Large-scale Assessments in Education

RESEARCH

Open Access

The effect of grade retention on performance in Spanish students: a propensity score matching approach

Jaime León[1*] and Fernando Martínez-Abad[2*]

*Correspondence: jaime.leon@ulpgc.es; fma@usal.es

[1] University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
[2] University Institute of Educational Sciences, University of Salamanca, Salamanca, Spain

Abstract

Background: Grade retention is an educational aspect that concerns teachers, families, and experts. It implies an economic cost for families, as well as a personal cost for the student, who is forced to study one more year. The objective of the study was to evaluate the effect of course repetition on math, science and reading competencies, and math self-efficacy.

Methods: We employed a causal approach using propensity score matching to compare the result in the PISA tests of retained versus non-retained students. We found a comparison group with a similar distribution in the control variables to the group of retained students.

Results: Course retention has a negative effect on the academic performance of students. Retained students showed lower results in math, science, reading, and math self-efficacy compared to the non-retained group.

Conclusions: In line with previous research, evidence shows that grade repetition in Spain is not obtaining the expected results in retained students. This evidence suggests a rethinking of grade retention policies in Spain.

Keywords: Causal inference, Performance, PSM, PISA, Propensity score matching, Grade retention, Secondary education

Introduction

Grade retention is a practice in schools where teachers either require or recommend that students repeat a year of school. This recommendation can stem from various factors, such as poor academic performance, irregular attendance, developmental delays, inadequate social interactions, health issues, and more. (Valbuena et al., 2021). Although, formally, grade retention should be grounded in purely academic and maturational criteria, scientific literature indicates a strong association with contextual factors, including socio-economic, cultural, and demographic influences (López-Rupérez et al., 2021; Nieto-Isidro & Martínez-Abad, 2023). Considering the significant economic, social, and personal costs associated with grade retention (Valbuena et al., 2021), alongside the fact that Spain is one of the European Union and OECD countries with the highest repetition rates (Nieto-Isidro & Martínez-Abad, 2023), it is particularly relevant to study this phenomenon.

Many teachers believe that grade retention can reinforce a student's knowledge or development and support their academic success (Eide & Showalter, 2001; Santos & Monteiro, 2021; Young et al., 2019). However, both international meta-analyses (Allen, 2001; Goos et al., 2021; Jimerson, 2001) and empirical studies (Goos et al., 2013; Tingle et al., 2012) have not consistently demonstrated a clear relationship between grade retention and academic success. In Spain, some studies even suggest that grade retention may have detrimental effects (Jerrim et al., 2022; López-Rupérez et al., 2021; Rodriguez-Rodriguez, 2022).

Researchers aiming to infer causality regarding grade retention face the challenge of identifying a suitable comparison group of students who have not been retained. To address this challenge, researchers rely on causal inference, which involves drawing conclusions about cause and effect based on the observed associations between two or more variables. (Hernán & Robins, 2022; Imbens & Rubin, 2010; Pearl & Mackenzie, 2018). In the context of grade retention, researchers seek to use causal inference to assess the effect of this practice on a student's academic success by comparing outcomes between students who have been held back and those who have not.

One statistical technique that can assist in elucidating the relationship between grade retention and student outcomes in the absence of a traditional control group is matching with propensity scores (Hernán & Robins, 2022). Matching involves creating pairs of observations that are similar on all observed variables, except for the outcome of interest (Rosenbaum & Rubin, 1983). By employing this technique, researchers can construct a control group of non-retained students that can be compared to a group of retained students, thus gaining a better understanding of the impact of grade retention.

Grade retention in Spain

Grade retention in Primary and Secondary education is governed by the Ley Orgánica 3/2020, de 29 de diciembre, por la que se modifica la Ley Orgánica 2/2006, de 3 de mayo, de Educación. Specifically, article 20 regulates grade retention in Primary Education and article 28 in Secondary Education. Specifically, article 20 addresses grade retention in Primary Education, while article 28 pertains to Secondary Education. In the context of Primary Education, students may be retained if the teaching staff, after implementing sufficient ordinary measures, determine that retention will enable the student to achieve the necessary level of competency acquisition. In Secondary Education, students can be retained up to three times, provided that teachers believe this measure supports the acquisition of the competencies required for that stage.

