Studies in Educational Evaluation 83 (2024) 101386

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

New evidence on the impact of learning in a foreign language on educational outcomes

Luis Pires [a,*], María-Jesús Mancebón [b], Mauro Mediavilla [c], José-María Gómez-Sancho [b]

[a] University Rey Juan Carlos, Spain
[b] University of Zaragoza, Spain
[c] University of Valencia and EVALPUB, Spain

ARTICLE INFO

Keywords:
Student evaluation
Bilingual education
CLIL
PISA 2015 and 2018
Spanish region of Madrid

ABSTRACT

A Content and Language Integrated Learning (CLIL) method, in which some subjects are taught in a foreign language (English), was initiated in Spain in 2005 and has progressively extended to half of public schools. The results have been very positive; however, it has been argued that studying subjects in a foreign language may reduce the educational outcomes of students. This paper evaluates this criticism in the Madrid bilingual program adopting a Propensity Score Matching approach, with the PISA 2015 and 2018 data. The model defines a homogenous student subsample from bilingual and nonbilingual schools in terms of the observable characteristics that may jointly influence both the selection of school type and educational scores. Our results, robust to the sample and unobservable hidden bias, indicate that the Madrid bilingual program, in addition to improving students' English level, does not reduce the skills of subjects taught in English or in Spanish.

1. Introduction

One of the major changes in the Spanish education system over the past two decades involved expanding bilingual education programs to public schools, which were previously only available in private fee-paying schools (Pires & Gallego, 2022). This move aimed to address historical shortcomings in Spaniards' foreign language skills by adopting the Content and Language Integrated Learning (CLIL) model. CLIL involves teaching subjects, or parts of subjects, through a foreign language with the dual goals of learning content appropriate for the learners' age and acquiring a foreign language (Coyle et al., 2010; Morton & Llinares, 2017). The adoption of CLIL programs has experienced a significant increase in publicly funded Spanish schools over the past two decades.

Despite widespread support from families, evidenced by a significant increase in enrollment in bilingual public schools, some sectors criticize this educational policy. One concern is the potential for cream skimming in the student population, both through school selection mechanisms and students' self-selection (Mediavilla et al., 2023). Additionally, opponents argue that bilingual programs may hinder student learning in both subjects taught in a foreign language and those taught in Spanish (Zafra, 2023).

In this context, our study aims to address the second criticism by evaluating the impact of a bilingual educational program (Spanish/English) launched in Madrid in 2004 on students' educational outcomes. Two key questions arise in this context. First, do students enrolled in bilingual programs show lower educational competencies compared to those in monolingual programs for subjects taught in English? Second, does the program have a negative impact on learning in the subjects taught in Spanish?

Answering these questions is important because if bilingual programs are as successful as intended, students educated in bilingual paths will benefit from academic improvements which, together with the returns produced by bilingualism itself, will have positive long-term effects on their careers. However, if such programs are not correctly designed and implemented, students may suffer a double disadvantage: losses in their native language skills and losses in their overall academic performance, due to the double burden of learning specific curricular content in a foreign language (Patrinos & Velez, 2009).

More specifically, our study scrutinizes the impact of the Madrid Bilingual Educational Program (hereinafter referred to as MBP) on students' learning outcomes. This assessment concentrates on a subject instructed in English (science) and two subjects (reading and mathematics) taught in Spanish, the mother tongue of the majority of students.

* Corresponding author.
E-mail address: luis.pires@urjc.es (L. Pires).

https://doi.org/10.1016/j.stueduc.2024.101386
Received 18 October 2023; Received in revised form 10 June 2024; Accepted 6 July 2024
Available online 22 July 2024

To conduct our research, we leverage microdata published in the 2015 and 2018 waves of the Program for International Student Assessment (PISA). In this sense, our approach is based on registry data and not on an experiment controlled by the researchers.

The evaluation performs a Propensity Score Matching (PSM) analysis to define a homogenous student subsample, in terms of the observable characteristics that may jointly influence both the selection of school type and educational scores. This technique addresses the potential endogeneity issue affecting our estimates, thereby reducing the likelihood of biased impact estimation. To evaluate the robustness of the PSM estimations, our study incorporates a double sensitivity analysis: Coarsened Exact Matching and Rosenbaum procedure for bounding the estimates of the treatment effect.

The current study has four main contributions. Firstly, and in distinction to other research, it uses a statistical matching technique to define a homogenous student subsample from bilingual and non-bilingual schools, in terms of the observable characteristics that may jointly influence both the selection of school type and educational scores.[1] Secondly, a robustness analysis is included. Thirdly, our research employs the PISA 2015 and 2018 assessments to evaluate the progress of the program between the two periods. Finally, we evaluate not only the impact of the bilingual program on the subjects taught in English but also on those imparted in Spanish, to test whether the greater effort required by studying subjects in a foreign language negatively impacts on the results obtained for the remaining subjects.

The structure of the paper is as follows. Section 2 reviews the studies conducted in the field of Education Economics on educational bilingual programs, paying special attention to those contributions which have evaluated the program in Madrid. The MBP is described in Section 3. An outline of the methodological approach is presented in Section 4. Section 5 supplies the data and describes the variables. Section 6 offers the empirical results. Finally, Section 7 summarizes the main conclusions and offers suggestions for further research.

2. Literature review

Traditionally, bilingual educational programs have been implemented in countries that have historically received significant volumes of immigrant populations with a language different to the official language of the host country. The case of the United States with the Hispanic community is the most representative. Additionally, there are territories that are multilingual for historical reasons, such as former European colonies or states formed after the dissolution of the USSR. Finally, educational bilingualism programs have been implemented in countries where various indigenous but minority ethnic and linguistic groups coexist, as is the case in some countries in Latin America and Africa.

The objectives pursued by the programs implemented in these countries are diverse. For example, bilingual education in the United States has primarily been a program whose goal is to teach English rather than to develop bilingualism. Hence, most USA bilingual programs are designed for students who come to school speaking native or home languages (mainly Spanish). These programs (labeled transitional bilingual education) were developed as a way of responding to various local, state, and federal mandates that required schools in the USA to provide equal access to educational opportunities for students who enter US schools with limited proficiency in English. The underlying rationale for these programs is to utilize students' native languages to teach content so that these students do not fall behind in their learning of content while they are learning English (Gándara & Escamilla, 2017).

Other bilingual programs in the USA have been developed with the aim of maintaining the native language of students whose mother tongue is not English. Their goal is to develop both languages equally (additive bilingualism), so as not to lose the home language one, but rather, use it to support English. These programs are advocated by those who believe that education should help preserve individual cultural identities because they contend that cultural diversity enriches society and that each culture contributes something valuable (Kim et al., 2015).[2]

In other countries, implemented programs have consisted of providing education in the mother tongue of the minority (L1) instead of in the official language of the country (L2). This is the case of several Latin American governments who have implemented bilingual or Indigenous language education programs, specifically targeted the Indigenous population. In these cases, the goal has been to reduce the persistent achievement gap between Indigenous and non-Indigenous children and contribute to reduce social in- equities (Hynsjö & Damon, 2016).

The studies on bilingualism have been conceived with the aim to determine whether it is better to educate the minority population in their mother tongue during the early years of schooling and gradually introduce the official language of the country in later years (the bilingual option) or whether a total immersion in the majority language from an early age is more beneficial in terms of educational outcomes (the monolingual option). The introduction of bilingual programs stems from the observation that minority groups usually have low rates of educational achievement (Bradley et al., 2007; Glewwe et al., 2017 among others). The objective is to attempt to facilitate the learning of minority linguistic groups by combining the benefits of learning in the child's mother tongue and the acquisition of the language skills necessary to develop adequate proficiency in the official and majority language of the host country (Ivlevs & King, 2014).

