



Does the educational management model matter? New evidence from a quasiexperimental approach

María Jesús Mancebón¹  · Domingo P. Ximénez-de-Embún²  ·
Mauro Mediavilla³  · José María Gómez-Sancho¹ 

Received: 6 November 2015 / Accepted: 15 August 2017
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract A growing literature has appeared in the last 2 decades with the aim to explore if the way in which publicly funded private schools are managed (a very autonomous mode) is more effective, than that applied in public schools (where decisions are highly centralized), concerning the promotion of student's educational skills. Our paper contributes to this literature providing new evidence from the Spanish experience. To this end, we use the Spanish Assessment named “*Evaluación de Diagnóstico*,” a national yearly standardized test given to students in the fourth grade and administered by the Regional Educational Authorities. In particular, our data are those corresponding to the assessment conducted in the Spanish region of Aragón in 2010. Our methodological strategy is defined by the sequential application of two methods: propensity score matching and hierarchical linear models. Additionally, the sensitivity

The authors are grateful for the financial support received from the Spanish Government, Ministry of Economics and Competitiveness (Project EDU2013-42480-R). Mauro Mediavilla and Domingo P. Ximénez-de-Embún also acknowledge the support from Fundación Ramón Areces. We thank the editor, two anonymous referees and the associate editor for their helpful comments.

✉ María Jesús Mancebón
mjmanca@unizar.es

Domingo P. Ximénez-de-Embún
dpxe@unizar.es

Mauro Mediavilla
mauro.mediavilla@uv.es

José María Gómez-Sancho
jmgomez@unizar.es

¹ Department of Applied Economics, University of Zaragoza, Zaragoza, Spain

² Department of Economic Analysis, University of Zaragoza, Zaragoza, Spain

³ Department of Applied Economics, University of Valencia, Valencia, Spain

of our estimates is also tested with respect to unobserved heterogeneity. Our results underline the existence of a slight advantage of the private management model of schools in the promotion of scientific abilities of students and in the acquisition of foreign language (English) skills.

Keywords School choice · Propensity score matching · Hierarchical linear models · Unobservable variables bias · Science and Foreign Language (English) skills · Primary schools

JEL Classification I21 · I29

1 Introduction

Studies devoted to evaluating the impact of educational interventions have expanded notably worldwide over the last 2 decades. There are two factors explaining this: on the one hand, the availability of new, high quality, national and international data; on the other, the development of innovative and sophisticated econometric methods capable of confronting the principal methodological problems facing these studies. These factors have created new opportunities for academics worldwide to conduct research that addresses policymakers' concerns about the consequences of actions directed at improving educational outcomes (Murnane and Willett 2011).

One of the interventions that has focused the empirical work of many educational economists has been the public funding of private schools. This policy has been widely applied under different guises worldwide (vouchers, charter schools, state-funded private schools, free schools, academies, educational agreements between local authorities and private schools, etc.).¹ Of all these alternatives, the Spanish policy in the compulsory education level (primary and secondary schools) consists of educational agreements (called *conciertos*) between the regional educational authorities and some private schools. Under them, authorities pay to these private schools all the school operating cost. In return, they are subject to the same Schools Admissions Code as all other public schools and are impeded from charging any fees to families. However, these private schools are self-governing.

The aim of this paper is to evaluate the effectiveness of the self-governing model applied by the Spanish publicly funded private schools (versus that applied by the state funded and run schools) in the development of the educational skills of children in the fourth grade.² Research into this topic is extremely important in Spain, where two

¹ Recent reviews of these educational policies can be found in the Handbook of the Economics of Education (Bettinger 2011; Epple et al. 2016; Urquiola 2016).

² For the sake of simplicity, hereinafter we will refer to the publicly funded privately run schools simply as private schools, except when necessary to differentiate from completely private independent schools. Namely, in this paper a private school is understood as a type of publicly funded but self-governed school. Similarly, we will refer to state funded and run schools as public schools. The main difference between them lies mainly in the way they are administered, being private schools much more autonomous concerning the process and personnel decisions (deciding on the purchase of supplies and on budget allocations within schools, hiring and rewarding teachers, choosing textbooks, instructional methods, and the like).

models of school management (public and private) coexist and compete for limited public resources.

Our paper fits into a wide literature that originates with Coleman et al. (1982) who found that private Catholic schools in the USA were more effective than public schools in the promotion of the scores obtained by pupils in standardized tests on basic cognitive skills (reading, writing and mathematics). These authors applied multiple regression analysis with a set of carefully selected variables, arguing that controlling for important preexisting differences between students attending public and private schools allowed them to overcome the selection bias that threatened their estimations.³ However, subsequent methodological advances have made clear that Coleman's empirical strategy does not permit to obtain the causal effect of attending to a private school.

On this basis, a considerable number of empirical studies, involved in the evaluation of the impact of private schools on student achievement, have emerged in recent decades using different empirical strategies (random assignment, instrumental variables, regression, matching techniques or differences in differences). Nevertheless, the results of this literature are so far inconclusive.

While a number of studies in various settings find that private schools outperform public ones (for instance, Kim 2011; Crespo and Santín 2014), other research has found that student performance is not different, in terms of statistical significance, between private and public schools (Chudgar and Quin 2012, among others). In addition, there exists other studies which have concluded that private schools perform worse than public ones (Mancebón et al. 2012, among others).

In some papers, the effect encountered varies depending on the specific evaluated outcomes (Zimmer et al. 2012; Davies 2013), academic level of students (Gronberg and Jansen 2001) or methodological approach (Davies 2013). To these studies must be added those which have shown that the effects of school type vary over time (Hanushek et al. 2007) and those which fail to find a consistently positive (or negative) effect of state-funded Catholic schools on overall area-wide educational performance (Allen and Vignoles 2015).

To summarize, technical studies comparing the performance of publicly and privately run schools are thus far inconclusive (Hanushek and Woessmann 2014), which have hitherto prevented the identification of the optimal model of educational management (public versus private).

In this context, our study makes a further contribution to this controversial issue. Its main aim is to provide new evidence on the topic from the Spanish experience.⁴ Our study's idiosyncrasy lies in the methodological way that it follows to deal with the empirical challenges overshadowing our estimates. Indeed, this is the first study in

³ This bias has its origin in the fact that attendance at a school, whether private or public, is not random but instead is conditioned by characteristics of the family background, which in turn are extremely important in the determination of educational outcomes (the family socioeconomic level, for example).

⁴ In this sense, our paper meets Davies' claim (2013, p. 880): "As debates over school choice become increasingly transnational, we need studies from a variety of settings to build a stockpile of international knowledge about school sectors and student achievement."

this literature, as far as we know, where the potential selection bias, in observable and unobservable variables, and the nested structure of the dataset are jointly addressed.

Specifically, our methodological strategy is defined by the sequential application of two methods: propensity score matching (PSM) and hierarchical linear models (HLM). The first of these will allow us to delimit a homogenous subsample of students attending private and public schools and to obtain an unbiased estimation of the effect of the private management model of schools. The application of a post-matching analysis (by means of HLM) to this subsample of students will allow us to obtain a more efficient estimation of the effect of the private management model of schools upon the student's academic skills

The remaining sections of the paper are organized as follows. The subsequent section provides a description of the data used and the institutional setting. Section 3 outlines the methodological strategy employed to identify the effect of private schools on educational achievement. Section 4 presents the results of our estimations. Section 5 concludes the paper with a summary of our findings.

2 Data and institutional background

One of the defining characteristics of the school system in Spain is its dual nature, consisting predominantly of public sector provision but with a substantial private sector. The largest segment of the latter is represented by schools publicly funded by regional education authorities but privately owned and run. This school model arose in 1985 through the *Right to Education Act*—LODE—where the families right to school choice was acknowledged for the first time [for a detailed description and historical evolution of the Spanish non-higher education system, see Bernal (2005)].

The distribution of students enrolled in primary education among different school types in Spain in 2010 was as follows: 67.4% of students attended a public school, 28.5% a state-funded and privately run school and 4.1% completely private independent schools (Spanish Ministry of Education 2013). This situation is practically identical for the region of Aragón, which constitutes the main core of our study, where these percentages are 67.5, 29.5 and 3.0%, respectively. A detailed explanation of the differences between Spanish public and private schools is offered by Green et al. (2014).

Our empirical study employs census data for students in the fourth grade in the Spanish region of Aragón in 2010. These data come from the *Evaluación de Diagnóstico* (ED), a national assessment of the educational skills of fourth grade pupils established by the Spanish Education Act (LOE) in 2006 (hereinafter we refer to our database as *2010 Aragón ED*). This regional perspective permits a more appropriate comparison between private and public schools in terms of the region's funding, governance arrangements and student populations. Data were provided by the Aragón Educational Authority, which is responsible for the implementation of the Assessment in its district.