Despite these regulations, grade retention remains a common practice in Spain, with 24.5% of 15 year-old students having been retained at least once, most frequently during their first year of Secondary Education (Ministerio de Educación y Formación Profesional, 2022). This practice is likely sustained by the prevailing belief among teachers that retention is beneficial for improving social and academic outcomes, particularly as a preventive measure against future failures (Range et al., 2012; Santos & Monteiro, 2021; Young et al., 2019). Consequently, while students may be retained based on their competency acquisition in a specific year, many characteristics unrelated to their academic performance increase the likelihood of retention. For example, Jerrim et al. (2022) concluded that grade retention is more common among students born later in the year. ound that students born later in the year are more likely to be retained, a disadvantage linked to their younger age at school entry that diminishes as they progress through the educational system. Similarly, González-Betancor & López-Puig (2016) identified factors such as a student's month of birth, parental education, and employment status as significantly increasing the probability of being retained.

Previous research about grade retention in Spain

Recent studies on grade retention in Spain primarily utilize PISA data to explore two main areas. Some investigate the reasons behind student retention, while others focus on how individual characteristics, such as previous retention, influence academic performance. Among the researchers exploring why students are retained, Jerrim et al. (2022) observed that students born at the end of the year are more likely to repeat a grade. They performed regression discontinuity analysis using the PISA data for students participating in the 2006, 2009, 2012, and 2015 assessments. And concluded that the main reason a younger student has a lower PISA score than an older student in Spain is not relative age per se, but because their knowledge is less as they have been taught less contents. Zinovyeva et al (2014) found that immigrant students perform significantly worse and were at higher risk of grade retention than natives, although their scores improve with the years spent in Spain. Using Oaxaca-Blinder decompositions, they showed that more than half of the gap is explained by individual and family characteristics. Choi et al. (2018) observed that gender and type of household increased the risk of being retained. They used PISA and PIRLS data set to estimate the effect of different variables of nine years old children on the risk of being retained until they were 15 years old. They observed that boys and children without a common household were more prone to repeat an academic year, and concluded that individual and household characteristics are highly relevant for explaining grade repetition in compulsory secondary education. Asensio et al. (2018) observed that being retained was one of the strongest predictors of PISA scores. They explored the effect of different PISA variables on student's performance using decision trees. They concluded that being retained, or more specifically, the academic year a student is in, can effectively summarize the student's performance history, which explains its predictive value. Rodriguez-Rodriguez (2022) also gather evidence of the negative effect of being retained. He assessed grades and motivational variables four times in more than a thousand student during two academic years. He concluded that research like this shows the detrimental effects of being retained and that batch application to poorly performing students subject to repetition has no positive effect, at least for Spanish students. Lopez-Agudo et al. (2023) instead of just comparing retained versus non-retained students, they followed a cohort of over 6000 students who were retained in year 8. They observed no differences in reading competences, that is, one year later retained students had not improved their reading competences and observed slight improvement in math. They also observed that retained students display much lower competences in math and reading. To sum up, researchers gathered evidence that characteristics outside student's control make them susceptible to be retained, and what is worse, that grade retention is one of the strongest factors on the explanation of student's performance.

Effects of grade retained: a causal inference approach

Thus, in this manuscript we wonder what the learning cost of being grade retained is. However, the answer to this question is not straightforward. The raw comparison of retained versus non-retained students would lead to erroneous conclusions. This would happen since there is no random assignment of grade retention, and the decision to retain a student is not independent of the student's academic performance. Researchers can handle this issue by constructing a counterfactual sample of students who are not grade retained, and then compare the academic performance of the actual grade retained students with the performance of the students in the counterfactual sample. By comparing the academic performance of the actual retained students with the performance of the students in the counterfactual sample, researchers can infer causal evidence of the learning cost of being retained.

To infer causality some conditions need to be met, being the most classic that the cause precedes the effect. Thus, if we want to test the effect of retention on future academic performance, obviously, retention must come before the assessment of performance, for example, students might be retained in Primary Education or Secondary Education and assess performance when they are 18 years old. While temporal precedence is a crucial condition for establishing causality, it is not the only one.