This strand of literature have shown that immigrant children who are educated in their mother tongue perform better academically than their counterparts who attend schools where the majority and official language of the country is used as the language of instruction. This has been demonstrated, for the United States, among others, by Adesope et al. (2010), Slavin et al. (2011), and Chin et al. (2013); for Europe by Reljić et al. (2015), and Ivlevs and King (2014); and for developing countries, by Patrinos and Velez (2009) in Guatemala, Hynsjö and Damon (2016) in Peru, Seid (2016) and Ramachandran (2017) for Ethiopia, Eriksson (2014), Taylor and von Fintel (2016) for South Africa, and Mohapatra (2016) for India.[3]

In recent decades, most of the bilingual programs implemented in European countries have been based on the CLIL method, where various curricular subjects are taught in a foreign language, mainly English, instead of in children's native language. Studies of this new phenomenon can be divided between those analyzing issues related to equity, understood as the possibility of covert segregation (Bruton, 2011; Van Mensel et al., 2020; Mediavilla et al., 2023), and those that focus on analyzing whether the educational outcomes obtained by those who follow a CLIL program differ from those achieved by students who receive a monolingual education. Among the latter, some studies evaluate how the CLIL method improves language proficiency in the foreign language: Admiraal et al. (2006), Lasagabaster (2008), and Diez Nieto De Diezmas (2016). Other studies have analyzed whether the CLIL method hinders competencies in the subjects taught in a foreign language, finding that students following the CLIL method obtained worse results (Marsh et al., 2000).

The MBP, one of the most advanced educational bilingual programs in Spain, has been the subject of numerous evaluations of its impact on

[1] Only one previous study of the Madrid bilingual program (Sotoca & Muñoz, 2015) has used a matching technique, but that analysis was only conducted in the small Eastern school district in Madrid.

[2] Kim et al. (2015) present an extensive overview of the existing modalities of educational bilingualism in the United States.

[3] For more details on the effects of bilingual education, refer to the recent meta-analysis by Gunnerud et al. (2020).

Table 1
Descriptive statistics.

PISA 2015										
Variables	Obs.	Mean			t-test		Std. Dev.	D values	Min.	Max.
		Total	No biling.	Biling.	t	p > t				
Bilingual school	864	0.41	0	1			0.49	2,04	0	1
Reading	864	510.20	502.16	521.91	-3.12	0.002	80.42	0,25	253.04	692.74
Mathematics	864	492.63	484.66	504.22	-3.61	0.000	75.42	0,26	297.03	681.93
Science	864	507.37	499.53	518.78	-3.09	0.002	83.65	0,23	266.90	715.23
Immigrant	864	0.24	0.27	0.19	2.87	0.004	0.43	-0,20	0	1
Education parents (low)	864	0.19	0.22	0.15	2.13	0.033	0.39	-0,17	0	1
Education parents (medium)	864	0.24	0.24	0.23	-0.24	0.811	0.43	-0,02	0	1
Education parents (high)	864	0.57	0.54	0.61	-1.45	0.148	0.50	0,15	0	1
Occupation parents	864	48.53	45.55	52.87	-4.58	0.000	22.38	0,33	12.00	89.00
Books at home (low)	864	0.24	0.29	0.16	3.82	0.000	0.42	-0,29	0	1
Books at home (medium)	864	0.56	0.53	0.59	-1.29	0.197	0.50	0,13	0	1
Books at home (high)	864	0.21	0.18	0.24	-2.24	0.026	0.40	0,15	0	1
Female	864	0.48	0.49	0.47	0.29	0.769	0.50	-0,03	0	1
Repetition	786	0.42	0.51	0.29	4.33	0.000	0.66	-0,34	0	3
Foreign language at home	863	0.08	0.08	0.08	0.46	0.000	0.27	-0,01	0	1
Parental support	859	3.48	3.45	3.51	-1.14	0.000	0.65	0,09	1	4
Student motivation	843	62.17	62.10	62.27	-0.01	0.988	18.97	0,01	0	100
Absenteeism	859	1.33	1.36	1.27	2.08	0.000	0.62	-0,14	0	4
PISA 2018										
Variables	Obs.	Mean			t-test		Std. Dev.	D values	Min.	Max.
		Total	No biling.	Biling.	t	p > t				
Bilingual school	1852	0.39	0.00	1.00			0.49	2,05	0	1
Reading	1852	463.31	449.24	484.95	-8.13	0.000	87.00	0,41	208.63	720.94
Mathematics	1852	475.59	463.01	494.95	-7.97	0.000	79.44	0,40	213.76	711.06
Science	1852	478.58	466.83	496.63	-7.17	0.000	81.15	0,37	193.75	767.86
Immigrant	1852	0.25	0.29	0.17	5.89	0.000	0.43	-0,29	0	1
Education parents (low)	1852	0.17	0.20	0.13	3.64	0.000	0.38	-0,17	0	1
Education parents (medium)	1852	0.18	0.20	0.15	3.00	0.002	0.38	-0,14	0	1
Education parents (high)	1852	0.65	0.60	0.72	-5.25	0.000	0.48	0,25	0	1
Occupation parents	1852	47.79	44.26	53.23	-8.44	0.000	22.29	0,40	11.74	88.96
Books at home (low)	1852	0.28	0.33	0.20	6.07	0.000	0.45	-0,30	0	1
Books at home (medium)	1852	0.50	0.49	0.52	-0.84	0.399	0.50	0,06	0	1
Books at home (high)	1852	0.22	0.18	0.28	-5.17	0.000	0.41	0,25	0	1
Female	1852	0.49	0.48	0.51	-1.30	0.194	0.50	0,06	0	1
Repetition	1667	0.39	0.46	0.29	4.55	0.000	0.65	-0,26	0	3
Foreign language at home	1851	0.08	0.10	0.05	3.71	0.000	0.27	-0,17	0	1
Parental support	1682	3.35	3.31	3.41	-2.71	0.000	0.79	0,12	1	4
Student motivation	1686	64.38	63.97	65.04	-1.36	0.174	13.93	0,08	0	100
Absenteeism	1679	1.37	1.38	1.34	1.03	0.304	0.66	-0,06	0	4

Note: P values "0.000″ are positive values < 0.001

student performance in the subjects taught in English. These analyses have used regional and international evaluations (PISA, TIMMS, PIRLS). Most such studies use an adaptation of the difference-in-difference technique, they compare the first cohort of students in the first year of introduction of the MBP with students from the same school, but from the previous year, who have not received CLIL tuition, and with the control group consisting of those schools which do not participate in the MBP (Quecedo, 2015; Anghel et al., 2016; Montalbán, 2016; Pires & Gallego, 2022). These studies conclude that, compared to non-bilingual students, the first cohort of students in schools implementing the MBP, worsened their results in the subject taught in English (science) when finishing primary education (6th grade). However, this deterioration is more notable in schools which were the first to instigate the MBP. In other words, there is an improvement over time in its application in schools which have recently implemented the program, as they can learn from the experience of forerunning schools. Pires and Gallego (2022) also examined students completing compulsory secondary education (10th grade) and found no negative effect on the subjects taught in English.

Other studies have used different methods, such as the Nonequivalent Control Group technique, to match students from bilingual schools with students of similar characteristics from non-bilingual schools (Sotoca & Muñoz, 2015). Alternatively, Mixed Effects Models combine the inference of the principal effects with estimates of the characteristics of secondary sources, such as the school or the municipality (Tamariz and Blasi, 2016), while Multinomial Logit Models measure the variables influencing the probability of obtaining improved results (García-Centeno et al., 2020). The results of the latter studies are similar to the former: although understanding of the subjects taught in English is slightly poorer in primary education, in secondary education there are no longer significant differences between students from bilingual and non-bilingual schools. The slight deterioration in the acquisition of knowledge in science in primary education is compensated for later in compulsory secondary education (12–16-year-old students).

3. The Madrid bilingual program

The bilingual educational program of the Community of Madrid (MBP) was initiated in the 2004–2005 academic year in public schools in

the first year of primary education (first grade with 6- or 7-year-old students). The program has been gradually extended to the remaining years, one academic year per year. The first 26 bilingual public primary schools completed their bilingual education in primary education in the 2009–2010 academic year (when their first bilingual students reached the sixth grade). Bilingualism was initiated in secondary education schools[4] in the 2010–2011 academic year, following the same progressive implementation during the four years of compulsory Secondary Education, from seventh to tenth grade.[5] In the 2015–2016 academic year the first students who had embarked upon the bilingual program twelve years earlier completed their twelfth grade. They were the first students to have undergone all their education (compulsory and pre-university non-compulsory) in a bilingual program.