The ED is involved in the evaluation of several educational competencies that rotate every 2 years: Spanish, Maths, Science, Digital Skills, Foreign Language, Social Interaction and Citizenship, Art, Learning by Oneself and Personal Autonomy. In 2010,

the second year of the application of the ED, the competencies evaluated were science and foreign language (English). In addition to the assessment of pupils' skills, the ED includes data on a wide range of students' characteristics. These include basic demographics (gender, age, immigration status, etc.), information on children's socioeconomic backgrounds (parents' occupational status, parents' years of schooling, household possessions, etc.), data on the academic profile of students (if the child has repeated any academic year, if he or she needs help in doing homework, the daily time devoted to study, etc.) and data on parents' involvement in education, on students' perceptions of themselves and the class environment and on students' satisfaction with the school.

Table 1 includes the descriptive statistics of all the variables drawn from the 2010 *Aragón ED*, grouped by school type (public and private schools). As it shows, raw results are higher for private school students (in both science and foreign language—English).⁵ But raw differences are insufficient to judge the relative quality of public versus private schools because they do not take into account the differences between students attending each type of school, which are very important. As shown in Table 1, students' academic potential in private schools is much greater than in public schools. Further, this pattern holds for variables such as parents' occupational status, parents' years of schooling, household possessions, the immigration status of pupils, parents' involvement in education, students' motivation and so on. These differences between public and private schools are almost all statistically significant, a result which mirrors the conclusions reached by other studies which have analysed this topic in Spain, using distinct databases (Mancebón et al. 2012; Doncel et al. 2012).⁶

3 Research strategy

When evaluating the impact of the self-governing management model of private schools 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, as shown in Table 1, the distribution of students between schools in the region of Aragón (as in the rest of Spain) is not random. This is due to the fact that schools are freely chosen by families (LODE 1985). Among other influences, family socioeconomic characteristics have been proven to be one of the main determinants of the selection pattern in Spain (Escardibul and Villarroya 2009; Mancebón and Ximénez-de-Embún 2014), being private schools chosen mainly by families belonging

⁵ The average score of each competence for the total number of schools is 500 and the standard deviation 100, given that as established by the General Report on Diagnostic Evaluation in Aragón 2010 "the evaluation of each competence in Aragón as a whole is established at the level of the average scores transformed into a reference value which has been fixed at 500, with a standard deviation of 100." Here, the approach of the Spanish Diagnostic Evaluation is similar to that of the evaluations of the PISA Project of the OECD. In Table 1, the average score differs from 500 due to the elimination from the sample of completely private independent schools and of those situated in municipalities in which there exists no choice between public and private schools.

⁶ Escardibul and Villarroya (2009) and Mancebón and Ximénez-de-Embún (2014) offer some suggestions to cope with these inequalities in the distribution of students between public and private schools.

Table 1 Descriptive statistics grouped by school type. *Source:* Authors' calculations, from 2010 Aragón ED (Aragón Regional Education Authority)

Description variable	Mean			Difference (Private–public)
	Total	Public	Private	
Outcomes				
Science: Achievement in Science	512.37	501.97	526.27	24.30***
Foreign language: Achievement in foreign language (English)	513.02	499.18	531.50	32.32***
Individual				
Gender (female = 1, male = 0)	0.49	0.49	0.48	– 0.01
Student has repeated one or more academic years (repeater = 1, non-repeater = 0)	0.08	0.09	0.06	– 0.03***
Involvement in education				
2-h study every day	0.16	0.15	0.17	0.02
More than 2-h study every day	0.48	0.48	0.48	0.00
Student needs help in homework	0.22	0.22	0.22	0.00
Parents do not check either diary or homework	0.21	0.23	0.19	– 0.04***
Parents check diary but not homework	0.10	0.07	0.13	0.06***
Parents check homework but not diary	0.16	0.20	0.12	– 0.08***
Parents check both diary and homework	0.53	0.50	0.57	0.07***
Private tutoring	0.09	0.08	0.09	0.01
Attitude: Student always finishes homework	0.93	0.92	0.94	0.02*
Aptitude Student answers homework correctly	0.85	0.84	0.87	0.03***
Household I: socioeconomic and cultural characteristics				
Mother white collar highly skilled	0.29	0.24	0.37	0.13***
Mother white collar low skilled	0.41	0.42	0.39	– 0.03*
Mother blue collar high skilled	0.04	0.04	0.03	– 0.01***
Mother blue collar low skilled	0.26	0.30	0.21	– 0.09***

Table 1 continued

Description variable	Mean			Difference (Private–public)
	Total	Public	Private	
Father white collar high skilled	0.39	0.31	0.49	0.18***
Father white collar low skilled	0.25	0.26	0.23	–0.03*
Father blue collar high skilled	0.30	0.35	0.23	–0.12***
Father blue collar low skilled	0.06	0.07	0.05	–0.02***
Education mother (years)	11.45	10.77	12.35	1.58***
Education father (years)	11.45	10.79	12.33	1.54***
Student born in Spain	0.87	0.84	0.91	0.07***
Student born in Africa	0.01	0.01	0.00	–0.01***
Student born in Asia	0.01	0.01	0.01	0.00
Student born in Europe	0.05	0.06	0.03	–0.03***
Student born in Latin America	0.05	0.06	0.04	–0.02***
Student born in an Arab country	0.01	0.02	0.01	–0.01***
More than 5 years living in Spain	0.94	0.93	0.95	0.02***
Number of TVs at home	2.11	2.08	2.15	0.07***
Number of computers at home	1.55	1.49	1.63	0.14***
Number of pay TVs at home	0.44	0.43	0.46	0.03
Number of video games at home	1.73	1.66	1.82	0.16***
Number of MP4s at home	1.01	0.93	1.11	0.18***
Household 2: educational Resources and their use				
More than 100 books at home	0.54	0.50	0.60	0.10***
Student uses books frequently	0.72	0.70	0.75	0.05***
Student has own room to study	0.95	0.94	0.96	0.02***
Internet at home	0.86	0.84	0.88	0.04***
<i>N</i> (students)	6724	3845	2879	
<i>N</i> (schools)	205	124	81	

T test equality of means: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

to better socioeconomic status than those that select public schools. The “school type” predictor is therefore an endogenous variable, which gives rise to correlations between this predictor and the residuals of the regressions, creating OLS-biased estimates.⁷

On the other hand, our database have a hierarchical structure, as the sample selection of individuals in *2010 Aragón ED* occurs at two levels (student and school); that is data are clustered. Because of this, some of the characteristics of students attending the same school are correlated, which violates the hypothesis of independence of the observations.

The two characteristics mentioned above are the basis for our empirical strategy. This consists of a two-step procedure.

Firstly, a propensity score matching (PSM) analysis is conducted. This 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 will manage to correct the endogeneity problem affecting the predictor of our interest (i.e., type of school) and to obtain an unbiased estimate of the average effect of attending a self-governing private school (average treatment effect or ATE⁸). In addition, our study also controls for the impact of unobservable variables on results. To do this, a sensitivity analysis is employed. One of the benefits of matching is that it produces lower variance in the estimates and remains more robust to departures from assumptions than model-based methods used on random samples (Rosenbaum and Rubin 1985; Rubin and Thomas 2000).

Secondly, a post-matching analysis is undertaken. With this second step, we intend to improve the performance of the ATE supplied by the PSM, by cleaning up the small residual covariance imbalance between the groups and/or improving precision (Stuart 2010). Certainly, by including in a regression all those variables that influence the outcomes, the residual variance is reduced, which in turn lowers the standard error of the estimates. Hence, the post-matching analysis provides us with an opportunity to increase precision, which is very important because greater precision brings increased statistical power (Murnane and Willett 2011).⁹

⁷ The key point is that these characteristics are also chief determinants of educational outcomes. This circumstance is the cause underlying the self-selection bias problem threatening our estimates. Selection bias and/or endogeneity are widespread in educational research and is the main methodological problem encountered when trying to evaluate the effect of private schools on the academic performance of children (Lefebvre et al. 2011). This is a methodological problem inherent in all impact evaluations in non-experimental studies (such as *2010 Aragón ED*).

⁸ The most common estimands in non-experimental studies are 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 being this what is exactly captured by the ATE (Austin 2011). This parameter allows us to know what the performance of the Spanish students would be if they attended a self-governing private school.

⁹ Such as Imbens (2004, p. 11) states when he refers to the combination of methods to estimate ATE “The motivation for these combinations is that although in principle any one of these methods can remove all of the bias associated with the covariates, combining two may lead to more robust inference. For example, matching leads to consistent estimators for average treatment effects under weak conditions, so matching and regression can combine some of the desirable variance properties of regression with the consistency of matching.”

For this post-matching analysis, a multilevel equation model (HLM) is estimated. This allows us to cope with the clustered structure of the data supplied by the 2010 *Aragón ED*. This model, applied to the subsample defined through PSM, permits differentiation between those influences affecting at different levels: student, class and school.

4 Empirical results

This section presents the principal results obtained from the empirical analysis performed. Firstly, the estimations obtained from the application of the PSM are commented upon. Next, we present the principal contributions to these estimations offered by the application of the HLM.