Another condition is that the cause could potentially be different, that is, if we had chosen another value for the cause, the effect would have changed. The cause needs to be a treatment that in different conditions could be manipulated by an investigator, recalling to a randomized control trial, where the effect would be a change in outcome, that is, a change in the value according to the treatment received. In "another reality" to be retained or not to retained could be in the researcher hands. In our study, we analyze the potential effects of treatment, in this case, whether being retained impacts student performance and self-efficacy, under the premise that the cause could have been different. For instance, in an ideal experimental setup, such as a randomized controlled trial, the researcher could manipulate retention to observe its effect on outcomes. In this hypothetical scenario, whether a student is retained or not would be under the control of the investigator, allowing for a direct assessment of how different treatments lead to changes in performance. But it would not be possible to infer causality if we want to compare boys and girls, as sex is not amenable to intervention.

In addition to these two conditions, we need pairs of students with similar covariates, one student for the treatment group and another for the control. Ensuring this similarity is vital to meeting the ignorability assumption, as it requires accounting for all potential covariates that could influence the causal relationship. However, as Rosenbaum and Rubin highlight quoting Fisher, this criterion is completely unrealistic. People exhibit a wide range of life experiences, genes, neurons configuration, etc., all of which are incredibly complex. Fortunately, these differences are irrelevant for successful causal inference, as our goal is to gather evidence of the average causal effect of grade retention on performance, which stands for the expected decrease of student's performance in response to being retained. An approximation to this ideal design is to select subjects who are as similar as possible in terms of the propensity to respond to the intervention. Thus, what researchers need is balanced groups in terms of covariates. More specifically, students in both groups need to have similar distributions of covariates. The transition from association to causation, which considers potential biases from unmeasured covariates, is preceded by controlling for observable covariates. Bias will occur from any unobserved or confounding variables that are connected to the treatment assignment. Independent of any confounders, each student must be given the option of being held or not. Students must have the same likelihood of being retained or not with the same variables (Kaplan, 2016).

Since propensity score matching (PSM) is based on this logic, it is essential to identify and control all the key covariates (Rubin, 2008) in the causal relation between the independent variable (assignment to treatment or control) and the dependent variable. Rosenbaum and Rubin (2022) stresses the importance of carefully planning the research design, paying particular attention to the covariates to be included in the data collection. In fact, the main criticism of this technique as a measure of evidence of causal effects is its bias associated with the existence of unobserved or unmeasurable variables. Although there is an open debate in this respect (Guo et al., 2020; Oyenubi, 2020), with some studies showing that this bias is small (Tübbicke, 2023), the evidence suggests that it is possible to minimize the error based on a good prior design (Guo et al., 2020). Researchers using this method attempt to reduce selection bias by making the treatment and control groups more similar to each other. The idea is that if two students have the same propensity score but belong to different groups (treatment or control), both subjects can be considered balanced, and therefore the assignment random (Cordero et al., 2018). The most common statistic used to determine the balance between groups is the standardized mean difference. This statistic is consistent under multiple data conditions and usually is presented in the literature as the most appropriate measure of balance (Ali et al., 2014). The standardized mean difference informs about the effect size of the covariate means difference between groups and can be interpreted as a Cohen's *d* statistic (Austin, 2009). In the context of propensity score matching analysis, standardized mean differences below 0.1 indicate that the covariate is well balanced in groups (Zhang et al., 2019) Along with balancing the main effects of the covariates, Belitser et al. (2011) find that inclusion of their interactions and squares terms returns models with smaller biases.

Present study

The aim of this study is to assess the effect of grade retention on students' academic performance in Spanish schools through causal inference techniques, by providing an analytical approach that contributes to reduce biases and endogeneity issues in its estimation and study. To do so, we use PISA data that includes information on students who were grade retained and those who were not. We match students in the two groups using propensity score matching, which allows us to create a control group that can be compared to a group of retained students.

This study contributes to the literature on grade retention in Spain by providing evidence of the effect of grade retention on academic performance. Addressing this topic is crucial, as few studies in educational journals utilize causal methods, making it challenging to accurately determine the impact of grade retention.

Methods

Participants

The target population of this research are the Spanish students aged 15–16 years old at the time of data collection for the 2022 PISA assessment. Thus, the sample of this study, drawn from the database provided by the OECD (2022), is composed of all those 15–16-year-old students (born between January and December 2006) who participated in the PISA 2022 tests in Spain. The PISA assessments apply stratified two-stage sampling design, sampling schools in the first stage and selecting students by school in the second stage. Since the strata are defined according to region (Autonomous Community) and school ownership, the sample distribution by these two variables is equivalent to the population distribution.