In the 2021–2022 academic year, the MBP encompassed 734 public schools (403 primary schools, 194 secondary schools, 10 vocational training schools and 127 early childhood education schools), in addition to 223 grant-maintained schools. These represent 50.4 % of public primary schools, 63.6 % of public secondary schools and 59.7 % of grant-maintained schools. The number of students on the MBP is approximately 385,000, of which a quarter of a million are in public centers: 14,968 in early childhood education, 116,748 in primary education, 89,136 in compulsory secondary education, 27,351 in the Baccalaureate, and 757 in vocational training centers. The financing of bilingual teaching in the Community of Madrid has enjoyed a consolidated and increasing budget which, in the 2019–2020 academic year, amounted to almost 47.50 million euros (Comunidad de Madrid, 2022).

The regulation of bilingual public schools is determined by Order 5958/2010 in Primary Education and Order 972/2017 in compulsory Secondary Education. In line with this legislation, all bilingual public schools must teach completely in English subjects that amount to at least 30 % of the weekly teaching curriculum, including English subject. Although schools can choose which subjects to teach in English to reach this minimum of 30 %, the educational authorities recommend that science and social science be taught in English. Mathematics and Spanish subjects can only be taught in Spanish.

Those teachers wishing to teach MBP subjects must obtain an English language credential in English by passing linguistic tests at the C1 CEFRL level. The school principal is responsible for supervising the correct development of the MBP and bilingual schools possess considerable supplementary resources, such as specific learning material, digital whiteboards, certificates of linguistic competence in English with international recognition for students and participation in European programs.

The Regional Ministry of Education publishes an annual Order for the selection of new bilingual public schools, which establishes the requirements that schools are obliged to meet to optimally develop the MBP. The decision to apply for participation in the bilingual program emanates from the school itself, but the final choice is made by the Regional Ministry of Education.

4. Methodological approach

When evaluating the impact of an educational program on students' educational outcomes, it is important to take into consideration certain empirical features that challenge observational studies addressing this question. In our specific case, the main methodological challenge to overcome stems from the fact that the distribution of students between public bilingual and non-bilingual schools in the region of Madrid (as in the rest of Spain) is not random. This is because schools are freely chosen by families. Among other influences, family socio-economic characteristics have been proven to be one of the main determinants of the school selection pattern in Spain, bilingual schools being chosen mainly by families of higher socio-economic status than families selecting public non-bilingual schools (Van Mensel et al., 2020; Mediavilla et al., 2023). This situation leads to a potential problem of endogeneity concerning the "Bilingual School" predictor, i.e., to potential correlations between this predictor and the residuals of the regressions, creating OLS-biased estimates.[6]

Overcoming this problem is the basis for our empirical strategy, consisting of performing a Propensity Score Matching (PSM). The PSM method tries to mimic the randomized assignment to treatment and comparison groups by choosing for the comparison group those units that have similar propensities to the units in the treatment group. Since propensity score matching is not a randomized assignment method but tries to imitate one, it belongs to the category of quasi-experimental methods (Gertler et al., 2016).

The PSM technique allows us to define a homogenous student subsample in terms of the observable characteristics that may jointly influence both the selection of school type and educational scores. In this way, we reduce the endogeneity problem affecting the predictor of interest (school type) and obtain an unbiased estimate of the average effect of attending a bilingual public school. One of the benefits of matching is that it produces lower variance in the estimates and is more robust to departures from assumptions than model-based methods used on random samples (Rosenbaum & Rubin, 1983; Rubin & Thomas, 2006).

In addition, our study controls for the impact of unobservable variables on results. To do this we offer a double robustness check. First at all, we replicated these estimations using a complementary approach, Coarse Exact Matching. Subsequently, we apply Rosenbaum (2002) procedure to establish boundaries on the estimates of treatment effects.

The purpose of PSM is to proxy a credible value of the counterfactual for each of the individuals belonging to the treatment group (Rosenbaum & Rubin, 1985). In our case, this consists of selecting a group of students from bilingual public schools (the school treatment group or TG) which is comparable to a group of students attending a public non-bilingual school (control group or CG) in all those covariates (X) which can potentially condition both school choice and the scores obtained in the PISA evaluation. The principal advantage of the PSM resides in its capacity to perform matchings between treated and non-treated individuals when the number of covariates is high (Rosenbaum & Rubin, 1983). This is because matchings are performed upon a single magnitude, the propensity score, which synthesizes all the information contained in the X control variables. The propensity score was defined by Rosenbaum and Rubin (1983) as the conditional probability of assignment to treatment, given covariates,[7] i.e.:

$$e(x) = P\,(Z = 1|X) \tag{1}$$

where e(X) is the propensity score, Z is the indicator of participation in treatment (treatment group Z = 1 and control group Z = 0) and X are the observable characteristics of individuals that affect both participation in treatment and the outcomes evaluated. The propensity score is a balancing score: conditional on the propensity score, the distribution of

[4] In the Spanish education system, public schools are separated into primary schools (from first to sixth grade) and secondary schools (from seventh to tenth grade) in compulsory education, to then continue with non-compulsory studies the Baccalaureate or Vocational Training.

[5] Grant-maintained private schools initiated the MBP in primary education in the 2008–2009 academic year and in secondary education in 2015–2016. The present article focuses on the highly demanding bilingual program in primary and secondary public schools.

[6] The key point is that household socioeconomic characteristics are also the chief determinants of educational outcomes. This underlies the problem of self-selection bias which threatens our estimates. Selection bias or endogeneity is a widespread methodological challenge in educational research (Murnane and Willet, 2011).

[7] The assumption of selection on observables requires that conditional on the observed variables, the assignment to treatment is random.

measured baseline covariates is similar between treated and untreated individuals. The propensity score exists in both randomized experiments and observational studies. In the former, the true propensity score is known and is defined by the design of the study. In observational studies, the true propensity score is unknown. However, it can be estimated using the study data (Austin, 2011). Econometric literature offers various methods to estimate the conditional probability of receiving a treatment (Guo & Fraser, 2010). In practice, the propensity score value is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics. This is the method we use to calculate the propensity score which indicates the probability of attending a bilingual school.

Having obtained two comparable samples of students in bilingual centers (the treated group) and in non-bilingual centers (the control group), the second step involves the application of different matching algorithms to the two groups (treated and control). The third step is to compare the outcomes of the treated and untreated individuals belonging to the matched subsample (average treatment effect or ATE[8]).

Once the ATE had been estimated we performed two robustness checks, as stated above. This is very important when impact is evaluated by PSM, since the observed positive relationship between a student's treatment status and test score outcomes may not necessarily indicate a causal effect (Bradley et al., 2013).

First at all, we apply the Coarse Exact Matching (CEM). The key idea behind CEM is to "coarsen" or group the values of the covariates into a smaller number of categories, reducing the dimensionality of the matching problem. In so doing, it becomes easier to identify similar units across treatment and control groups (Iacus et al., 2012).

The second robustness check is the Rosenbaum sensitivity analysis. With this we try to overcome the primary challenge associated with cross-sectional matching analysis, namely the potential presence of hidden bias resulting from selection effects on unobserved differences. The Rosenbaum sensitivity analysis permits the assessment of the robustness of our findings, which are faced with potential imbalances in unobservable factors (Altonji et al., 2008; Peel, 2014, among others). The Rosenbaum (2002) procedure functions by establishing boundaries to the estimates of the treatment effect. Specifically, it reports the p-values obtained from Wilcoxon sign-rank tests for the average treatment effect (ATE), while assuming a certain value γ to quantify hidden bias. This reflects our assumptions regarding unmeasured differences or endogeneity in treatment assignment (expressed as the odds ratio of differential treatment assignment, due to an unobserved covariate). For each γ value, we calculate a hypothetical significance level we term the "critical p-value," which represents the upper limit for the significance level of the treatment effect in cases of endogenous self-selection into treatment status.

