of modern practices is only significant (and positive) in urban and private

[26] For the sake of simplicity, we only show the results for the main variables of interest, although all the estimations are calculated with the same control variables.
[27] Again, we only show the results for the main variables of interest, but estimations also include control variables.

schools. This finding suggests that students coming from those contexts might have some higher-order thinking skills that facilitate their engagement with this style of teaching (Gao, 2014; Caro et al., 2016). Therefore, the implementation of this type of strategies needs still more development in order to be effective in schools with students not having those skills.

Table 4.8 Estimates using a student fixed effects model: teaching strategies interactions with school variables

		Coefficient	Standard Error
Traditional style index	Low ESCS	0.726***	(0.0291)
	High ESCS	0.756***	(0.0280)
Modern style index	Low ESCS	-0.148***	(0.0301)
	High ESCS	0.358***	(0.0299)
Traditional style index	Urban	0.729***	(0.0263)
	Rural	0.785***	(0.0514)
Modern style index	Urban	0.0849***	(0.0270)
	Rural	-0.0549	(0.0547)
Traditional style index	Public	0.763***	(0.0283)
	Private	0.660***	(0.0518)
Modern style index	Public	0.0193	(0.0269)
	Private	0.177***	(0.0500)

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

4.6. Conclusions

This paper has examined the relationship between different styles of teaching applied by teachers in their classes and the results achieved by pupils using data about 14 countries participating in PISA 2015. This is the first wave of PISA that includes information supplied directly by teachers about a broad spectrum of teaching activities conducted by (science) teachers within the classroom. Regarding this point, it is worth mentioning that the characteristics of the database do not allow us to establish a link between each teacher and their students, thus our analysis have been focused on exploring the effectiveness of different indicators representing the teaching style of the school, i.e. assuming that teaching activities conducted by teachers belonging to the same school are similar. This is a major difference with regard to most part of the existing evidence in the literature, which refers to the teaching practices carried out by a single teacher.

Our estimations are based on a student-fixed effects approach holding also country effects constant to disentangle the role of teaching practices from national educational policies. The results suggest that both classic and modern strategies have a positive impact on students´ science achievement, although the impact of the former is clearly more relevant. In other words, students whose teachers

who report a more frequent use of classical activities such as lecturing or repetition of tasks perform better than those using more modern activities focused on promoting the involvement of students. Additionally, a more detailed analysis of these effects has allowed us to fine-tune these conclusions with some new insights. In particular, we detect that, while the positive influence of traditional practices maintains almost constant regardless of the characteristics of students or schools, the implementation of modern strategies only appears be effective in high-achieving students and schools with certain characteristics (private, urban and with students from more favourable socioeconomic backgrounds). These results has important policy implications for the design of educational policies, especially in educational systems in which authorities are promoting a more intensive use of modern constructivist approaches in preference to more traditional learning methods, since we provide evidence suggesting that that this policy might widen the gap between disadvantaged and advantaged students and schools. In particular, we should reflect on the potential educational inequalities that may arise if only better-off schools are able to implement successfully modern strategies in terms of improving achievement.

Finally, we should note certain limitations of our analysis that are common in most studies using data from large-scale assessments like PISA. First, we rely on teachers' self-reported instructional activities, which cannot capture complexity of teaching processes as much as direct classroom observations (Caro et al., 2017). Moreover, the responses given by teachers might include some potential bias due to social desirable responses, especially with regard to the actual implementation of modern strategies that usually require more time than is available. However, we believe that the consideration of aggregate data at school level mitigates this potential shortcoming to a certain extent. Second, the impossibility of establishing a link between students and teachers due to the characteristics of the dataset may question the validity of causal inference estimates. However, the aggregation of data brings other advantages as well. For example, grouped data at school level have the advantage of not being as severely influenced by errors of measurement. Finally, our measure of the educational outcome only reflects the academic achievement, which implies that other higher order skills that may be equally important such as critical thinking or the ability to express and reason ideas are not considered in our empirical analysis (Muijs, 2006). This is serious limitation that might explain the smaller influence of modern teaching practices that have been previously identified in the literature to be linked with the successful completion of higher level tasks (Bietenbeck, 2014).

CONCLUDING REMARKS

The main objective of this PhD thesis has been the analysis of the influence of teaching practices on students' achievement. As we were interested in obtaining evidence in terms of causality, the first step of the research was to carry out an extensive review of the previous studies that had used causal inference techniques with educational data from large-scale assessments. Subsequently, we have conducted three empirical studies using alternative methodological approaches, so that we can provide consistent evidence about how the implementation of innovative/modern or traditional/classical teaching strategies in secondary schools contributes to promote students´ attainment in standardized tests both in the specific case of Spain (chapters 2 and 3) and in the international context (chapter 4).