4.1 PSM results

4.1.1 Estimation of the propensity score model

The purpose of PSM is to proxy a credible value of the counterfactual for each of the individuals belonging to the treatment group (Rosenbaum and Rubin 1983). In our case, this consist of finding a group of students from public schools (control group or CG) which is comparable with the students who attend a private school (treatment group or TG) in all those variables X which can potentially condition both the choice of school and the scores obtained in the 2010 *Aragón ED*.

The principal advantage of the PSM resides in its capacity to perform matchings between the treated and non-treated individuals when the number of covariates (X) is numerous (Rosenbaum and 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) to be the conditional probability of assignment to treatment given covariates,¹⁰ i.e:

$$e(x) = P(Z = 1|X) \quad (1)$$

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.

The propensity score is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated individuals. The propensity score exists in both randomized experiments and observational studies. In randomized experiments, the true propensity score is known and is defined by the study design. In observational studies, the true propensity score is not known. However, it can be estimated using the study data (Austin 2011).

¹⁰ The assumption of selection on observables requires that conditional on the observed variables, the assignment to treatment is random.

Econometric literature offer various methods of estimation of the conditional probability of receiving a treatment (Guo and Fraser 2010). In practice, the propensity score is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics. Other proposed methods to estimate the propensity score are: bagging or boosting (McCaffrey et al. 2004), recursive partitioning or tree-based methods (Lee et al. 2010), random forests (Lee et al. 2010), neural networks (Setoguchi et al. 2008) and series logit estimator (Hirano et al. 2003). In our study, we use a generalized boosted model (GBM).¹¹ In this model, the analyst does not need to specify functional forms of the predictor variables. In addition, GBM permits nonlinear and interaction effects to be captured (McCaffrey et al. 2004). Finally, the data-adaptive algorithm on which this method is grounded leads to estimations of the propensity score that balance the observable covariates of the treatment and control group, which is a very valuable feature of GBM when it is used in the context of PSM. For all these reasons, GBM constitutes a highly suitable model to be used in the context of the PSM (Chowa et al. 2013).

Table 2 presents the results of the estimation of equation 1 (selection equation).¹² It can be observed that the variables which capture the greatest degree of influence in the probability of attending a Spanish private school are the years of study of the father and mother (21 and 16%, respectively), followed by the variables which proxy the degree of possessions in the household (number of TVs, PCs, video game consoles, MP4s, study room). The influence of the employment of the parents is also important. The dummies which approximate the employment of the mother account for 5.7% and those of the father 10.6%. More qualified occupations lead to an increase in the probability of attending a private school.

At the bottom of Table 2, several measures of the goodness of fit of the model are showed. This is because when using categorical dependent variables, it is difficult to evaluate the overall fit of the estimated model (Rehm 2005, p. 19). Pseudo- R^2 for the train and test samples and the percentage of correctly predicted observations (PCP) are reported on Table 2. Although PCP reaches practically 64%, it is difficult to evaluate whether this qualifies as a “good” or a “bad” model. In any event, this percentage shows that the model is at least more successful than a simplistic model that trivially predicts that all observations will be a one (Rehm 2005). As PCP suffers from various problems (Herron 1999), other additional measures have been calculated

¹¹ We are assuming homogeneity in response across observed covariates. Lehrer and Kordas (2013) demonstrate that when the treatment effects vary in an unsystematic manner with the true propensity score, there are gains from using a matching algorithm based on propensity scores estimated via binary regression quantiles.

¹² In the estimation of the propensity score, only those variables that could affect both the choice of a private school and the students’ academic performance were included (no consideration is taken of either the variables which can potentially contribute to explaining the differences in educational outcomes but which do not influence the choice of school, such as study habits, for example, nor those which could be determinants of that choice but do not influence the educational skills cited, such as the distance to the school, for example). In addition, only those variables which are potential predictors of educational outcomes and which occur prior to the choice of school (or were stable between the time of the choice of school and the time of the outcome assessment) were included as explanatory variables in equation 1 (Caliendo and Kopeinig 2008). Material that point out all the observables are listed in Table 1 and case-wise deletion was used to handle missing data.

Table 2 Results of the GBM

Variable	Influence
Parents education: Father's education (years)	21.08
Mother's education (years)	16.02
Household possessions	
Number of Videogames at home	9.36
Number of MP4s at home	8.20
Number of TVs at home	6.57
Student has own room to study	5.99
Number of PCs at home	4.51
Number PayTV at home	3.58
Parents occupation status (Reference Blue collar low skilled)	
Father blue collar high skilled	2.84
Father white collar low skilled	6.76
Father white collar high skilled	1.01
Mother blue collar high skilled	1.76
Mother white collar low skilled	1.23
Mother white collar high skilled	2.73
Student born in	
Spain (non-immigrant)	0.98
Africa	0.76
Asia	2.08
Europe	2.07
Latin America	0.57
Number of books	1.91
Best num iterations	16453
Train R^2	0.084
Test R^2	0.045
Percentage of correct predictions (PCP)	63.4%
Percentage reduction in error (PRE) ^b	36.2%
Expected percentage of correct predictions (ePCP) ^c	54.4%
Train fraction	0.5
Bag	0.5
Shrinkage factor	0.0005
Distribution	Logistic
Max num interactions	4 ^a
Max num iterations	20,000

Table 2 continued

Variable	Influence
Seed	0

Dependent variable $P (Z = 1)$

^aMcCaffrey et al. (2004) provides a detailed description of how GBM handles interaction terms and recommends a maximum of four splits for each simple tree used in the model, which allows all four way interactions between all covariates to be considered for optimizing the likelihood function at each iteration.

^bPRE (percentage reduction in error) is defined as the quotient $(PCP - PMC)/(1 - PMC)$, where PCP is the percentage correctly predicted and PMC is the percentage of observations in the modal category of the observed data. Therefore, PRE seeks to compare the information provided by probit fitted categories (PCP) with the classification errors a researcher would make if she naively assigned all fitted categories to the modal category (PMC).

^cePCP is the expected percentage of correct predictions and is defined as the sum over all observations i of the probability that $\hat{y}_i = y_i$, (see equation 12 in Herron 1999)

and are shown in Table 2: the percentage of reduction in error (PRE) and the expected percentage of correct predictions (ePCP). The final part of the table shows various parameters used in the estimation of the GBM.

Figure 1 shows the distribution of the predictions of the propensity scores estimated for the individuals from public (CG) and private schools (TG). It can be clearly observed, in the distribution graph, that there is a very broad area of common support. In other words, individuals in the treatment group have individuals in the control group with whom they can be compared, as their propensity scores are the same.

4.1.2 Matching and resampling estimation

After estimating the propensity score, the matching process was then undertaken. Various algorithms can be found in the literature regarding the performance of this process: greedy matching, optimal matching and fine balance (Guo and Fraser 2010). The present study uses the first of these, which may be applied via a range of variants

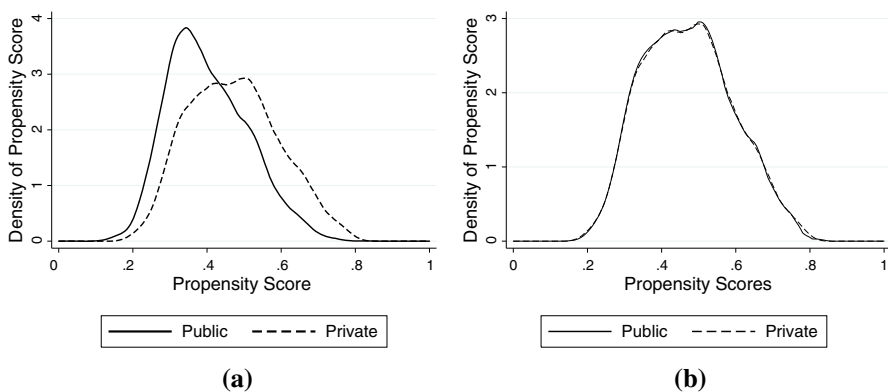


Fig. 1 Propensity score distribution by school type. **a** Full sample and **b** matched sample

(Smith and Tood 2005). Concretely, our study applied nearest neighbour matching (hereinafter NNM) and methods based on kernel functions (hereinafter KM), as well as several of the options permitted by NNM (with and without replacement, with caliper and without caliper, 1 to 1, 1 to 2 and 1 to 3).¹³ The KM was applied with different bandwidths. This was done in an attempt to test the sensitivity of the matching to the different estimation methods.¹⁴

Here, we present the results from estimations of the Epanechnikov kernel type KM with a bandwidth of 0.03, since this was the algorithm that best matched the individuals from both the treatment group and the control group.¹⁵ The sample was only reduced by 9 individuals from the control group who were not paired with any individual from the treatment group. The remaining individuals from the control group were weighted on the basis of the number of times that they were matched with treated individuals. These weights are required to be used in the subsequent statistical analyses.