The Propensity Score Matching technique and the robustness checks employed in this study allow for an improved analysis of the causality between enrolment in a bilingual or monolingual school and students' academic outcomes, although never fully demonstrating such causality. For the application of all the above-mentioned methodological issues, we use STATA software (version 15.0) and applicate commands "pscore" and "psmatch2".

[8] The most common estimator in non-experimental studies is the "average effect of the treatment on the treated" (ATT), which is the effect for those in the treatment group, and the "average treatment effect" (ATE), which is the effect on all individuals (treatment and control). Our focus of interest is to measure the expected effect on the outcome if individuals in the population were randomly assigned to treatment, this being exactly what is captured by the ATE (Austin, 2011). This parameter allows us to ascertain what the performance of Spanish students would be if they attended a Public bilingual school.

5. Data and variables

Our study uses a database based on the microdata published in the 2015 and 2018 editions of the Program for International Student Assessment (PISA). Concretely, we employ the microdata corresponding to the representative sample of the Spanish region of Madrid. PISA evaluates the ability of 15-year-old students to apply the knowledge and skills taught and learnt in the classroom to concrete situations and practical contexts. The comprehensive information supplied by PISA covers different aspects of the educational process and includes not only information on the scores obtained by students in diverse standardized tests, but also their personal contexts and their family and academic backgrounds. To provide valid estimates of student achievement and characteristics, PISA selects a sample of students that represents the full population of 15-year-old students in each participating country or education system. Some regions, including Madrid in the PISA editions of 2015 and 2018, conduct an expanded sample that allows for comparing their results with those of other participating regions and countries. Our study only includes the network of public centers, in order to create a more homogenous database in terms of institutional aspects such as staff recruitment, salaries, and the autonomy of school principals, among others. The PISA 2018 database identifies the schools participating in the bilingual program, while the PISA 2015 database does not. Consequently, to the PISA 2015 database we add the administrative information supplied by the Regional Ministry of Madrid for the year of initiation of each school in the MBP, in order to obtain data on which students belong to a bilingual school.

The final database of the study comprises, for PISA 2015, 864 students from 26 secondary public schools, of which 10 schools were participants in the MBP (with 358 students) and 16 were not (506 students). In turn, PISA 2018 includes 1852 students from 61 secondary public schools, of which 30 schools were participants in the MBP (with 960 students) and 31 were not (892 students).

Concerning the variables used in our analysis, the main predictor in our estimations is a dichotomous variable (bilingual school) that indicates whether students attend (1) or not (0) an MBP center.

Student educational outcomes are approximated by the three main competences evaluated in PISA: reading, mathematics and science. In PISA, these competencies have a mean of 500 and a standard deviation of 100. As explained above, in schools belonging to the MBP reading and mathematics are taught in Spanish while science is taught in English.

To account for individual heterogeneity that might affect the relationship between our main predictor (attending a bilingual school) and educational outcomes, we include a comprehensive set of control variables employed in the existing literature. Specifically, we base ourselves on previous studies that have analyzed the determinants of attendance at a bilingual school (Mediavilla et al., 2023) and we consider all the variables that have shown a statistically significant correlation with our dependent variables. From among these variables, we only consider for the matching analysis those that influence attendance at a bilingual school and the academic results of students; in addition, the variables included in the matching process must be stable in the period from entering a bilingual school to taking the assessment tests several years later. Specifically, these variables include the personal and family characteristics of students (whether they are immigrants, and the educational, occupational and cultural level of their parents). Other variables that can affect the PISA scores or which do change during the time of schooling will be used in a post- matching analysis; these are variables related to the students' academic background, that we proxy by repetition of an academic year, the academic support of parents, the level of educational motivation, or absenteeism, among others.

Table 1 supplies the descriptive statistics for all the variables employed and incorporates information on the differences between bilingual and non-bilingual schools. As Table 1 shows, these differences are considerable and statistically significant. Students attending bilingual schools achieve better outcomes in all the PISA competences.

Table 2
Determinants of attendance at a bilingual state school (logit model estimation).

VARIABLES	PISA 2015	Odds ratio	PISA 2018	Odds ratio
Immigrant	-0.199	0.808	-0.366*	0.722
	(0.123)		(0.207)	
Education parents (medium)	0.113	1.150	-0.053	0.960
	(0.213)		(0.147)	
Education parents (high)	0.023	0.934	0.157	1.162
	(0.236)		(0.167)	
Occupation parents	0.012**	1.012	0.013***	1.012
	(0.006)		(0.003)	
Books at home (medium)	0.503***	1.559	0.337*	1.344
	(0.178)		(0.175)	
Books at home (high)	0.514	1.666	0.522**	1.640
	(0.362)		(0.239)	
Constant	-1.352***	0.280	-1.348***	0.442
	(0.661)		(0.309)	
Log likelihood	17567.87***		-16392.26***	
	(1835.094)		(388.220)	
Correctly classified (%)	59.95		58.80	
Area Under ROC Curve	0.61		0.63	
Pseudo R2	0.027		0.038	
Observations	864		1852	

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

However, students at such schools are also from more select academic and socio-economic strata than those attending non-bilingual schools. Consequently, bilingual schools have a lower percentage of immigrant students and of students having repeated one or more academic years. Furthermore, bilingual schools attract a greater proportion of families with a high level of education and professional skills and more motivated students. Consequently, the comparison of raw educational results does not permit a correct evaluation of the impact of having attended a bilingual public school. To address this problem, we perform a Propensity Matching Score analysis.

6. Results

This section presents the principal results obtained from our empirical analysis. Firstly, the estimations obtained from the application of the PSM are offered. Next, we present the principal contributions to these estimations offered by the application of a post-matching analysis.

6.1. PSM results

As explained above, the purpose of PSM is to proxy a credible value of the counterfactual for each of the individuals belonging to the treatment group (Rosenbaum & Rubin, 1983). In our case, this consists of encountering a group of students from bilingual public schools (the treatment group) which is comparable with students attending non-bilingual public schools (the control group) in all those covariables which can potentially condition both school choice and PISA scores.

The first step in the PSM is to estimate the propensity score. To do this we construct a binomial logit model which allows us to calculate the probability of a student attending a bilingual school, conditional on its observable variables. This model is an auxiliary step aimed at generating a homogeneous subsample of students in bilingual and non-bilingual schools concerning the observable variables that affect both school choice and student performance. For this reason, only variables that simultaneously influence the participation decision and the outcome variable should be included (Caliendo & Kopeinig, 2008). The results are displayed in Table 2, showing that the model fits reasonably well as around 60 % of the observations are correctly classified and the area under ROC curve is higher than 0.60. Subsequently, we examine the distribution of the estimated propensity score by the treatment and the control groups, with a univariate Kernel density estimation. Table 2 shows that the strongest influence on the probability of attending a public bilingual school in Madrid are immigrant status (native children are more likely to attend a public bilingual school), parental occupation and the number of books at home. In other words, the children most likely to attend a bilingual public school are those from higher socio-economic and cultural backgrounds.

Fig. 1 displays the distribution of the predictions of the estimated PS for individuals from bilingual and non-bilingual public schools. As shown, there is a large common support zone (0.2 - 0.7, approximately), a fundamental requirement of the PSM technique.

Having estimated the propensity score, the matching process was undertaken. The literature suggests several algorithms for the performance of this process (Guo & Fraser, 2010). In the present study we have chosen to apply the Epanechnikov kernel type KM with a bandwidth of 0.06, since this was the algorithm that best matched the individuals from both the treatment group and the control group.[9]

Employing this method, the PS matched 596 observations (298 from each group) in PISA 2015, and 1406 observations (703 from each group) in PISA 2018. The control individuals were weighted on the basis of the number of times they were paired with treated individuals. These weightings are essential for the subsequent statistical analyses.