The sample comprised 29,775 students, with 50.42% female and 49.58% male. We addressed missing data in covariates using predcitive mean matching as implemented in MICE (Buuren & Groothuis-Oudshoorn, 2011). This approach allowed us to estimate and fill in missing values, enabling us to utilize the entire dataset for our analysis.

Variables

The independent variable is grade retention. It includes all students who are not in the expected grade for their age and encompasses those who may have been retained one or more times.

The dependent variables of the study are the scores obtained by the students in the mathematics, reading and science competence tests. These variables are defined by the PISA 2022 assessment and analytical framework as follows (OECD, 2023):

- Mathematics: “An individual’s capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain and predict phenomena” (p. 75).
- Science: “Is the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen. A scientifically literate person, therefore, is willing to engage in reasoned discourse about science and technology which requires the competencies of explaining phenomena scientifically […]; evaluating and designing scientific enquiry […]; interpreting data and evidence scientifically” (pp. 100–101).
- Reading: “Is understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one’s goals, to develop one’s knowledge and potential and to participate in society” (p.28).
- Math self-efficacy: Is the belief in their ability to successfully perform specific mathematical tasks. It is measured by asking students how confident they feel about completing various pure and applied mathematics problems.

The selection of observed covariates was based on their theoretical and empirical relationship with grade retention, as identified in studies on school effectiveness in

Spain and relevant literature using PISA assessments. These covariates are known to influence retention decisions and educational trajectories (e.g., Choi & Calero, 2013; Gamazo et al., 2018). Specifically, we used the variables listed in Table 1.

The OECD (2009) highlights that the joint use of plausible values and the replicate weights is the most efficient alternative for the estimation of parameters and standard errors. Thus, we used both plausible values of math, reading and science student competencies and replicate weights of student weights in the regression models computed.

Data analyses

Matching is based on the propensity score (PS), defined as the conditional probability of assignment to treatment, from a group of covariates (Martínez-Abad & León, 2023):

$$PS = p(Z = 1|X_i)$$

where X_i are the covariates included, Z is the assignment of the subject to the treatment (Z=1) or control group (Z=0), this group refers to the non-treated group resulting from the matching procedure.

To facilitate optimal balance between groups, we initially considered full matching, which uses all available data by applying a weight to each subject, allowing any treated student to be matched with one or more control student. In addition, we explored nearest neighbour matching, beginning with a 1:1 ratio where each treated subject is matched with one control subject having the closest propensity score. We also tested a 1:2 ratio, matching each treated subject with two control subjects to broaden the comparison group for each treated individual. To further test the robustness of our results and ensure stability and reliability, we varied the length of the caliper from 0.1 to 0.5. Adjusting the caliper, which defines the maximum allowable difference in propensity scores for matched pairs, allowed us to examine how sensitive our results were to the tightness or looseness of matches between treated and control units. Through these diverse matching options, we aimed to achieve the best balance between treated and control groups.

After matching, we compared univariate distribution of the full versus the matched sample. Next, we collected evidence of reasonable balance by checking all

Table 1 List of control variables used in the study

Variable	PISA database name	Values
Economic, social, and cultural status (ESCS)	ESCS	Continuos
Years of early education	DURECEC	Discrete
Month of birth	ST003D02T	1 (Jan)–12 (Dec)
Gender (*recoded*)	ST004D01T	0: Male 1: Female
Migratory status (*recoded*)	IMMIG	0: Native 1: 1st generation immigrant
Home language (*recoded*)	ST022Q01TA	0: Language of the test 1: Other language

standardized mean differences of covariates and all standardized mean differences of squares and mutual interactions between them.

To estimate the treatment effect and its standard error, we focused on comparing the treated observations, students who experienced grade retention, with non-treated observations included in the matched sample. We achieved this by fitting linear regression models, using each of the three competencies (math, science, and reading) and math self-efficacy as the outcome variables, with retention and covariates serving as predictors. To account for the sampling used in PISA assessment we included the propensity score weights in the estimation procedure. Specifically, we undertook a two-step process for recalculating weights. First, we derived the propensity score matching weights, excluding the initial PISA weights to focus purely on the matching component. In the second step, we calculated new weights for the replicate samples by integrating these adjusted weights with the existing replication structure. This recalibration allowed us to maintain the integrity of both the sampling design and the matching procedure, ensuring accurate and representative results in our analysis.