All the previous analysis permit us to be confident about the quality of our matching as they show the fulfilment of one of the key assumptions in PSM applications: overlap or common support.¹⁶

Having selected the subsample of comparable individuals, the following step in the PSM is to calculate the matching estimator of the average treatment effect (ATE). We apply the *psmatch2* module in Stata to obtain the ATE.¹⁷ Results are displayed in Table 3 (column 2) for science and foreign language (English). It shows that private schools have a positive and statistically significant effect on both science and foreign language (English) scores for students evaluated in 2010 Aragón ED.

¹³ The first of these (NNM) matches each treated individual with that non-treated individual having the most similar propensity score value. This is to say, in nearest neighbour matching, Stata selects the control(s) nearest to each treated observation for comparison. KM constructs matches using all the individuals in the potential control sample in such a way that it gathers more information from those who are closer matches and less from distant observations. In so doing, KM uses comparatively more information than other matching algorithms (Guo and Fraser 2010, chapter 7).

¹⁴ We applied the coarsened exact matching as a robustness technique obtaining worse results in terms of similarity between treatment and control groups generated. Results are available upon request.

¹⁵ Results supplied by the different matching estimation methods led to similar conclusions. They are not supplied here but are available from the authors upon request.

¹⁶ Other results devoted to test the matching quality are shown in Appendix. In particular, Table 5 shows the differences in the average values of propensity scores and covariates for the whole sample and the paired sample. The last two rows in this table show the median absolute standardized bias (Rosenbaum and Rubin 1985) before and after matching. As can be inferred, KM has reduced covariate imbalance on all variables. Figure 2 shows graphically the pre- and post-matching bias for each of the variables included in the estimation of the propensity score. Figure 3 depicts the distribution of these same variables by type of school for the complete sample (figures on the left) and the matched sample (figures on the right).

¹⁷ This command calculates the ATE as a weighted average of the ATT (average effect on treated) and the ATU (average effect on untreated). This is a very common definition of the ATE in the literature (see for instance, Böckerman et al. 2013; Gangl 2014). An alternative way to calculate the ATE is by weighting observations by the inverse of the calculated propensities scores (Hirano et al. 2003). In order to check the robustness of the ATE, we also calculated it applying this last method, i.e., using the propensities as sampling weights. For this, we used the Stata's *teffects* module. Results are similar to those shown in Table 3 and are available upon request.

Table 3 Average treatment effect of private schools

Scores	ATE	S.E.	T-stat
Science	10.65***	2.43	4.39
Foreign language (English)	13.69***	2.38	5.76

Matching algorithm used is an Epanechnikov kernel with a bandwidth of 0.03

ATE standard errors corrected by bootstrapping (nrep=1000)

*** p value=0.000

4.1.3 Sensitivity analysis: selection on unobservables

The main issue with cross-sectional matching analysis is that there may be a problem of hidden bias due to the effect of selection on unobserved heterogeneity and any positive association between a pupil's treatment status and test score outcomes may not represent a causal effect (Bradley et al. 2013). To deal with this problem, we perform a sensitivity analysis which allows us to evaluate to what extent our results are robust to a potential imbalance in the unobservable factors (Altonji et al. 2008; Peel 2014, among others).

We apply the Rosenbaum's (2002) procedure for bounding the treatment effect estimates. There we give the results of the p value from Wilcoxon sign-rank tests for the ATE while setting the level of hidden bias to a certain value γ , which reflects our assumption about unmeasured heterogeneity or endogeneity in treatment assignment (expressed in terms of the odds ratio of differential treatment assignment due to an unobserved covariate).¹⁸ At each γ , we calculate a hypothetical significance level "p value critical," which represents the bound on the significance level of the treatment effect in the case of endogenous self-selection into treatment status.

Table 4 shows that robustness to hidden bias varies across the two variables. The finding of a positive effect of private management on science is the least robust to the possible presence of selection bias. The critical level of γ at which we would have to question our conclusion of a positive effect is between 1.12 and 1.15, i.e., it is attained if an unobserved covariate caused the odds ratio of treatment assignment to differ between treatment and control cases by a factor of about 1.15. For foreign language model, it would require a hidden bias of γ between 1.30 and 1.33 to render spurious the conclusion of a positive benefit effect on "private school".¹⁹

A critical value of 1.15 suggests that individuals with the same X -vector differ in their odds of participation by a factor of 1.15 or 15%. It is important to note that these are worst-case scenarios. Hence, a critical value of 1.15 does not mean that unobserved heterogeneity exists and that there is no effect of treatment on the outcome variable. This result only states that the confidence interval for the effect would include zero if an unobserved variable caused the odds ratio of treatment assignment to differ

¹⁸ For a mathematical demonstration, see DiPrete and Gangl (2004).

¹⁹ In addition, we calculated the Hodges–Lehmann point estimates and its confidence intervals obtaining the same critical values.

Table 4 Rosenbaum bounds for management benefit treatment effects

Variable	γ^a	<i>P</i> value critical
Science (<i>N</i> = 2879 matched pairs)	1	8.5e−07
	1.03	.000021
	1.06	.000296
	1.09	.002644
	1.12	.015351
	1.15	.060581
	1.18	.169812
	1.21	.353833
	1.24	.575602
	Foreign language (English) (<i>N</i> = 2879 matched pairs)	1
1.03		4.7e−13
1.06		4.8e−11
1.09		2.8e−09
1.12		9.8e−08
1.15		2.2e−06
1.18		.000031
1.21		.000306
1.24		.002087
1.27		.010261
1.30	.037393	
	1.33	.103961

^agamma: log odds of differential assignment due to unobserved factors

between treatment and comparison groups by 1.15. Moreover, this variable's effect on the outcome would have to be so strong that it almost perfectly determines the outcome in each pair of matched cases in the data. However, even if there is unobserved heterogeneity to a degree of 15% in the group of science, the inference about the treatment effect would not be changed.

To repeat, the Rosenbaum bounds are in this sense a “worst-case” scenario. Nonetheless, they convey important information about the level of uncertainty contained in matching estimators by showing just how large the influence of a confounding variable must be to undermine the conclusions of a matching analysis.²⁰

²⁰ Additionally, we test our estimation with another sensitivity analysis proposed by Ichino et al. (2008) This consists in calculating the ATE under different possible scenarios of deviation of conditional independence assumption (CIA). To do so, the authors impose values to parameters that characterize the U distribution in order to simulate the ability to generate bias in the unobservable and recalculate the parameter value with the inclusion of the influence of simulated unobserved variable. Results are available upon request. This approach has been widely used in the literature (Binder and Coad 2013; Caliendo and Künn 2015, among others). Others types of sensitivity analysis have been proposed in the literature. For example Altonji et al. (2005) applied a similar idea to the Heckman selection model.

In conclusion, this analysis allows confirm that only a large amount of unobserved heterogeneity would alter the inference about the estimated effects. Even so, as suggested by Lee and Lee (2009), it is always advisable to be very cautious when interpreting the results.

4.2 Post-matching analysis: HLM results

The application of the PSM has permitted us to obtain an unbiased estimation of the ATE with regard to the observable variables (X) which distinguish the members of the treatment and the control groups and which are potentially important in the determination of outcomes (Y).

However, the potential influences on the educational outcome include, usually, more variables than those that simultaneously influence attendance to a private school, that is to say those considered in the construction of the propensity score. Given this situation, the calculation of a more robust effect of an intervention in the educational context requires the contrast of the influence of those other factors (X') which are potentially important in the determination of Y but do not influence school choice. For this, it is fundamental to carry out a post-matching analysis. Two types of influence deserve attention: the characteristics of the schools at which individuals are educated and the attributes of the students not incorporated into the calculation of the propensity score (those contemporary to the receipt of the treatment).

The testing of the importance of these two characteristics can be performed via a regression model on the matched sample. Of all the available regression models, the HLMs adapt best to the peculiarities of our dataset.²¹ Their main advantage is that they permit differentiation between those influences acting at the student level (first level of analysis) and those acting at the class and school level (second and third levels).²² They are, therefore, models which are especially appropriate for working with data nested at various levels, such as those supplied by almost all educational databases, including *2010 Aragón ED*.²³

²¹ HLM are similar to OLS concerning the way in which they weigh the observations (see Yitzhaki 1996 for a discussion of OLS weights). Both weigh the observations differently to PSM. We are grateful to an anonymous referee for making this point. In any case, our purpose with the HLM is not to compare the ATE that it supplies with that obtained via PSM.

²² Multilevel models, such as HLM, are built on the Moulton's (1990) work of clustering. The insight provided by Moulton's work was that when individuals within the aggregated level are clustered, so that they are in fact more similar to one another than individuals belonging to another cluster group, the OLS assumption that observations are independent and identically distributed is violated. For this reason, the estimation by OLS can result in a downward bias in the estimated standard errors of estimates leading the analyst to conclude the aggregate level effects are statistically significant when they are in fact not. Multilevel models have the benefit of allowing for partial pooling of coefficients towards the completely pooled OLS estimate which according to Gelman (2006) can be a more effective estimation strategy. Simulations using a dataset with students clustered within classrooms and classrooms within schools suggest that modelling the clustering of the data using a multilevel method is a better approach than clustering the standard errors of OLS estimate (Cheah 2009).