Other results devoted to testing the quality of matching[10] are given in Table 3 and Fig. 2. Table 3 shows the differences in the average values of the propensity scores and the covariates for the whole sample and the matched sample. As can be inferred, the matching reduced covariate imbalance for all variables. Following the matching process, the

[9] We also performed the analysis with two other matching algorithms: with replacement (dropping treatment observations whose p-score is higher than the maximum or less than the minimum p-score of the controls), without replacement (1-to-1 matching without replacement) and with the nearest neighbour matching. We also used different bandwidths in the Kernel estimations, but in no case do they affect the number of matched individuals in the sample. Consequently, we decided to use the bandwidth that STATA applies by default, 0.06. The results are available on request.

[10] Since we do not condition on all covariates but on the propensity score, it has to be checked if the matching procedure is able to balance the distribution of the relevant variables in both the control and treatment group (Caliendo & Kopeinig, 2008).

covariables do not display statistically significant differences between the treated and control groups. Hence, after the matching process, students are balanced in both groups (bilingual and non-bilingual schools). In addition, Fig. 2 depicts the distribution of the propensity score by type of school for the matched sample. As shown, a perfect overlap is achieved between bilingual and non-bilingual schools following the matching process.

All the previous analyses allow us to be confident about the high quality of our matching, as they show the fulfilment of one of the key assumptions in PSM applications: overlap and common support (Caliendo & Kopeinig, 2008).

Having selected the subsample of comparable individuals, the following step in the PSM was to calculate the matching estimator of the average treatment effect (ATE); the results are displayed in Table 4 for PISA 2015 and 2018 and for the outcomes in science, mathematics and reading.

To verify the effect of the bilingual program, we first analyzed the results in science, as students receive this subject in English in bilingual schools (treatment group), while in non-bilingual schools they receive it in Spanish (control group). The results are different between PISA 2015 and PISA 2018. In the former, the differences between the two groups are very slight (6.7 points) and not statistically significant. Thus, in 2015, students' outcomes from bilingual schools do not differ significantly from those for students from non-bilingual schools, despite receiving their classes in English and the fact that the PISA test is conducted in Spanish. On the other hand, in PISA 2018 the differences in science scores are positive (13.5) and statistically significant, indicating that bilingual school students obtain better results than those from non-bilingual schools. Although our model does not allow for fully demonstrating causality, the results indicate that learning a subject in English does not harm the acquisition of skills.

To complete the study of the effect of attending a bilingual school, we next evaluated the impact on the subjects taught in Spanish, i.e., reading and mathematics. This permitted us to verify whether studying a subject in English (i.e., science) can have negative repercussions on the educational competences of subjects taught in Spanish. The underlying hypothesis in this analysis is that the greater effort required for a student to study a subject in a foreign language could negatively influence the results obtained in the remaining subjects (a negative external effect). If this is true, the overall assessment of the success of the program would be called into question. The results of this analysis, once again, differ between 2015 and 2018. In 2015, with the exception of mathematics where bilingual schools achieve better results, there are no differences in the two competencies (reading and mathematics) between bilingual and non-bilingual schools, in the same way that there were no differences in the subject taught in English (science). On the other hand, in 2018 there is a difference in favor of students from bilingual schools (13.7 in mathematics and 19.9 in reading), although this difference is similar to that for science, taught in English. Therefore, the joint results show that in bilingual schools the teaching of a subject in English does not worsen the results in the subjects taught in Spanish (i.e., there is no negative external effect).

To better interpret the differences between bilingual and non-bilingual schools in PISA test scores, we need to understand how PISA translates its results into comparable terms. All PISA results are scaled to fit approximately normal distributions, with means around 500 score points and standard deviations around 100 score points. Elsewhere, according to the (OECD, 2019), the scores obtained can be translated into "years of formal education" and can be compared if the circumstances in which this education is provided are the same for all students. In 2018, the OECD established that, on average, a difference of 40 points represented the distance between contiguous grades of education and can therefore be translated into one year of formal education. Taking into account the 10-month duration of a school year in Spain, a student at a bilingual school experiments an increase of approximately 3 months (13.5 points) and 5 months (19.9 points) of formal education, compared to a student at a monolingual school.

6.2. Sensitivity analysis and robustness check

To assess the validity of our estimations, that is, to discover whether the use of the PSM has identified two groups of students sufficiently similar to allow reliable observations of the differences between them,

Table 3
Covariables after PSM.

PISA 2015	Mean			t-test	
	Treated	Control	%bias	t	p > t
Propensity Score (PS)	0.435	0.434	1.8	0.25	0.801
Immigrant	0.182	0.181	0.1	0.01	0.993
Education parents (low)	0.145	0.145	0	-0.01	0.995
Education parents (medium)	0.240	0.235	1.2	0.16	0.870
Education parents (high)	0.615	0.620	-1	-0.14	0.889
Occupation parents	53.542	53.267	1.2	0.17	0.868
Books at home (low)	0.156	0.160	-1	-0.14	0.887
Books at home (medium)	0.584	0.586	-0.4	-0.05	0.957
Books at home (high)	0.260	0.254	1.4	0.18	0.858
PISA 2018	Mean			t-test	
	Treated	Control	%bias	t	p > t
Propensity score (PS)	0.543	0.541	2.3	0.52	0.606
Immigrant	0.165	0.169	-1.1	-0.26	0.797
Education parents (low)	0.127	0.130	-0.9	-0.22	0.824
Education parents (medium)	0.148	0.150	-0.5	-0.11	0.916
Education parents (high)	0.725	0.720	1.1	0.25	0.803
Occupation parents	53.566	53.149	1.9	0.41	0.681
Books at home (low)	0.193	0.201	-1.9	-0.44	0.658
Books at home (medium)	0.522	0.521	0.1	0.03	0.976
Books at home (high)	0.285	0.278	1.7	0.36	0.721

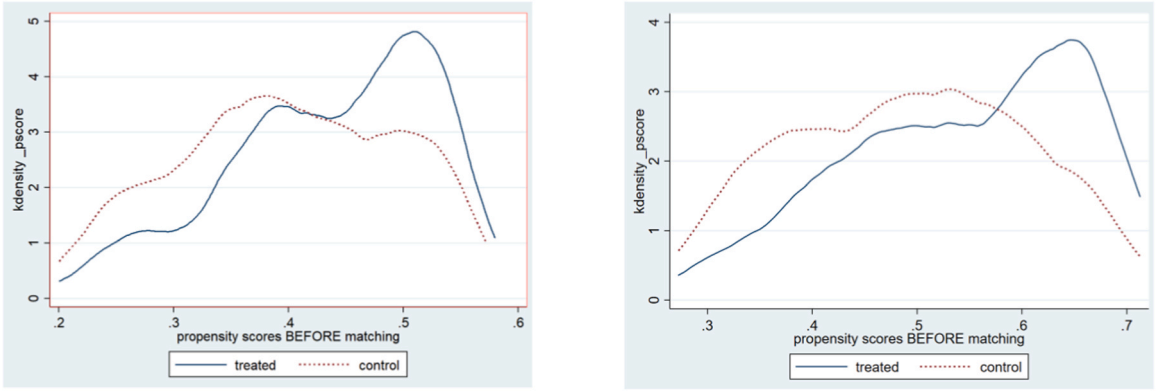

Fig. 1. Kernel density estimation before PSM. PISA 2015 (left) and PISA 2018 (right)

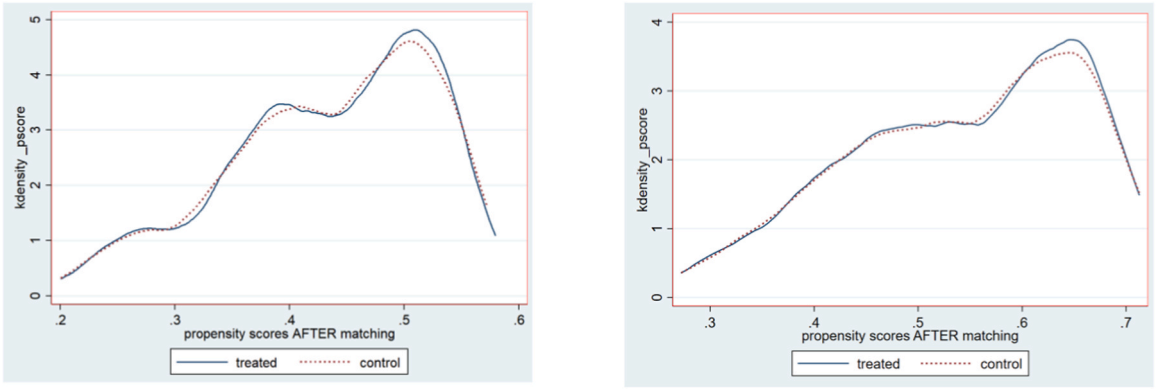

Fig. 2. Kernel density estimation after PSM. PISA 2015 (left) and PISA 2018 (right)

Table 4
ATE in science, mathematics, and reading competences between bilingual (treated) and non-bilingual (control) public schools.