All analyses were conducted using RStudio. We started our analysis by estimating the propensity scores using the *MatchIt* package (Ho et al., 2011). Once we decided the estimation procedure, we continued with the linear regression models to estimate the effect of grade retained on math, science, reading and math self-efficacy. We used the *Survey* package (Lumley, 2004) to define the structure of weights, replicate weights, and groupings of students within the PISA 2022 database. After developing an independent regression model for each plausible value in each competency, we calculated and reported the averages of the obtained parameters, adhering to the recommendations from the OECD (2022). The complete implemented R code can be found in the Appendix.

Results

Data from 29,775 students were retrieved from the Spanish PISA data. Among these, 5642 (18.95%) had been grade retained, while 24,133 (81.05%) had not. As shown in Table 2, before propensity score matching, retained students showed notable differences in covariates, such as lower ESCS and less early education. They also varied in immigrant status, month of birth, gender, and Spanish language use at home compared

Table 2 Characteristics of the study population before propensity score matching

	Retained	Non-retained
N	5642	24,133
ESCS	− 0.65	0.14
Early education	2.87	2.96
Migratory status	0.12	0.05
Month of birth	7.01	6.45
Gender	0.42	0.52
Home language	0.79	0.85
Math	397.04	495.62
Science	413.48	505.43
Reading	400.27	497.11
Self-efficacy	− 0.60	− 0.10

to non-retained students. These variations underscore the need for propensity score matching to balance these covariates.

Propensity score estimation

Selection of matching estimation approach

Results indicate that all estimation methods (see Table 3), whether nearest neighbor approaches or full matching with varying caliper widths, produce similar outcomes in covariates between retained and control groups. Key factors such as ESCS, early education, immigrant status, month of birth, gender, and the prevalence of Spanish language use demonstrate consistent balance across all methods. Given these similar results, we have decided to rely on the nearest neighbour 1:1 matching approach for our final analysis. This method offers effective covariate balance while maintaining simplicity and efficiency in our matching process.

Covariate balance assessment post-matching

In Fig. 1, we observe the distribution of covariates before and after matching. Post-matching, the distributions for variables such as ESCS, month of birth, gender, and home language align more closely between the retained and control groups. This indicates improved balance, with the matched sample showing a distribution that resembles the retained group more accurately. Notably, the density plots reveal adjustments in the propensity score distribution, particularly for ESCS and month of birth, suggesting that the applied matching has created comparable groups.

After the propensity score matching differences between retained and non-retained students diminished. Standardized differences were less than 0.10 for all variables, suggesting a balance between the two matched groups after propensity score matching (Fig. 2).

Impact of grade retention on academic performance and self-efficacy

As shown in Table 4, retained students demonstrate lower average scores across key competences: math (397.04 vs. 479.57), reading (400.27 vs. 478.55), and science (413.48 vs. 490.87). Additionally, math self-efficacy scores are markedly lower for retained students (− 0.59 vs. − 0.21), indicating diminished confidence in their

Table 3 Comparison of covariate balance across different matching procedures and calipers

	Full	Nearest						
		Ratio 1:1	**Cal. 0.1**	**Cal. 0.2**	**Cal. 0.3**	**Cal. 0.4**	**Cal. 0.5**	**Ratio 1:2**
N	24,133	5642	5519	5565	5609	5642	18,168	11,284
ESCS	− 0.70	− 0.67	− 0.65	− 0.66	− 0.66	− 0.67	− 0.67	− 0.67
Early education	2.86	2.85	2.86	2.86	2.85	2.85	2.85	2.85
Migratory status	0.13	0.12	0.11	0.11	0.11	0.12	0.12	0.12
Month of birth	7.01	7.03	7.03	7.03	7.03	7.03	7.03	7.03
Gender	0.41	0.42	0.42	0.42	0.42	0.42	0.42	0.42
Home language	0.21	0.20	0.20	0.20	0.20	0.20	0.20	0.20