The results presented in the second chapter of this thesis can be considered as a starting point, since they are based on a multilevel regression analysis using PISA 2015, thus they not provide evidence in terms of causality. However, they point out some relevant ideas such as the positive and significant association found between using more intensively traditional practices and students´ test scores or the opposite finding for modern teaching strategies and results. In the third chapter we have fine-tuned the purpose of the analysis by focusing on schools specialized in the use of each of those strategies. Hence, we have divided our sample of schools according to the style of teaching adopted, so that we can identify a treated and a control group. Then, we have applied propensity score matching to control for potential estimation bias. Again, the results indicate that schools focused on implementing modern instructional activities obtain worse results, although only in some competences. Likewise, schools specialized in traditional strategies achieve similar results or better. Therefore, maybe the Spanish educational authorities should rethink the idea of promoting that teachers should mainly apply innovative practices in their classes to the detriment of traditional teaching strategies, since the latter seems to be more effective.

The international evidence reported in the fourth chapter suggests the existence of important similarities with the Spanish case. In particular, after conducting a cross-country analysis using data about 12 countries participating in PISA 2015 and adopting a fixed-effect approach to mitigate potential bias in the estimation, we found that traditional teaching practices have a positive and significant effect on student achievement. Modern strategies present a positive but less relevant impact on performance, although we detect that the gap between disadvantaged and advantaged students seems to be larger in systems where modern strategies predominate over traditional methods, thus it seems that those practices are not effective for all types of students.

Despite the relevance of our results, there is still substantial scope for further empirical research on the relationship between teaching practices and student performance. However, a more in depth study of the effectiveness of specific activities carried out by some teachers would require having available data about different teachers that can be linked to students and, ideally, to have access to longitudinal datasets that allow us to use some alternative approaches to deal with the potential existence of endogeneity in data such as difference-in-differences or value-added models. Unfortunately, these possibilities did not exist in the databases that we have used in the present thesis. Thus, we have been forced to focus on exploring the aggregate effects of the teaching culture of the school and being creative in attempting to emulate the characteristics of an experimental design to avoid potential estimation bias.

To conclude, the results of this PhD thesis aim to give empirical-based ground for a public debate about the effectiveness of educational policies. In particular, the most relevant issue addressed in this thesis has been the impact of teaching strategies, but there are many other aspects related to teachers that also deserve more attention in the literature. For instance, how should teachers interact with students? To what extent new technologies should be employed in the classes replacing traditional lectures? This research provides food for thoughts to new lines of future research. From our viewpoint, the role of a teacher should not only be limited to explain the contents included in the school curriculum, but also help students think and develop their capabilities and their own ideas about the concepts studied. Unquestionably, in the upcoming years new professional profiles will be demanded, since it is necessary that teachers are able to address correctly the needs of students with different characteristics. Fostering new and more flexible professionals seems to be nowadays one of the biggest challenges for our society.

REFERENCES

Agasisti, T. and 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.

Akiba, M., LeTendre, G. K. and Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher*, 36(7), 369-387.

Algan, Y., Cahuc, P. and Shleifer, A. (2013). Teaching practices and social capital. *American Economic Journal: Applied Economics*, 5(3), 189-210.

Álvarez-Morán, S., Carleos, C.E., Corral, N.O. and Prieto, E. (2017). Metodología docente y rendimiento en PISA 2015: Análisis crítico, *Revista de Educación,* 379, 85-114.

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

Arikan, S., van de Vijver, F. and 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.

Austin. B., Adesope, O., French, B., Gotch, C., Belanger, J. and Kubacka, K, (2015). Examining school context and its influence on teachers: linking TALIS 2013 with PISA 2012 student data. *OECD Education Working Papers*, No. 115, Paris, OECD Publishing.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A. and Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133-180.

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

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates?. *The Quarterly Journal of Economics*, 249-275.

Betz, T. (2013). Robust Estimation with Nonrandom Measurement Error and Weak Instruments. *Political Analysis*, 21(1), 86-96.

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

Boonen, T., Van Damme, J. and Onghena, P. (2014). Teacher effects on student achievement in first grade: which aspects matter most?, *School Effectiveness and School Improvement*, 25(1), 126-152.