ATE	PISA 2015			PISA 2018		
	Difference	S.E.	D	Difference	S.E.	D
Science	6.7	4.7	1.4	13.5 ***	4.0	3.4
Mathematics	8.3 *	4.8	1.7	13.7 ***	3.7	4.0
Reading	6.2	4.9	1.3	19.9 ***	4.3	4.7

ATE: Average treatment effect; Difference: between treated (bilingual) and non-treated (non-bilingual); S.E.: Standard Error
* $p < 0.1$
*** $p < 0.01$

we performed a sensitivity analysis, repeating the estimations by means of Coarsened Exact Matching (CEM). This methodology simulates a random assignment of the treatment group (bilingual students) over a larger population than the initial matching, in order to reduce the imbalance that the selected covariates may have (Guarcello et al., 2017). Table 5 offers the comparison of the results between the estimation of PSM (Table 4) and the estimation using CEM. The results show that the values of the PSM are similar to the results obtained by CEM. In conclusion, the reliability of our estimations is assured.

In addition to the above, the principal issue with cross-sectional propensity matching analysis is that there may be a problem of hidden bias, due to the effect of selection on unobserved heterogeneity. Consequently, we perform the Rosenbaum (2002) procedure for bounding the estimates of the treatment effect. Table 6 shows that robustness to hidden bias varies across the PISA 2015 and PISA 2018 estimations. In PISA 2015, where we found no effect (neither positive nor negative) on academic results for attendance at a bilingual school, the critical level is exceeded at gamma = 1. On the other hand, in PISA 2018 we verified that there may be a positive effect on academic results due to attendance at a bilingual school. The critical gamma level at which we might question this conclusion is 1.33. This implies that only a considerable unobserved heterogeneity (33 %) would alter the inference about the estimated effects. That is to say, our results would not be robust if there existed an unobservable variable which caused a 33 % variation in academic performance. Since our calculations incorporate several covariables for the estimation of the ATE, it is very unlikely that there exists an unobservable variable possessing an influence of 33 % in itself. Moreover, it is important to note that the Rosenbaum bounds are a "worst case" scenario.

Table 5
Results from PSM and CEM.

		PISA 2015		PISA 2018	
		Difference	t	Difference	t
Science	PSM	6.7	0.9	13.5 ***	3.4
	CEM	5.5	0.84	14.3 **	3.4
Mathematics	PSM	8.3 *	1.4	13.7 ***	3.6
	CEM	7.0	1.17	16.6 ***	4.1
Reading	PSM	6.2	0.8	19.9 ***	4.7
	CEM	6.1	0.95	20.6 ***	4.5

* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Table 6
Rosenbaum bounds.

	Gamma	P value critical
PISA 2015	1	0.056
(N = 358 matched pairs)	1.03	0.089
	1.06	0.133
PISA 2018	1	0.000
(N = 960 matched pairs)	1.03	0.000
	1.06	0.000
	1.09	0.000
	1.12	0.000
	1.15	0.000
	1.18	0.001
	1.21	0.002
	1.24	0.005
	1.27	0.012
	1.3	0.026
	1.33	**0.051**
	1.36	0.090
	1.39	0.146
	1.42	0.219

On the basis of the results of the CEM and the Rosenbaum (2002) procedure for bounding the treatment effect estimates we can conclude that our PSM results are robust.

6.3. *Post-matching analysis*

The application of PSM has permitted us to obtain an unbiased estimation of ATE with regard to the observable variables (X) which affect both the selection of a bilingual school and educational outcomes. However, the potential influences on educational outcomes usually include more variables than those considered in the construction of the propensity score. Thus, the calculation of a more precise effect of attendance at a public bilingual school requires considering the influence of those other factors which, although potentially important in the determination of the PISA scores, do not influence school choice (the attributes of students not incorporated into the calculation of the propensity score, i.e., those coinciding with education in a bilingual school.). To do this, it is essential to perform a post-matching analysis.

The testing of the influence of the characteristics not included in the estimation of the ATE can be undertaken via a mixed-effects model for

Table 7
Post-matching regression.

	PISA 2015			PISA 2018		
	Science	Math	Reading	Science	Math	Reading
Bilingual school	-2.260	-0.924	-3.022	17.42***	18.78***	24.59***
	(6.537)	(5.105)	(6.384)	(5.107)	(4.508)	(8.473)
Immigrant	-10.570	-13.03*	-5.456	4.097	-5.397	10.84*
	(8.295)	(7.227)	(7.123)	(5.980)	(6.042)	(6.583)
Education parents (medium)	-2.006	-3.074	-3.699	2.927	5.779	0.770
	(5.890)	(6.753)	(5.347)	(7.534)	(6.407)	(7.546)
Education parents (high)	-5.792	-1.147	-8.112	-8.952	-6.628	-15.61***
	(8.549)	(9.209)	(7.373)	(5.898)	(5.842)	(5.970)
Occupation parents	0.088	0.096	0.147	0.343***	0.357***	0.281**
	(0.118)	(0.117)	(0.134)	(0.122)	(0.138)	(0.128)
Books at home (medium)	29.09***	20.55***	31.53***	21.02***	12.82***	16.60***
	(7.509)	(5.336)	(6.792)	(5.095)	(4.752)	(5.238)
Books at home (high)	44.35***	35.18***	41.37***	41.65***	31.13***	38.19***
	(11.170)	(7.969)	(9.660)	(5.849)	(5.637)	(7.745)
Female	-19.51***	-24.52***	8.387**	-27.04***	-23.10***	6.650
	(4.096)	(2.477)	(3.946)	(3.867)	(4.375)	(4.672)
Repetition	-57.07***	-55.61***	-57.18***	-49.14***	-60.26***	-49.47***
	(4.253)	(4.375)	(3.826)	(3.789)	(3.479)	(2.956)
Foreign language at home	-9.894	0.944	-8.665	-10.440	-7.319	-8.549
	(10.230)	(7.806)	(8.825)	(7.939)	(8.230)	(9.900)
Support parents	-5.990	-6.463	-5.156	-0.816	-2.523	0.068
	(5.659)	(4.541)	(5.483)	(2.760)	(2.834)	(2.941)
Student motivation	0.758***	0.723***	0.604***	0.183	0.377	0.401
	(0.154)	(0.140)	(0.131)	(0.194)	(0.231)	(0.286)
School absence	-2.897	-4.771	-6.857	-9.621***	-11.33***	-11.22**
	(4.765)	(3.408)	(4.211)	(3.093)	(3.743)	(4.702)
Constant	502.7***	499.0***	501.4***	484.7***	482.8***	444.7***
	(27.380)	(23.320)	(25.240)	(16.480)	(15.030)	(20.770)
ICC	5.3 %	4.1 %	6.5 %	13.7 %	14.9 %	19.9 %
Number of groups	26	26	26	61	61	61
Observations	762	762	762	1401	1401	1401
Null model: Schools Variance	363.3	230.0	409.5	970.9	1033.5	1612.8
Null model: Students Variance	6515.7	5411.1	5836.8	6101.5	5902.7	6490.3
Null model: Total Variance	6879.0	5641.2	6246.3	7072.4	6936.2	8103.1
Complete model: Schools Variance	140.6	72.7	163.1	196.0	173.1	803.8
Complete model: Students Variance	3561.1	2680.1	3041.6	3811.3	3334.1	4534.4
Complete model: Total Variance	3701.6	2752.8	3204.7	4007.3	3507.1	5338.2
Total variance explained by variables	46.2 %	51.2 %	48.7 %	43.3 %	49.4 %	34.1 %
Level 1 (students) variance explained by variables	45.3 %	50.5 %	47.9 %	37.5 %	43.5 %	30.1 %
Level 2 (schools) variance explained by variables	61.3 %	68.4 %	60.2 %	79.8 %	83.3 %	50.2 %

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

the matched sample.[11] Table 7 offers the results of these estimations. The dependent variables in the regression are the PISA scores in science, mathematics and reading. The independent variables, calculated from the PISA students' questionnaire, are attendance at a public bilingual school and several variables that the prolific literature on the educational production function[12] has shown to be drivers of educational achievement: gender, repetition, foreign language at home, parental support, student motivation, and absenteeism.