Cal. = caliper

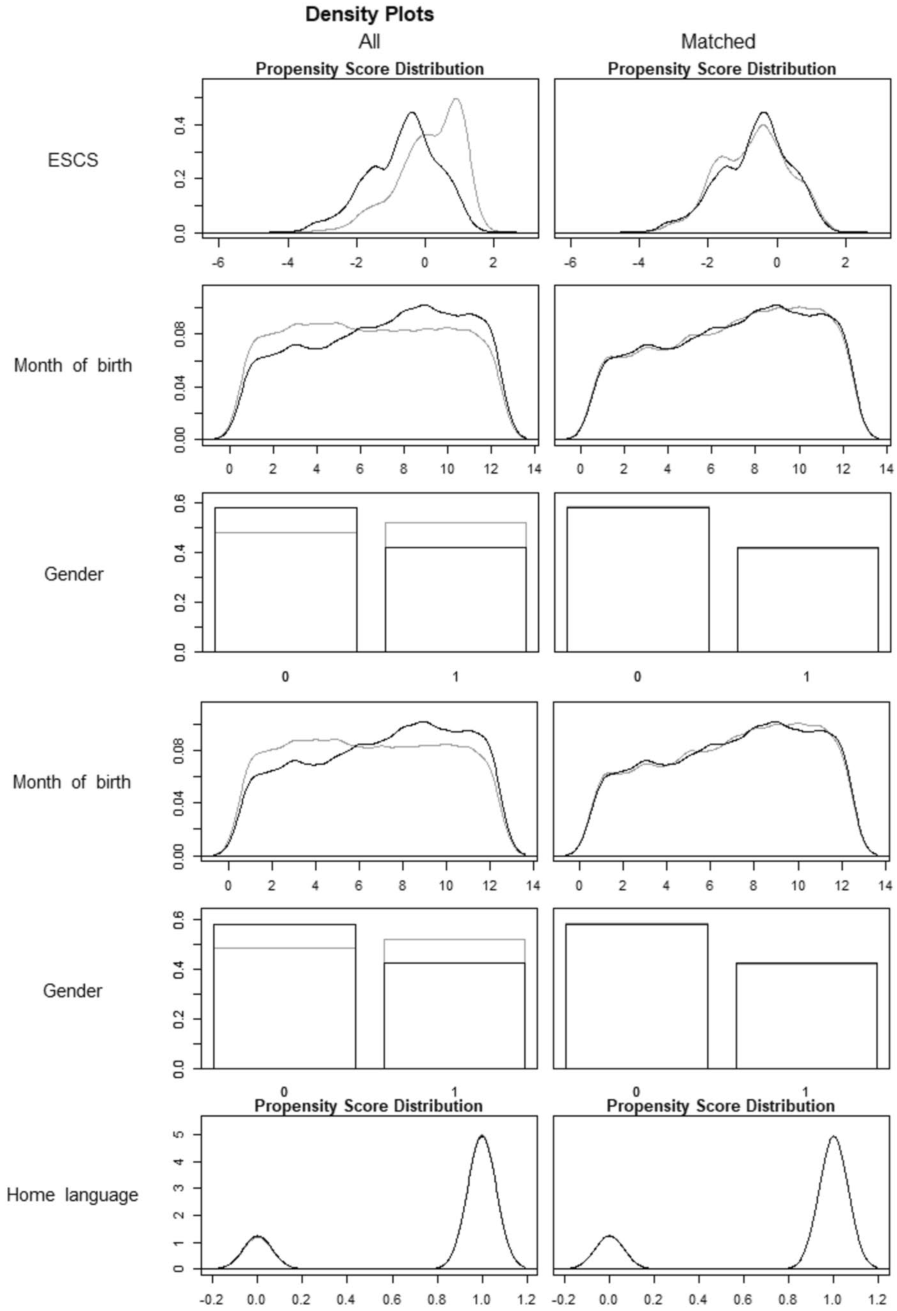

Fig. 1 Distributions of retained versus non retained before and after the matching. Retained students are in dark lines. Non-retained students are in grey lines

mathematical abilities. These findings underscore relevant disparities in academic achievement and self-efficacy, highlighting the challenges faced by retained students.

These results illustrate the differences between retained students and the non-retain counterparts on PISA competencies and math self-efficacy. The results show a negative effect on all measured outcomes. Retained students performed considerably lower than their non-retained peers. Math scores were 82.77 points lower, reading