²³ Bryk and Raudenbusch (1988) recommend the use of this type of general model when analysing the effects of schools on educational outcomes. There exist multiple applications of this methodology to the

Table 8 in Appendix shows the results of the application of HLM to the matched sample.²⁴ The dependent variables in the regression are the scores in science and foreign language (English). The left-hand side of the table presents the results from the two-level model for foreign language (English). The right-hand side offers the results for the three-level model, which is more adequate for the estimation of the determinants of the outcomes in science. The models were estimated by imposing fixed effects on the parameters (with the exception of the independent term), after rejecting the null hypothesis that there existed statistically significant random effects.

The predictor which has greatest interest in our study is attendance at a private school. It can be observed that this effect is positive and significant both for science and foreign language (English). The coefficient estimated in science is 23.53 points, equivalent to two-tenths of the standard deviation (0.23), which corresponds to a small level size (Cohen 1988). Similarly, the value of the coefficient estimated in foreign language (English) is small (0.2 points of the standard deviation).

5 Conclusions

This paper has carried out an evaluation of the effectiveness of an educational policy that has been widely applied under different guises worldwide: public funding of privately owned and run schools. Empirical studies devoted to evaluating the impact of this educational intervention have expanded notably over the last 2 decades using different methodological approaches. However, results of this research remain inconclusive. Our work contributes to this literature providing new evidence from the Spanish experience. Our methodological strategy has been guided by our database's empirical features: (a) non-random distribution of students between Spanish private and public schools and (b) clustered data. These two characteristics led us to apply, in the first step, a propensity score-matching analysis that permitted us to define a homogeneous subsample of students attending public and private schools. In the second step, we estimated the impact of attending to a publicly funded privately run school by means of a hierarchical linear model that allowed us to cope with the clustered structure of our database.

The analysis performed in this study has underlined the existence of a slight and significant advantage in performance of the self-governing model of private schools

educational context. Among these are Willms (2006), Somers et al. (2004) and Mancebón et al. (2012), the last of these being applied to Spanish data from PISA 2006.

²⁴ Previously to the estimation of the HLM, we evaluated the appropriateness of applying it to our data. For this, we calculated the intra-class correlation (ICC) values of the null model of science and foreign language (English) performance (the two being the dependent variables of the regression). If the ICC were zero, a hierarchical model would not be necessary, since in this case the total variance of the scores would not be explained by the differences existing between students attending different classes or schools. Results of these calculations for an HLM at two levels and three levels are offered in Appendix (Tables 6 and 7). These results (which show that the class level explains a small percentage of the variance of the results in foreign language (English), but does explain a higher percentage of the results in science) leads us to apply a two-level model for achievement in a foreign language (English) and a three-level model for science. At any rate, results for three-level model for English and two-level model for science lead to the same conclusions and are available upon request.

in the region of Aragón (Spain). As the main difference between the Spanish private and public schools in our sample is the way in which they are run, our results point to a slight superiority of autonomy against centralization in the management of primary schools.

At any rate, we cannot lose sight of the magnitude of the effect of self-governing private schools on students' abilities is quite low (a little over a 2% of a standard deviation) which puts our work very close to those studies where no effect of privately run schools has been found. The results coming from such a diversity of studies on the topic (different countries, methodological approaches, type of schools and outcomes) might be actually pointing that the model of schools' management (public or private) is not a key factor in the explanation of the elusive concept of school quality. Concerning this last topic, what our estimations from the HLM have shown (see Tables 6 and 7 in Appendix) is that school itself matters for students' outcomes, a result in concordance with the international evidence on education production functions (Hanushek and Woessmann 2014). The most robust conclusion of this literature to date is that the influence of schools does not occur through traditional inputs and that the discovery of relevant inputs of school quality remains a true Gordian knot. This is an important result as it points to the necessity of further research inside the black box of schools. In our opinion, the focus should be reoriented to obtain new indicators that truly capture the essence of what is occurring inside the classroom.

Appendix

See Tables 5, 6, 7, 8 and Figs. 2 and 3.

Table 5 Average differences based on school type for the variables in the pre- and post-matching samples and bias reduction

Variable		Mean		% Bias	% Reduct Bias	<i>t</i> test	
		Treated	Control			<i>t</i>	<i>p</i> > <i>t</i>
Propensity score	Unmatched	0.47	0.40	60.8		24.87	0.00
	Matched	0.47	0.47	0.9	98.5	0.32	0.75
Mother white collar high skilled	Unmatched	0.37	0.24	29.7		12.15	0.00
	Matched	0.37	0.37	0.7	97.6	0.26	0.80
Mother white collar low skilled	Unmatched	0.39	0.42	- 5.4		- 2.20	0.03
	Matched	0.39	0.40	- 2.0	62.5	- 0.77	0.44
Mother blue collar high skilled	Unmatched	0.03	0.04	- 7.6		- 3.06	0.00
	Matched	0.03	0.03	- 1.7	77.5	- 0.71	0.48

Table 5 continued

Variable		Mean		% Bias	% Reduct Bias	<i>t</i> test	
		Treated	Control			<i>t</i>	<i>p</i> > <i>t</i>
Mother blue collar low skilled	Unmatched	0.21	0.30	-21.7		-8.74	0.00
	Matched	0.21	0.20	2.3	89.6	0.93	0.36
Father white collar high skilled	Unmatched	0.49	0.31	36.7		14.98	0.00
	Matched	0.49	0.50	-1.2	96.8	-0.44	0.66
Father white collar low skilled	Unmatched	0.23	0.26	-6.5		-2.61	0.01
	Matched	0.23	0.24	-2.0	68.8	-0.77	0.44
Father blue collar high skilled	Unmatched	0.23	0.35	-27.5		-11.07	0.00
	Matched	0.23	0.21	3.4	87.8	1.38	0.17
Father blue collar low skilled	Unmatched	0.05	0.07	-10.7		-4.29	0.00
	Matched	0.05	0.05	-0.3	97.2	-0.13	0.90
Mother's education (years)	Unmatched	12.34	10.78	33.9		13.66	0.00
	Matched	12.34	12.49	-3.2	90.7	-1.26	0.21
Father's education (years)	Unmatched	12.34	10.78	33.5		13.53	0.00
	Matched	12.34	12.47	-2.9	91.4	-1.13	0.26
Student born in Spain	Unmatched	0.91	0.84	21.0		8.36	0.00
	Matched	0.91	0.91	0.9	95.8	0.38	0.70
Student born in Africa	Unmatched	0.00	0.01	-9.3		-3.65	0.00
	Matched	0.00	0.00	1.0	89.4	0.60	0.55
Student born in Asia	Unmatched	0.01	0.01	-1.0		-0.42	0.68
	Matched	0.01	0.00	2.1	-101.6	0.89	0.37
Student born in Europe	Unmatched	0.03	0.06	-12.8		-5.11	0.00
	Matched	0.03	0.04	-2.5	80.9	-1.06	0.29
Student born in Latin America	Unmatched	0.04	0.06	-10.9		-4.36	0.00
	Matched	0.04	0.04	0.2	97.8	0.10	0.92
Student born in Arab country	Unmatched	0.01	0.02	-8.1		-3.19	0.00
	Matched	0.01	0.01	-0.7	91.8	-0.30	0.76

Table 5 continued

Variable		Mean		% Bias	% Reduct Bias	<i>t</i> test	
		Treated	Control			<i>t</i>	<i>p</i> > <i>t</i>
Number of books at home	Unmatched	0.60	0.50	18.7		7.56	0.00
	Matched	0.60	0.61	-2.0	89.4	-0.76	0.45
Student has own room to study	Unmatched	0.96	0.94	8.1		3.23	0.00
	Matched	0.96	0.96	-1.2	84.9	-0.51	0.61
Internet at home	Unmatched	0.88	0.84	12.7		5.06	0.00
	Matched	0.88	0.89	-3.4	73.4	-1.40	0.16
Number of TVs at home	Unmatched	2.15	2.08	9.8		3.96	0.00
	Matched	2.15	2.15	0.8	92.1	0.30	0.77
Number of PCs at home	Unmatched	1.63	1.49	17.2		6.97	0.00
	Matched	1.63	1.66	-4.1	76.2	-1.56	0.12
Number of pay TVs at home	Unmatched	0.46	0.43	4.5		1.82	0.07
	Matched	0.46	0.48	-3.4	24.1	-1.24	0.21
Number of video games at home	Unmatched	1.82	1.66	16.3		6.62	0.00
	Matched	1.82	1.84	-1.9	88.6	-0.71	0.48
Number of MP4s at home	Unmatched	1.11	0.93	18.3		7.47	0.00
	Matched	1.11	1.14	-3.0	83.4	-1.11	0.27
Abs (bias)	Unmatched			17.7		617.20	0.00
	Matched			1.9		31.47	0.09