Brewer, D. J. and Goldhaber, D. D. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 32(3), 505-523.

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

Brown, G., Micklewright, J., Schnepf, S. V. and Waldmann, R. (2007). International surveys of educational achievement: how robust are the findings?, *Journal of the Royal Statistical Society: Series A (statistics in society)*, 170(3), 623-646.

Calero, J. and Escardíbul, J.O. (2007). Evaluación de servicios educativos: el rendimiento en los centros públicos y privados medido en PISA-2003. *Hacienda Pública Española / Revista de Economía Pública*, 183-4, 33-66.

Caliendo, M. and Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1), 31-72.

Camburn, E.M. and Han, S.W. (2011). Two decades of generalizable evidence on U.S. instruction from national surveys. *Teachers College Record*, 113(3), 561–610.

Campbell, J., Kyriakides, L., Muijs, D. and Robinson, W. (2012). *Assessing teacher effectiveness: different models*. London: Routledge Falmer.

Capps, D. K., Crawford, B. A. and Constas, M. A. (2012). A review of empirical literature on inquiry professional development: Alignment with best practices and a critique of the findings, *Journal of Science Teacher Education*, 23(3), 291-318.

Carbonaro, W. J. and Gamoran, A. (2002). The production of achievement inequality in high school English. *American Educational Research Journal*, 39, 801–827.

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

Caro, D. H., Lenkeit, J. and Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012, *Studies in Educational Evaluation*, 49, 30-41.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633-2679.

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. and 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. and Vigdor, J.L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26, no. 6: 673–82.

Clotfelter, C.T., Ladd, H.F. and Vigdor, J.L. (2010). Teacher credentials and student achievement in high school: a cross-subject analysis with student fixed effects. *Journal of Human Resources*, 45(3), 655–682.

Coleman, J., Campbell, E.Q., Hobson, C.F. McPartland, J. and Mood, A.M. (1966): *Equality of Educational Opportunity*. Washington: U.S. Office of Education.

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

Cordero, J. M., Crespo, E. and Pedraja, F. (2013). Rendimiento educativo y determinantes según PISA: una revisión de la literatura en España. *Revista de Educación*, 362, 273-297.

Cordero, J.M., Cristóbal, V. and Santín, D. (2017). Causal Inference on Education Policies: A Survey of Empirical Studies Using PISA. TIMSS and PIRLS. *Journal of Economic Surveys (in press)*. doi: 10.1111/joes.12217.

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. and Kyriakides, L. (2008). *The dynamics of educational effectiveness*, London, Rouledge.

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

Croninger, R. G., Rice, J. K., Rathbun, A. and Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review*, 26(3), 312–324.

De Witte, K. and López-Torres, L. (2015). Efficiency in education: a review of literature and a way forward. *Journal of the Operational Research Society,68(4), 339-363.*

De Witte, K. and Van Klaveren, C. (2014). How are teachers teaching? A nonparametric approach. *Education Economics*, 22(1), 3-23.

Dee, T. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, 42 (3) 528–554.

Dehejia, W. and Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.

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

Dronkers, J. and 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.

Echazarra, A., Salinas. D., Méndez, I., Denis, V. and Rech, G. (2016). How teachers teach and students learn: Successful strategies for school. *OECD Education Working Papers*, No. 130. Paris: OECD Publishing. http://dx.doi.org/10.1787/5jm29kpt0xxx-en.

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

Ehrenberg, R. G. and Brewer, D. J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. *Economics of Education Review*, 13, 1–17.

Falck, O. and 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. and 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.

Fendick, F. (1990). *Correlation between teacher clarity of communication and student achievement gain: a meta-analysis*, University of Florida, George A. Smathers Libraries.

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

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

Gao, S. (2014). Relationship between science teaching practices and students' achievement in Singapore, Chinese Taipei, and the US: An analysis using TIMSS 2011 data. *Frontiers of Education in China*, 14(4), 519–551.

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

Gee, K. and 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. and Vermeersch, C. M. (2016). *Impact evaluation in practice*. *2nd edition*, World Bank Publications.

Gil-Izquierdo, M. and Cordero, J. M. (2018). Guidelines for data fusion with international large scale assessments: Insights from the TALIS-PISA link, *Studies in Educational Evaluation*, 59, 10-18.

Gil, M. and Cordero, J.M. (2018b). The effect of teaching strategies on student achievement: An analysis using TALIS-PISA-link, Journal of Policy Modeling, forthcoming, https://doi.org/10.1016/j.jpolmod.2018.04.003

Gil, M., Cordero, J. M. and Cristóbal, V. (2018). Las estrategias docentes y los resultados en PISA 2015, *Revista de Educación*, 379, 32-55.