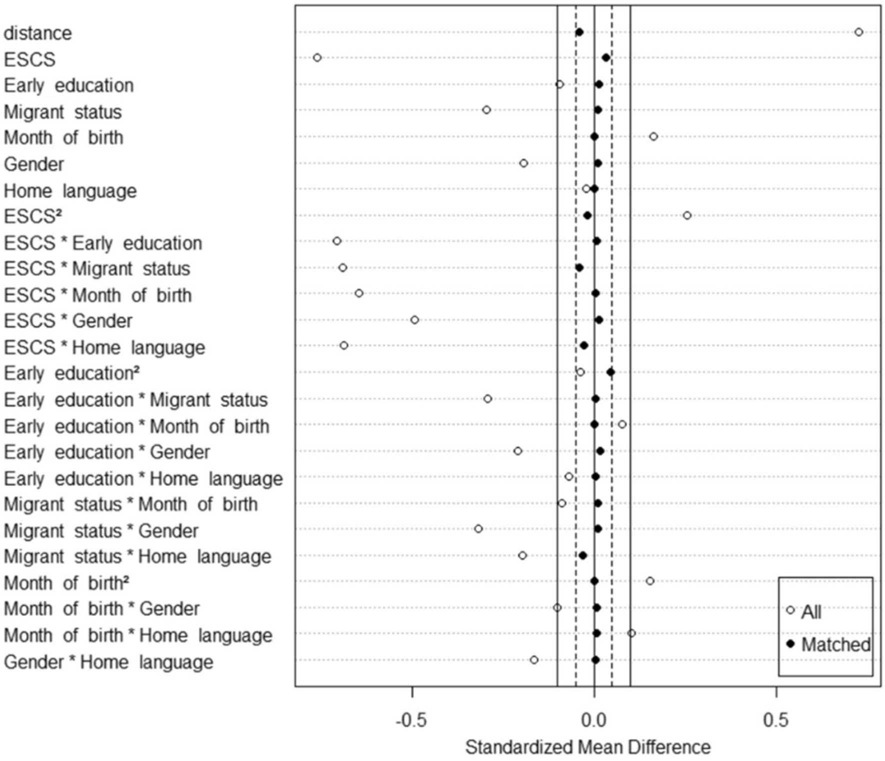

Fig. 2 Standardized mean difference of covariates and interactions

Table 4 Comparion of retained and non retained groups

		Retained	Non retained
	N	5642	5642
Outcomes	Math	397.04	479.81
	Read	413.48	490.72
	Science	400.27	479.19
	Math self-efficacy	− 0.59	− 0.20
Covariates	ESCS	− 0.65	− 0.66
	Early childhood education	2.88	2.85
	Migrant status	0.12	0.11
	Month of birth	7.01	7.03
	Gender	0.42	0.41
	Home language	0.20	0.20

scores 77.24 points lower, and science scores 78.92 points lower. Math self-efficacy was also lower by 0.39 points.

These differences are further highlighted by the effect sizes presented in Table 5. Retained students show substantial negative effect sizes across all outcomes, with math at − 1.18, indicating a large impact. Science and reading also exhibit relevant negative effects, with effect sizes of − 0.98 and − 0.95, respectively. Math self-efficacy, although smaller, still shows a notable effect size of − 0.37. These effect sizes

Table 5 Cohen's d Effect sizes

Outcomes	Effect Size
Math	− 1.18
Science	− 0.98
Reading	− 0.95
Self-Efficacy	− 0.37

underscore the profound impact of grade retention on both academic performance and self-confidence in mathematical abilities.

The regression analysis shows that grade retention has a significant negative impact on academic outcomes and math self-efficacy, even when accounting for covariates (see Table 6). Retained students score markedly lower in math ($\beta = -82.57$, $p < 0.001$), science ($\beta = -77.26$, $p < 0.001$), and reading ($\beta = -78.69$, $p < 0.001$), and they exhibit reduced math self-efficacy ($\beta = -0.39$, $p < 0.001$). These outcomes highlight the substantial challenges faced by retained students, indicating that the negative effects of retention persist even after considering various demographic and socio-economic factors.

Discussion

The aim of this study was to assess the effect of grade retention on students' academic performance in Spanish schools. To do so, we used PISA data that included information on students who were grade retained and those who were not. We matched students in the two groups using propensity score matching, which allowed us to create a control group that can be compared to a group of retained students.

Several leading researchers in the field of applied quantitative research methods in Social Sciences recommend the use of causal inference techniques to carry out analyses with large-scale educational assessment databases (e.g., Cordero et al., 2018; Kaplan, 2016; Rutkowski & Delandshere, 2016). Such is the case of PISA, PIRLS or TIMSS, for example. In fact, there are some studies in the literature that propose the use of propensity score matching to study the causal effects of diverse educational factors on achievement, such as ICT use (Agasisti et al., 2020), immigrant status (Arikan et al., 2020), school climate (Rizzotto & França, 2022) or school ownership (Crespo-Cebada et al., 2014). However, there are very few studies that propose the use of this technique to analyze the efficiency of retention based on large-scale assessments (Ehmke et al., 2017; Goos et al., 2013).