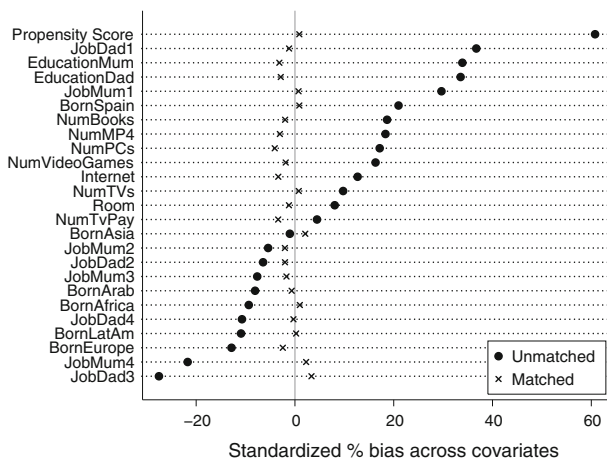


Fig. 2 Pre- and post-matching bias between public versus private schools standardized

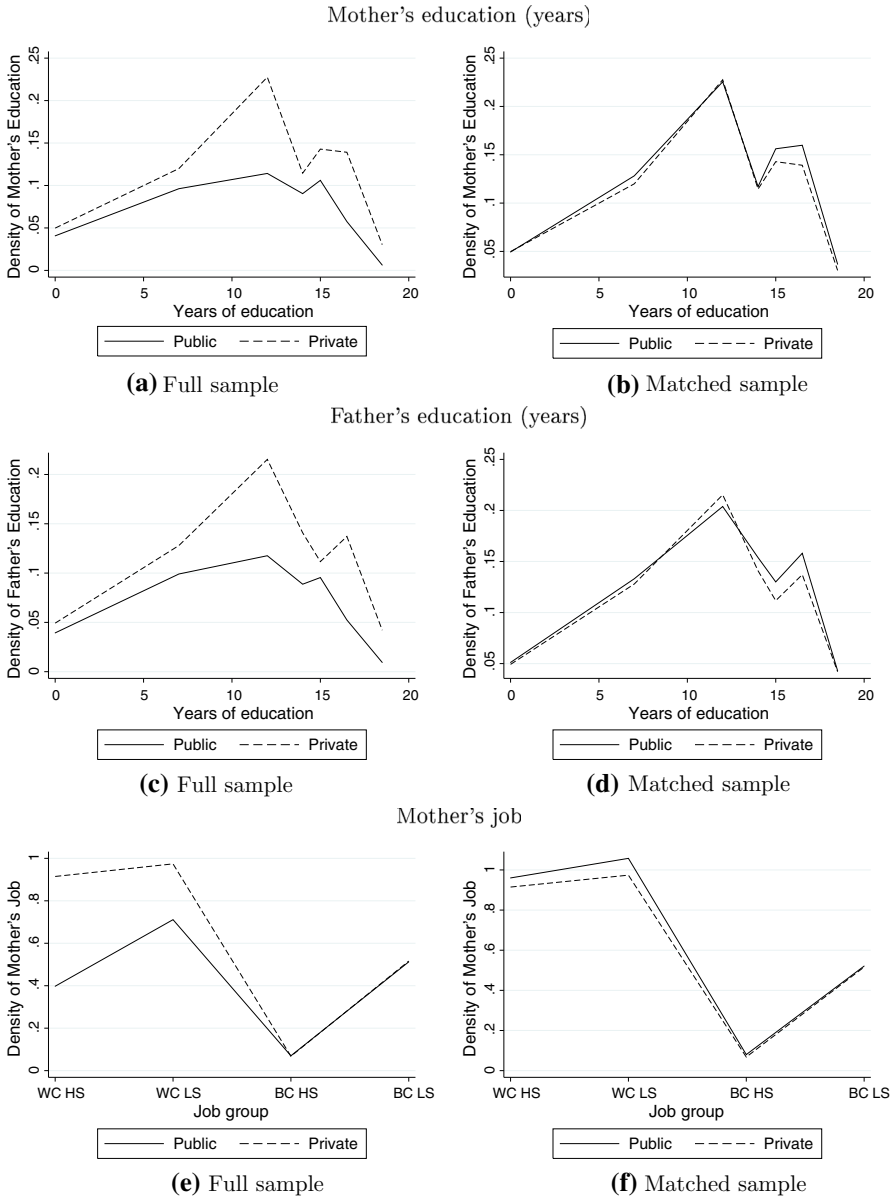


Fig. 3 Distribution of the variables in the unmatched and matched samples

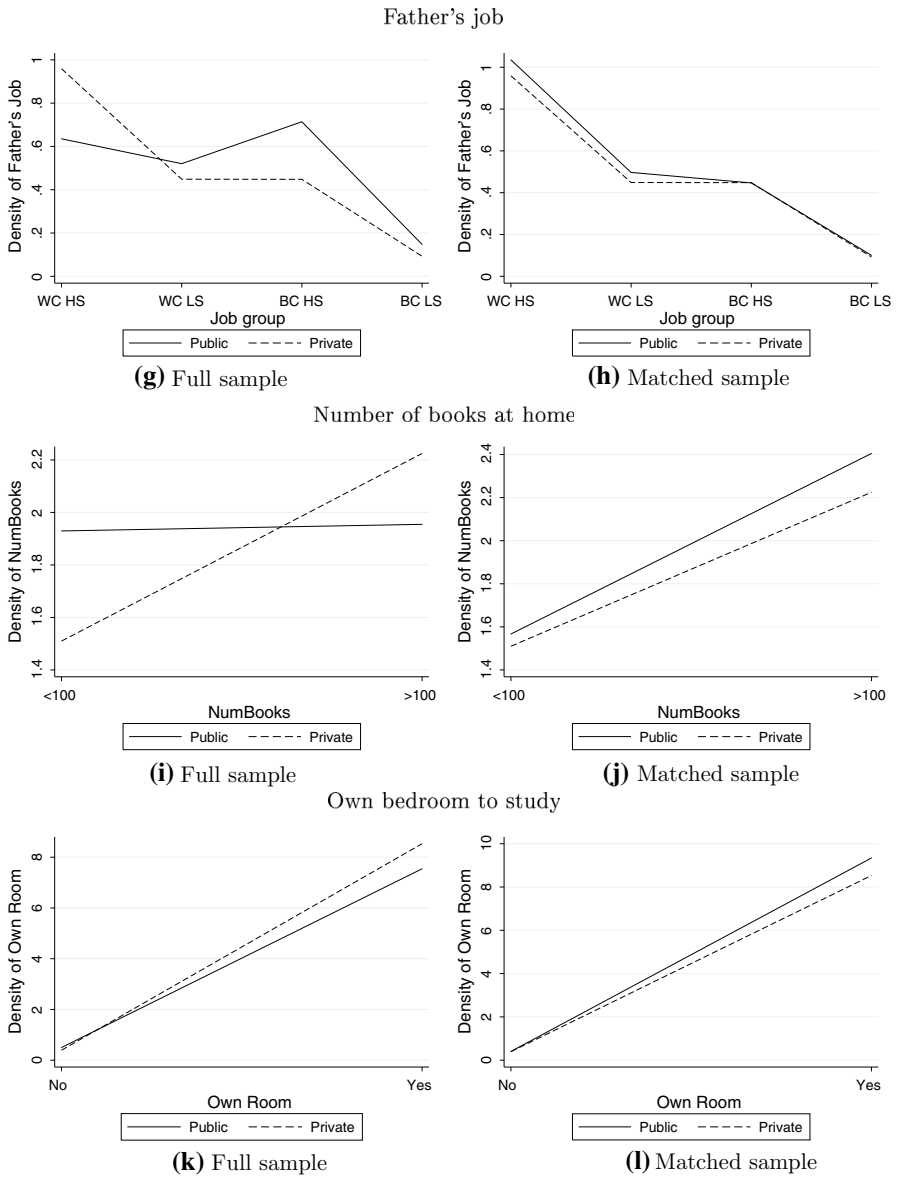


Fig. 3 continued

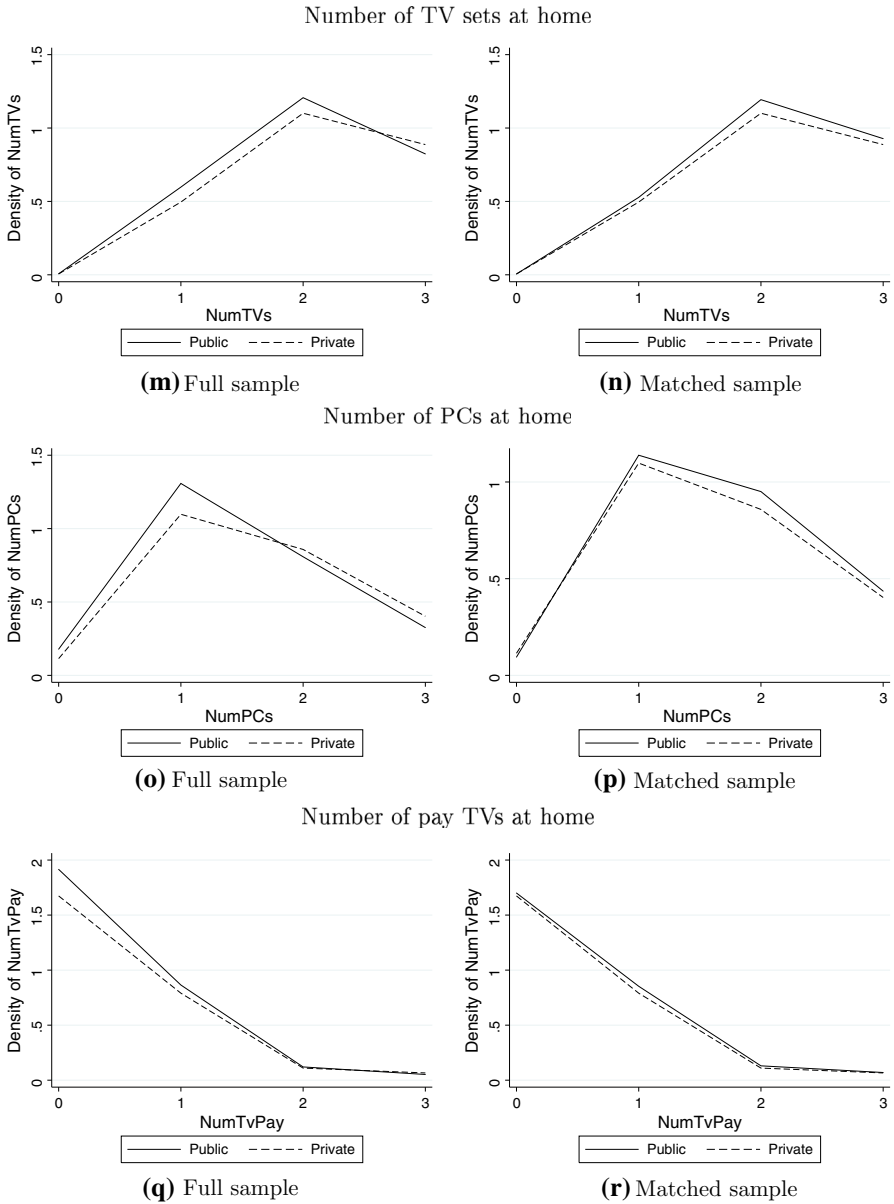


Fig. 3 continued

Table 6 HLM regression: random effects (3-levels)

	Science		Foreign language (English)	
	Null model	Complete model	Null model	Complete model
Schools	1765.68	1563.41	3127.05	2199.61
Classes	1261.67	1139.06	462.25	543.82
Students	5635.50	3923.77	5136.59	3462.15
Total	8662.86	6626.24	8725.89	6205.58
ICC (school)	20.4%		35.8%	
ICC (class)	14.6%		5.3%	
% of total variance explained by variables		23.5%		28.9%
% of level 1 (students) variance explained by variables		30.4%		32.6%
% of level 2 (classes) variance explained by variables		9.7%		-17.6%
% of level 3 (schools) variance explained by variables		11.5%		29.7%

Table 7 HLM regression: random effects (2-levels)

	Science		Foreign language (English)	
	Null model	Complete model	Null model	Complete model
Schools	2625.54	2264.81	3386.08	2418.67
Students	6160.68	4403.65	5333.38	3688.13
Total	8786.22	6668.46	8719.46	6106.81
ICC (schools)	29.9%		38.8%	
% of total variance explained by variables		24.1%		30.0%
% of level 1 (students) variance explained by variables		28.5%		30.8%
% of level 2 (schools) variance explained by variables		13.7%		28.6%

Table 8 Estimation of fixed effects with robust standard errors in the HLM

Two-level model Foreign language (English)		Three-level model Science	
Variable	Coeff.	Variable	Coeff.
<i>School variables (Level 2)</i>		<i>School variables (Level 3)</i>	
Intercept	368.63*** (35.3)	Intercept	507.46*** (46.31)
Private school	20.65*** (7.81)	Private school	23.53*** (7.93)
TERUEL province ^a	9.91	TERUEL province ^a	2.03
Ref: Huesca province	(16.29)	Ref: Huesca province	(16.32)
ZARAGOZA province ^a	6.03	ZARAGOZA province ^a	- 11.61
Ref: Huesca province	(15.06)	Ref: Huesca province	(15.29)
Municipality size (inhabitants in thousands)	1.2*** (0.44)	Municipality size (inhabitants in thousands)	- 0.05 (0.44)
ZARAGOZA city	- 776.9 (290.29)	ZARAGOZA city	11.9 (287.66)
		<i>Class variables (Level 2)</i>	
Percentage girls at school	14.66 (16.83)	Percentage girls in class	- 30.65 (24.91)
Percentage repeaters at school	36.21 (25.93)	Percentage repeaters in class	- 21.12 (38.88)
Percentage of pupils living over 5 years in Spain at school at school	5.28 (27.33)	Percentage of pupils living over than 5 years in Spain at school in class	0.13 (39.03)
Percentage of students whose mother is white collar high skilled at school	9.33 (18.38)	Percentage of students whose mother is white collar high skilled in class	- 21.71 (28.76)
Percentage of students whose mother is white collar low skilled at school	21.81 (15.61)	Percentage of students whose mother is white collar low skilled at class	- 14.76 (24.43)