Goe, L., Bell, C. and Little, O. (2008). *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. National Comprehensive Center for Teacher Quality.

Goldhaber, D. and Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics*, 89(1), 134-150.

Green, A. D. and 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.

Guo, S. and Fraser, M.W. (2010): *Propensity Score Analysis. Statistical Methods and Applications*. SAGE publications. London.

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, 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. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24. 1141–1177.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, 19, 141–164.

Hanushek, E. A. (2011). The economic value of higher teacher quality, *Economics of Education Review*, *30*(3), 466-479.

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, 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.

Hattie, J. A. C. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Oxon: Routledge.

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

Hidalgo, A. and López-Mayan, C. (2015). Teaching styles and achievement: Student and teacher perspectives. *Economic Analysis Working Paper Series 2/2015*. Universidad Autónoma de Madrid.

Hiebert, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. Washington: DIANE Publishing.

Hogrebe, N. and 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.

House, J. D. (2009). Elementary-school mathematics instruction and achievement of fourth-grade students in Japan: Findings from the TIMSS 2007 assessment. *Education*, 130(2), 301-308.

Hox, J. (2002). *Multilevel Analysis. Techniques and Applications.* Psychology Press.

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

INEE (2011). *Evaluación General de Diagnostico 2010. Informe de Resultados*. Ministerio de Educación. Madrid (España).

Isphording, I. E., Piopiunik, M. and 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, 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. and 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. and 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*, 58(2), 139-148.

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

Jürges, H., Schneider, K. and 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. and 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.

Koenker, R. and Bassett Jr. G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1), 33-50.

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

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

Kuzmina, J. and 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. and 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. and 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. (2011). *What makes an effective teacher? Quasi-experimental evidence*. NBER Working Paper 16885.

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.

Le Donné, N., Fraser. P, and Bousquet, G. (2016). Teaching Strategies for Instructional Quality: Insights from the TALIS-PISA Link Data. *OECD Education Working Papers*, No. 148, Paris: OECD Publishing. http://dx.doi.org/10.1787/5jln1hlsr0lr-en

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. and 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. and 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. and 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. and 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.

Marina, J.A., Pellicer, R. and Manso, J. (eds.) (2015). *Libro Blanco de la profesión docente y su entorno escolar*. Madrid: Ministerio de Educación. Cultura y Deporte.

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

Martínez-Arias, R. (2006). La metodología de los estudios PISA. *Revista de Educación*, Número Extraordinario, 111-129.

Méndez, I. (2015). *Prácticas Docentes y Rendimiento Estudiantil: Evidencia a partir de PISA 2012 y TALIS 2013*. Madrid: Instituto Nacional de Evaluación Educativa y Fundación Santillana.

Meroni, E. C., Vera-Toscano, E. and Costa, P. (2015). Can low skill teachers make good students? Empirical evidence from PIAAC and PISA. *Journal of Policy Modeling*, 37(2), 308-323.

Mislevy, R. J., Beaton, A. E., Kaplan, B. and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133-161.

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

Mullis, I. V., Martin, M. O., Foy, P. and Arora, A. (2012). *TIMSS 2011 international results in mathematics*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Murnane, R. J. and Phillips, B. R. (1981). What do effective teachers of inner-city children have in common? *Social Science Research*, 10(1), 83-100.

Lavrijsen, J. and 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.

OCDE (2009). PISA Data Analysis Manual. SPSS Second Edition. Paris: *OECD Publishing*.

OCDE (2013). Advancing National Strategies for Financial Education. Paris: OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*, PISA, OECD Publishing, Paris.

OCDE (2016). PISA 2015 Technical Report. PISA. Paris: *OECD Publishing*.

OECD (2017). *How Your School Compares Internationally: OECD Test For Schools*. OECD Publishing, Paris.

O'Dwyer, L. M., Wang, Y. and Shields, K. A. (2015). Teaching for conceptual understanding: A cross-national comparison of the relationship between teachers' instructional practices and student achievement in mathematics. *Large-scale Assessments in Education*, 3(1), 1-30. doi: 10.1186/s40536-014-0011-6.

Opdenakker, M.C. and Van Damme, J. (2006). Teacher characteristics and teaching styles as effectiveness enhancing factors of classroom practice. *Teaching and Teacher Education*, 22, 1–21.

Palardy, G. J. and Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, 30(2), 111-140.