As shown in Table 7, our model explains a high percentage of the variance of student's performance in PISA. While in PISA 2015 the effect of attending a bilingual school is not significant, in PISA 2018 the effect in the three competences is positive and statistically significant, 17.42 points in science, 18.78 in mathematics and 24.59 in reading, an increase of approximately 4 and 6 months of formal education compared to a student at a monolingual school.

Furthermore, if we consider that all PISA tests are performed in Spanish (including scientific competences), the results obtained for the subject taught in English in bilingual schools could be downwardly biased.

The remaining results in Table 7, although they do not allow for fully demonstrating causality, are predictable and similar to those obtained in the numerous studies of PISA results already performed. Thus, boys obtain better results in science and mathematics, while girls do so in reading, although this last effect is not statistically significant in PISA 2018. In turn, the family variables that most positively influence the

[11] The implementation of a mixed-effects model is necessary because PISA data have a hierarchical nature. This is due to the sampling structure in each PISA country, which is not purely random but involves intricate random draws from the sampling frame conducted in multiple stages. Specifically, PISA employs a stratified two-stage sample design, where schools are sampled using probability proportional to size (school enrollment of 15-year-olds) sampling, and students are sampled with equal probability within schools. Consequently, the data lack independence (students enrolled in the same school share the same values for the school variable), violating the OLS assumption of Independence of Errors. The regression was performed using the STATA "MEGLM" command, specifying sampling weights in the school level (weight provided by PISA with the variable W_SCHGRNRABWT), and in the observation level (student), including the weights provided by PISA (W_FSTUWT) and the weights obtained in the PSM. We are very grateful to one of the reviewers for providing insightful comments regarding the most suitable model for the post-matching analyses.

[12] See Hanushek (2020).

results are the parents' occupation and the number of books at home, variables that reflect the economic and cultural level of the family. Finally, repetition of one or more academic years negatively influences results, as does students' academic attitude, especially absenteeism and motivation.

7. Conclusions

In the last fifteen years, the Madrid region has implemented one of the most significant educational innovations ever undertaken in Spain, namely the introduction of bilingual education programs into publicly funded schools.

This article has attempted to answer two questions about this program. The first is whether participation in the MBP results in a loss of competence in science, the subject studied in English (given the greater difficulty that studying in a foreign language presents for students). The second issue under consideration is whether participation in the program negatively affects academic performance in subjects taught in Spanish (given that the increase in time required to study a subject taught in English may generate negative external effects on the remaining subjects). These questions concern many Spanish families faced with the choice of the most convenient school for their children.

Our results show that the greater effort required by studying subjects in a foreign language does not impacts on the results obtained by the students. In the MBP, in addition to improving students' English level, learning a subject in a foreign language (English) does not harm the acquisition of skills in those subjects or in the subjects taught in Spanish (mathematics and reading).

Choosing a bilingual public school, as against a non-bilingual center, could be a sensible decision for families. Although students in bilingual centers do not worsen their academic results on average, families must also consider the situation and personal characteristics of their children, seeing if they will be able to take advantage that bilingual centers offer, or if they will be able to withstand the pressure and the effort needed in that program.

Most studies conducted on the MBP, except for those analyzing the program's early years of implementation, reach the same conclusions as this study, that bilingualism does not reduce students' competencies despite teaching in English. The more positive effect in PISA 2018 compared to PISA 2015 could be because the students tested in PISA 2015 were 10th graders from the first wave of bilingual centers (which started the program in 2004 and 2005); as the program was developed and its efficiency improved (with greater knowledge on the part of the participants), students from later waves achieved better results in PISA 2018. Our analysis even shows that the content subjects taught in English are positively affected. There are several possible explanations for this circumstance. Firstly, the results of this study indicate that bilingual schools possess some characteristics favoring their students' learning. Since participation in the MBP is voluntary for schools, it is possible that the management team and teachers at these schools are highly motivated. The analysis of these schools' characteristics is a promising field of future study on the Madrid bilingual program. For example, the work of teachers can be analyzed to verify their different performance between bilingual and non-bilingual schools. Secondly, the CLIL system may present learning advantages, as the subjects are taught in a manner that enhances processing and remembering. For example, Surmont et al. (2016) show how CLIL appears to have a positive impact on the mathematical performance of pupils even after a short period of time because CLIL stimulates students in more than one aspect of the learning process. Similar conclusions are reached in Gunnerud et al. (2020). Therefore, students do not have to be highly proficient in the speaking languages and the mathematical language to benefit from an increased metalinguistic awareness.

Statements and Declarations

None.

Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

We have received funds, and I have already conveyed that information: 1) Conselleria d'Hisenda i Model Econòmic, Generalitat Valenciana, Grant GIUV2016-318; 2) Spanish Regional Government of Aragón, Grant S23_20R and S23_23R.

CRediT authorship contribution statement

Luis Pires: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Investigation, Formal analysis, Data curation, Conceptualization. **María-Jesús Mancebón:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Data curation. **Mauro Mediavilla:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **José -María Gómez-Sancho:** Visualization, Validation, Investigation.

Data availability

The data used in this work come from the PISA database (2015 and 2018 editions), published by the Organization for Economic Cooperation and Development (OECD) and freely accessible on its website (https://www.oecd.org/ steps/data/).

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245. https://doi.org/10.3102/0034654310368803

Admiraal, W., Westhoff, G., & De Bot, K. (2006). Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English 1. *Educational Research and Evaluation, 12*(1), 75–93. https://doi.org/10.1080/13803610500392160

Altonji, J., Elder, T., & Taber, C. (2008). Using selection on observed variables to assess bias from unobservables when evaluating Swan–Ganz catheterization. *American Economic Review, 98*(2), 345–350. https://doi.org/10.1257/aer.98.2.345

Anghel, B., Cabrales, A., & Carro, J. M. (2016). Evaluating a bilingual education program in Spain: The impact beyond foreign language learning. *Economic Inquiry, 54*(2), 1202–1223. https://doi.org/10.1111/ecin.12305

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Bradley, S., Migali, G., & Taylor, J. (2013). Funding, school specialization and test scores: An evaluation of the specialist schools policy using matching models. *Journal of Human Capital, 7*(1), 76–106. https://doi.org/10.1086/669203

Bradley, S., Draca, M., Green, C., & Leeves, G. (2007). The magnitude of educational disadvantage of indigenous minority groups in Australia. *Journal of Population Economics, 20*(3), 547–569. https://doi.org/10.1007/s00148-006-0076-9

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System, 39*(4), 523–532. https://doi.org/10.1016/j.system.2011.08.002

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Chin, A., Daysal, N. M., & Imberman, S. A. (2013). Impact of bilingual education programs on limited English proficient students and their peers: Regression discontinuity evidence from Texas. *Journal of Public Economics, 107*, 63–78. https://doi.org/10.1016/j.jpubeco.2013.08.008

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and Language Integrated Learning (CLIL*. Cambridge University Press. https://doi.org/10.1017/9781009024549

Comunidad de Madrid (2022). Datos y Cifras de la educación 2021–2022. Retrieved September 21, 2023, from http://edicion.comunidad.madrid/sites/default/files/doc/educacion/sgea_datosycifras_2021–22_1.pdf.