Percentage of students whose mother is blue collar high skilled at school	- 2.78 (36.19)	Percentage of students whose mother is blue collar high skilled in class	- 94.98 (57.4)
Average number of years of mothers education at school	4.84*** (1.56)	Average number of years of mothers' education ins class	1.25 (2.29)
<i>Student variables (Level 1)</i>		<i>Student variables (Level 1)</i>	
Female student	20.15*** (2.31)	Female student	- 11.83 (2.41)

Table 8 continued

Two-level model Foreign language (English)		Three-level model Science	
Variable	Coeff.	Variable	Coeff.
If student has repeated any year	-40.72 (5.50)	If student has repeated any year	-27.98 (5.72)
Student's mother is white collar high skilled	11.97*** (3.66)	Student's mother is white collar high skilled	11.3*** (3.82)
Student's mother is white collar low skilled	0.7 (3.15)	Student's mother is white collar low skilled	1.26 (3.27)
Student's mother is blue collar high skilled	2.21 (7.15)	Student's mother is blue collar high skilled	1.19 (7.44)
Pupil's father is white collar high skilled	9.34* (5.65)	Pupil's father is white collar high skilled	1.06 (5.93)
Student's father is white collar low skilled	1.48 (5.70)	Student's father is white collar low skilled	-2.26 (5.98)
Student's father is blue collar high skilled	3.21 (5.61)	Student's father is blue collar high skilled	0.76 (5.89)
Mother's years of education	1.31*** (0.31)	Mother's years of education	1.51*** (0.32)
Over 5 years living in Spain	-7.24 (5.64)	Over 5 years living in Spain	18.81*** (5.87)
Over 100 books at home	7.43*** (2.56)	Over 100 books at home	13.18*** (2.7)
Student uses books frequently	13.87*** (2.8)	Student uses books frequently	12.99*** (2.97)
Number of TVs at home	-5.16 (1.54)	Number of TVs at home	-6.36 (1.62)
2 h of homework every day	-3.53 (3.43)	2 h studying every day	-10.01 (3.63)
Over 2 h studying every day	-11.95 (2.62)	Over 2 h studying every day	-12.09 (2.78)
Student needs help in homework	-23.25 (2.87)	Student needs help in homework	-27.24 (3.02)
Parents check diary but not homework	-2.86 (4.4)	Parents check diary but not homework	-11.24 (4.62)
Parents check homework but not diary	-1.12 (3.7)	Parents check homework but not diary	-10.69 (3.93)

Table 8 continued

Two-level model Foreign language (English)		Three-level model Science	
Variable	Coeff.	Variable	Coeff.
Parents check both homework and diary	- 10.83 (3.06)	Parents check both homework and diary	- 18.06 (3.24)
Private tutoring	- 17.26 (4.18)	Private tutoring	- 19.68 (4.41)
Student Attitude (Student always finishes homework)	9.78* (5.26)	Student Attitude (Student always finishes homework)	16.46*** (5.57)
Student Aptitude (Student answers homework correctly)	11.75*** (3.81)	Student Aptitude (Student answer homework correctly)	16.53*** (4.00)
RELCEN ^b (Factor 1)	4.58*** (1.18)	RELCEN ^b (Factor 1)	0.76 (1.26)
SELFCONF ^b (Factor 2)	21.63*** (1.34)	SELFCONF ^b (Factor 2)	16.13*** (1.42)
PERCAMB ^b (Factor 3)	- 3.83 (1.14)	PERCAMB ^b (Factor 3)	- 7.38 (1.24)

Bootstrapped standard errors in parenthesis (calculated with R package lme4, nrep=1000)