Papanastasiou, C. (2008). A residual analysis of effective schools and effective teaching in mathematics, *Studies in Educational Evaluation*, 34(1), 24-30.

Pedraja-Chaparro, F., Santín, D. and 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. and Santín, 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. and 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. and 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. and 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.

Rasch, G. (1980). (1960/80). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Danish Institute for Educational Research.

Razak, N. A. and Shafaei, A. (2016). The Variation in Teaching and Learning Practices and their Contribution to Mathematics Performance in PISA 2012, in Thien, L.M., Razak, N.A., Keeves, J.P. and Darmawan (eds.). *What Can PISA 2012 Data Tell Us? Performance and Challenges in Five Participating Southeast Asian Countries*, Rotterdam: Sense Publishers, 123-157.

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

Rivkin, S. G. and 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.

Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247–52.

Rodríguez-Coma, M. (2012) Técnicas de evaluación de impacto: propensity score matching y aplicaciones prácticas con STATA. *Documentos de Trabajo del Instituto de Estudios Fiscales* 2/2012.

Rosen, M. and 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), 1-19.

Rosenbaum, P. R. and 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.

Rosenshine, B. and Stevens, R. (1986). Teaching functions, in Wittrock, M.C. (Ed.). *Handbook of research on teaching*, MacMillan, New York, 376-391.

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

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement, *Quarterly Journal of Economics*, 125(1), 175-214.

Rowan, B., Correnti, R. and Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teacher College Record*, 104, 1525–1567.

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. and 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. and 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. and 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.

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

Schacter, J. and Thum, Y. M. (2004). Paying for highand low-quality teaching. *Economics of Education Review*, 23, 411–430.

Schlotter, M., Schwerdt, G. and 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. and Winter-Ebmer, R. (2008). Peer effects in Austrian schools, *Empirical Economics*, 32, 387-409.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H. and 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 Working Paper nº 68.

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

Seidel, T. and Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454-499.

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

Strietholt, R., Gustafsson, J. E., Rosen, M., and 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.

Stuart, E. A. and Rubin, D. B. (2008). Matching methods for causal inference: Designing observational studies, in Osborne, J. (Ed.). *Best practices in quantitative methods*, Sage, London, 155–176.

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

Van de Grift, W. J. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*, 25(3), 295-311.

Van Klaveren, C. (2011). Lecturing style teaching and student performance, *Economics of Education Review*, 30(4), 729-739.

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

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

Von Davier, M. and Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models, in Rutkowski, L., Von Davier, M., Rutkowski, D. (eds.). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 155-174, CRS Press, London.

Von Secker, C. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *The Journal of Educational Research*, 95(3), 151-160.

Von Secker, C. E. and Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching*, 36(10), 1110-1126.

Wayne, A. J. and Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1), 89-122.

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

Wentzel, K. R. (2002). Are effective teachers like good parents? Teaching styles and student adjustment in early adolescence. *Child Development*, 73, 287–301.

West, M.R. and 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. and 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.

Wilde, E. T. and 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.

Willms, J. D. and Smith, T. (2005). *A manual for conducting analyses with data from TIMSS and PISA.* Report prepared for UNESCO Institute for Statistics.

Windschitl, M. and Sahl, K. (2002). Tracing teachers' use of technology in a laptop computer school: The interplay of teacher beliefs, social dynamics, and institutional culture. *American Educational Research Journal*, 39(1), 165-205.

Woessmann, L. (2003). School resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2), 117-170.

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

Woessmann, L. and 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.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2-3), 114-128.

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

Xue, Y. and Meisels, S. (2004). Early literacy instruction and learning in kindergarten: Evidence from early childhood longitudinal study-class of 1998-1999. *American Education Research Journal*, 41(2), 191-229.

Zabalza, M.A. (2011). Metodología docente. (Teaching Methodology), *Revista de Docencia Universitaria*. 9 (3), 75-98.

Zakharov, A., Carnoy, M. and Loyalka, P. (2014). Which teaching practices improve student performance on high-stakes exams? Evidence from Russia. *International Journal of Educational Development*, 36, 13-21.

Zemelman, S., Daniels, H. and Hyde, A. (2005). *Best practice: Today's standards for teaching and learning in America's schools*. Portsmouth: *Heinemann*.

Zuzovsky, R. (2013). What works where? The relationship between instructional variables and schools' mean scores in mathematics and science in low-, medium-, and high-achieving countries. *Large-scale Assessments in Education*, 1(1), 2.