Diez Nieto De Diezmas, E. (2016). The impact of CLIL on the acquisition of L2 competences and skills in primary education. *International Journal of English Studies, 16*(2), 81–101. https://revistas.um.es/ijes/article/view/239611.

Eriksson, K. (2014). Does the language of instruction in primary school affect later labour market outcomes? Evidence from South Africa. *Economic History of Developing Regions, 29*(2), 311–335. https://doi.org/10.1080/20780389.2014.955272

Gándara, P., & Escamilla, K. (2017). Bilingual education in the United States. *Bilingual and Multilingual Education, 12*(1), 439–452. https://doi.org/10.1007/978-3-319-02258-1_33

García-Centeno, M. C., de Pablos Escobar, L., Rueda-López, N., & Calderón Patier, C. (2020). The impact of the introduction of bilingual learning on sixth grade educational achievement levels. *PloS One, 15*(6), Article e0234699. https://doi.org/10.1371/journal.pone.0234699

Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact Evaluation in Practice, Second Edition*. Washington, DC: Inter-American Development Bank and World Bank. https://doi.org/10.1596/978-1-4648-0779-4

Glewwe, P., Kritikova, S., & Rolleston, C. (2017). Do schools reinforce or reduce learning gaps between advantaged and disadvantaged students? Evidence from Vietnam and Peru. *Physiology and Behavior, 176*(5), 139–148. https://doi.org/10.1016/j.physbeh.2017.03.040

Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing student success: Assessing supplemental instruction through coarsened exact matching. *Technology, Knowledge and Learning, 22*(3), 335–352.

Gunnerud, H. L., Ten Braak, D., Reikerås, E. K. L., Donolato, E., & Melby-Lervåg, M. (2020). Is bilingualism related to a cognitive advantage in children? A systematic review and meta-analysis. *Psychological Bulletin, 146*(12), 1059–1083. https://doi.org/10.1037/bul0000301

Guo, S., & Fraser, M. W. (2010). Propensity score analysis: Statistical methods and applications. *Psychometrika, 75*, 775–777. https://doi.org/10.1007/s11336-010-9170-8

Hanushek, E. (2020). Education Production Functions. In Steve Bradley & Colin Green (ed.), *Economics of Education* (2nd ed., pp. 161–170). Academic Press.

Hynsjö, D., & Damon, A. (2016). Bilingual education in Peru: Evidence on how Quechua-medium education affects indigenous children's academic achievement. *Economics of Education Review, 53*, 116–132. https://doi.org/10.1016/j.econedurev.2016.05.006

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Ivlevs, A., & King, R. M. (2014). 2004 minority education reform and pupil performance in Latvia. *Economics of Education Review, 38*, 151–166. https://doi.org/10.1016/j.econedurev.2013.08.010

Kim, Y. K., Hutchison, L. A., & Winsler, A. (2015). Bilingual education in the United States: An historical overview and examination of two-way immersion. *Educational Review, 67*(2), 236–252. https://doi.org/10.1080/00131911.2013.865593

Lasagabaster, D. (2008). Foreign language competence in content and language integrated courses. *The Open Applied Linguistics, 1*, 30–41. https://benthamopen.com/ABSTRACT/TOALJ-1-30.

Marsh, H. W., Hau, K. T., & Kong, C. K. (2000). Late immersion and language of instruction in Hong Konghigh schools: Achievement growth in language and non-language subjects. *Harvard Educational Review, 70*, 302–346.

Mediavilla, M., Mancebón, M. J., Pires, L., & Gómez-Sancho, J. M. (2023). Bilingual school choice and socio-economic segregation: An analysis for Spain based on PISA 2015. *Research Papers in Education*. doi: 10.1080/02671522.2023.2188247.

Mohapatra, D. (2016). Effect of Bilingual Education on Academic Achievement: Evidence from India. SSRN. https://ssrn.com/abstract=2840830.

Montalbán, J. (2016). *Improving students' reading habits and solving their early performance cost exposure: evidence from a bilingual high school program in the Region of Madrid.* Working Paper of the Paris School of Economics, pp. 1–33. https://www.parisschoolofeconomics.eu/docs/montalban-castilla-jose/evaluacion-de-institutos-bilingues-en-madrid(1).pdf.

Morton, T., & Llinares, A. (2017). Content and Language Integrated Learning: Type of Program or Pedagogical Model? In A. Llinares, & T. Morton (Eds.), *Applied Linguistics Perspectives on CLIL* (pp. 1–16). John Benjamins. https://doi.org/10.1075/lllt.47.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter*. New York: Oxford University Press.

OECD. (2019). *PISA 2018 Results. What Students Know and Can Do* (Volume I.). https://doi.org/10.1787/5f07c754-en

Patrinos, H. A., & Velez, E. (2009). Costs and benefits of bilingual education in Guatemala: A partial analysis. *International Journal of Educational Development, 29*(6), 594–598. https://doi.org/10.1016/j.ijedudev.2009.02.001

Peel, M. (2014). Addressing unobserved endogeneity bias in accounting studies: Control and sensitivity methods by variable type. *Accounting and Business Research, 44*(5), 545–571.

Pires, L., & Gallego, M. J. (2022). The bilingual program in Madrid and its effects on learning. *Revista Dèlelőtt Educación, 397*, 351–380. https://doi.org/10.4438/1988-592X-RE-2022-397-550

Quecedo, C. H. (2015). The impact of bilingual education on average school performance: an evaluation of Madrid's bilingual schools program. *The Public Sphere: Journal of Public Policy, 3*(2), 158–162. https://psj.lse.ac.uk/articles/38.

Ramachandran, R. (2017). Language use in education and human capital formation: Evidence from the Ethiopian educational reform. *World Development, 98*, 195–213. https://doi.org/10.1016/j.worlddev.2017.04.029

Reljić, G., Ferring, D., & Martin, R. (2015). A meta-analysis on the effectiveness of bilingual programs in Europe. *Review of Educational Research, 85*(1), 92–128. https://doi.org/10.3102/0034654314548514

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55. https://www.jstor.org/stable/2335942.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rosenbaum, P. (2002). *Observational studies*. Springer.

Rubin, D. B., & Thomas, N. (2006). Combining propensity score matching with additional adjustments for prognostic covariates. Matched Sampling for Causal Effects. *Journal of the American Statistical Association, 95*(450), 573–585. https://doi.org/10.1017/CBO9780511810725.025

Seid, Y. (2016). Does learning in mother tongue matter? Evidence from a natural experiment in Ethiopia. *Economics of Education Review, 55*, 21–38. https://doi.org/10.1016/j.econedurev.2016.08.006

Slavin, R. E., Madden, N., Calderón, M., Chamberlain, A., & Hennessy, M. (2011). Reading and language outcomes of a multiyear randomized evaluation of transitional bilingual education. *Educational Evaluation and Policy Analysis, 33*(1), 47–58. https://doi.org/10.3102/0162373711398127

Sotoca, E., & Muñoz, A. (2015). The impact of bilingual education on academic achievement of students enrolled in public schools in the autonomous community of Madrid. *Journal of Education Research, 9*(1), 27–40. https://doi.org/10.5209/rev_RCED.2014.v25.n2.41732

Surmont, J., Struys, E., Van Den Noort, M., & Van De Craen, P. (2016). The effects of CLIL on mathematical content learning: A longitudinal study. *Studies in Second Language, Learning and Teaching, 6*(2), 319–337. https://doi.org/10.14746/ssllt.2016.6.2.7

Taylor, S., & von Fintel, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. *Economics of Education Review, 50*, 75–89. https://doi.org/10.1016/j.econedurev.2016.01.003

Van Mensel, L., Hiligsmann, P., Mettewie, L., & Galand, B. (2020). CLIL, an elitist language learning approach? A background analysis of English and Dutch CLIL pupils in French-speaking Belgium. *Language, Culture and Curriculum, 33*(1), 1–14.

Zafra, I. (2023). School bilingualism fails in its objective of generalizing the learning of English. *El País. October 3*. https://elpais.com/educacion/2023-10-03/el-bilinguismo-escolar-fracasa-en-su-objetivo-de-democratizar-el-aprendizaje-de-ingles.html#?prm=copy_link.