^aThis variable proxies the location of the school in the region of Aragón. This region consists of three provinces: Zaragoza, Teruel and Huesca (this last being the category of reference)

^bRELCEN contains information regarding the evaluation the child makes of his or her school (if the centre has cultural and sports activities, if the pupil uses the school's library, if the installations are well maintained for, etc.). SELFCONF synthesizes the information offered by variables related to the self-perception of the pupils' academic capacity (if pupils understand what they read, if they express themselves well, if they write correctly, if they are good at languages, etc.). PERCAMB, finally, reflects the subjective perceptions of the school environment (if there is a good atmosphere in the pupil's class, if his/her classmates help each other, if the pupil has a good relationship with his/her teachers, if the teachers stimulate their pupils, etcetera). These variables were obtained from a principal components analysis applied to the data concerning the school environment. These data proceed from the answers supplied by the pupils evaluated in the 2010 Aragón ED

References

- Allen R, Vignoles A (2015) Can school competition improve standards? The case of faith schools in England. *Empir Econ* 50(3):959–973
- Altonji J, Elder T, Taber C (2005) Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. *J Polit Econ* 113(1):151–184
- Altonji J, Elder T, Taber C (2008) Using selection on observed variables to assess bias from unobservables when evaluating Swan–Ganz catheterization. *Am Econ Rev* 98(2):345–350
- Austin P (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res* 46(3):399–424
- Böckerman P, Bryson A, Ilmakunnas P (2013) Does high involvement management lead to higher pay? *J R Stat Soc Ser A Stat Soc* 176(4):861–885
- Bernal J (2005) Parental choice, social class and market forces: the consequences of privatization of public services in education. *J Educ Policy* 20(6):779–792

- Bettinger E (2011) Educational vouchers in international contexts. In: Hanushek E, Machin S, Woessmann L (eds) *Handbook of the economics of education*, vol 4. Elsevier Science & Technology, North-Holland, pp 551–572
- Binder M, Coad A (2013) Life satisfaction and self-employment: a matching approach. *Small Bus Econ* 40(4):1009–1033
- Bradley S, Migali G, Taylor J (2013) Funding, school specialisation and test scores: an evaluation of the specialist schools policy using matching models. *J Hum Cap* 7(1):76–106
- Bryk A, Raudenbusch S (1988) Toward a more appropriate conceptualization of research on school effects: a three-level hierarchical linear model. *Am J Educ* 97(1):65–108
- Caliendo M, Künn S (2015) Getting back into the labor market: the effects of start-up subsidies for unemployed females. *J Popul Econ* 28(4):1005–1043
- Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. *J Econ Surv* 22(1):31–72
- Cheah B (2009) Clustering standard errors for modeling multilevel data. Technical report, Columbia University, New York
- Chowa G, Masa R, Wretman C, Ansong D (2013) The impact of household possessions on youth's academic achievement in the Ghana Youthsave experiment: a propensity score analysis. *Econ Educ Rev* 33:69–81
- Chudgar A, Quin E (2012) Relationship between private schooling and achievement: results from rural and urban India. *Econ Educ Rev* 31(4):376–390
- Cohen J (1988) *Statistical power analysis for the behavioral sciences*, 2nd edn. Erlbaum Associates, Lawrence
- Coleman J, Hoffer T, Kilgore S (1982) *Secondary school achievement. Public, catholic and private schools compared*. Basic Books, Inc. Publishers, New York
- Crespo E, Santfín D (2014) Does school ownership matter? An unbiased efficiency comparison for regions of Spain. *J Prod Anal* 41(1):153–172
- Davies S (2013) Are there Catholic school effects in Ontario, Canada? *Eur Sociol Rev* 29(4):871–883
- DiPrete T, Gangl M (2004) Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. WZB discussion paper SP I 2004-101, Wissenschaftszentrum Berlin für Sozialforschung
- Doncel L, Sainz J, Sanz I (2012) An estimation of the advantage of charter over public schools. *Kyklos* 65(4):442–463
- Epple D, Romano R, Zimmer R (2016) Charter schools: a survey of research on their characteristics and effectiveness. In: Hanushek E, Machin S, Woessmann L (eds) *Handbook of the economics of education*, vol 5. Elsevier Science & Technology, North-Holland, pp 139–208
- Escardibul JO, Villarroya A (2009) The inequalities in school choice in Spain in accordance to PISA data. *J Educ Policy* 24(6):673–695
- Gangl M (2014) Matching estimators for treatment effects. In: Best H, Wolf C (eds) *The SAGE handbook of regression analysis and causal inference*. SAGE Publications Ltd, London, pp 251–276
- Gelman A (2006) Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics* 48(3):432–435
- Green C, Navarro-Paniagua M, Ximénez de Embún D, Mancebón M (2014) School choice and student wellbeing. *Econ Educ Rev* 38:139–150
- Gronberg T, Jansen D (2001) Navigating newly chartered waters. An analysis of charter school performance. Texas Public Policy Foundation, Austin
- Guo S, Fraser M (2010) *Propensity score analysis. Statistical methods and applications*. SAGE publications Ltd., London
- Hanushek E, Woessmann L (2014) Institutional structures of the education system and student achievement: a review of cross-country economic research. In: Strietholt R, Bos W, Gustafsson JE, Rosen M (eds) *Educational policy evaluation through international comparative assessments*. Waxmann, Munster
- Hanushek E, Kain J, Rivkin S, Branch F (2007) Charter school quality and parental decision making with school choice. *J Public Econ* 91(5–6):823–848
- Herron M (1999) Postestimation uncertainty in limited dependent variable models. *Polit Anal* 8(1):83–98
- Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189
- Ichino A, Mealli F, Nannicini T (2008) From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? *J Appl Econ* 23(3):305–327

- Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat* 86(1):4–29
- Kim Y (2011) Catholic schools or school quality? The effects of Catholic schools on labor market outcomes. *Econ Educ Rev* 30(3):546–558
- Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. *Stat Med* 29(3):337–346
- Lee M, Lee S (2009) Sensitivity analysis of job-training effects on reemployment for Korean women. *Empir Econ* 36(1):81–107
- Lefebvre P, Merrigan P, Verstraete M (2011) Public subsidies to private schools do make a difference for achievement in mathematics: longitudinal evidence from Canada. *Econ Educ Rev* 30(1):79–98
- Lehrer S, Kordas G (2013) Matching using semiparametric propensity scores. *Empir Econ* 44(1):13–45
- LODE (1985) Organic Law 8/1985, 3 July, Regulating education. Official Spanish State Bulletin 159
- Mancebón M, Ximénez-de-Embún DP (2014) Equality of school choice: a study applied to the Spanish region of Aragón. *Educ Econ* 22(1):90–111
- Mancebón M, Calero J, Choi A, Ximénez-de-Embún DP (2012) The efficiency of public and publicly-subsidized high schools in Spain. Evidence from PISA-2006. *J Oper Res Soc* 63(11):1516–1533
- McCaffrey D, Ridgeway G, Morral A (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 9(4):403–425
- Moulton B (1990) An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Rev Econ Stat* 72(2):334–338
- Murname R, Willett J (2011) *Methods matter*. Oxford University Press, New York
- Peel M (2014) Addressing unobserved endogeneity bias in accounting studies: control and sensitivity methods by variable type. *Account Bus Res* 44(5):545–571
- Rehm P (2005) Citizen support for the welfare state: determinants of preferences for income redistribution. WZB markets and political economy working paper SP II 2, Wissenschaftszentrum Berlin für Sozialforschung
- Rosenbaum P (2002) *Observational studies*. Springer, New York
- Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
- Rosenbaum P, Rubin D (1985) The bias due to incomplete matching. *Biometrics* 41(1):103–116
- Rubin D, Thomas N (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc* 95(450):573–585
- Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF (2008) Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* 17(6):546–555
- Smith H, Tood P (2005) Does matching overcome LaLonde's critique of non-experimental estimators? *J Econ* 125(1–2):305–353
- Somers M, McEwan P, Willms J (2004) How effective are private schools in Latin America? *Comp Educ Rev* 48(1):48–69
- Stuart E (2010) Matching methods for causal inference: a review and a look forward. *Stat Sci* 25(1):1–21
- Spanish Ministry of Education (2013) Spanish Education Statistics. Madrid, 2013. <https://www.meecd.gob.es>
- Urquiola M (2016) Competition among schools. Traditional public and private schools. In: Hanushek E, Machin S, Woessmann L (eds) *Handbook of the economics of education*, vol 5. Elsevier Science & Technology, North-Holland, pp 209–237
- Willms J (2006) Learning divides: ten policy questions about the performance and equity of schools and schooling systems. UIS working paper 5. UNESCO Institute for Statistics, Montreal
- Yitzhaki S (1996) On using linear regressions in welfare economics. *J Bus Econ Stat* 14(4):478–486
- Zimmer B R, Gill, Booker K, Lavertu S, Witte J (2012) Examining charter student achievement effects across seven states. *Econ Educ Rev* 31(2):213–224