The results of this study showed the possibility of finding a comparison group of retained students in terms of ESCS, early education, month of birth, gender, immigrant status, and the language spoke at home. These two groups showed significant differences in math, science, reading, and self-efficacy, suggesting that grade retention has a negative effect on academic outcomes. Students who were grade retained had significantly lower performance in PISA compared to their non-retained counterparts.

Our findings align with the existing literature on grade retention in Spain, reinforcing the evidence that grade retention may not be an effective educational practice and may have significant negative consequences for students. By employing propensity score matching, our study contributes to gather additional evidence of the impact of grade

Table 6 Regression Analysis of Retention and Covariates on Academic Outcomes and Math Self-Efficacy

Outcomes	Covariate	Beta	Std.Error	df	*T* value	*p*	CI.Lower	CI.Upper
Math	Intercept	498.59	2.94	24.31	169.48	< 0.001	492.52	504.66
	Retention	− 82.57	1.72	24.68	− 47.99	< 0.001	− 86.12	− 79.03
	ESCS	14.79	0.82	28.67	18	< 0.001	13.11	16.47
	Early education	2.26	0.68	33.97	3.33	< 0.001	0.88	3.63
	Migratory status	− 11.35	2.41	52.94	− 4.71	< 0.001	− 16.19	− 6.52
	Month of birth	− 0.73	0.23	19.3	− 3.13	0.01	− 1.22	− 0.24
	Gender	− 18.77	1.74	22.89	− 10.78	< 0.001	− 22.37	− 15.16
	Home language	− 7.09	2.09	74.52	− 3.38	< 0.001	− 11.26	− 2.91
Science	Intercept	509.73	3.6	23.25	141.52	< 0.001	502.29	517.18
	Retention	− 77.26	2.43	16.54	− 31.76	< 0.001	− 82.41	− 72.12
	ESCS	13.49	0.80	80.69	16.89	< 0.001	11.9	15.08
	Early education	2.58	0.70	40.46	3.70	< 0.001	1.17	3.99
	Migratory status	− 9.96	3.56	21.18	− 2.79	< 0.01	− 17.37	− 2.55
	Month of birth	− 1.00	0.28	22.08	− 3.61	< 0.001	− 1.58	− 0.43
	Gender	− 14.58	1.99	20.67	− 7.33	< 0.001	− 18.73	− 10.44
	Home language	− 16.33	2.8	28.95	− 5.84	< 0.001	− 22.05	− 10.6
Reading	Intercept	491.78	3.36	50.56	146.44	< 0.001	485.04	498.52
	Retention	− 78.69	2.59	16.71	− 30.33	< 0.001	− 84.17	− 73.21
	ESCS	14.83	1.03	28.93	14.44	<0.001	12.73	16.93
	Early education	1.09	0.75	141.91	1.45	0.15	− 0.39	2.57
	Migratory status	− 10.42	3.34	30.76	− 3.12	< 0.001	− 17.22	− 3.61
	Month of birth	− 1.05	0.36	14.41	− 2.92	< 0.01	− 1.82	− 0.28
	Gender	15.16	2.02	23.24	7.5	< 0.001	10.99	19.34
	Home language	− 20.56	3.21	27.19	− 6.4	< 0.001	− 27.15	− 13.96
Self-efficacy	Intercept	− 0.11	0.03	72	− 3.95	< 0.001	− 0.17	− 0.06
	Retention	− 0.39	0.01	72	− 26.83	< 0.001	− 0.42	− 0.36
	ESCS	0.12	0.01	72	15.37	< 0.001	0.11	0.14
	Early education	0.03	0.01	72	3.87	< 0.001	0.01	0.04
	Migratory status	0.05	0.02	72	2.52	<0.01	0.01	0.10
	Month of birth	0.00	0.00	72	− 0.87	0.39	− 0.01	0.01
	Gender	− 0.26	0.02	72	− 16.15	< 0.001	− 0.29	− 0.23
	Home language	0.12	0.02	72	6.08	< 0.001	0.08	0.15

retention compared to previous studies that primarily relied on non-causal focused methods (e.g., Resino et al., 2019). Specifically, the large negative effect sizes observed in math (− 1.18), science (− 0.98), and reading (− 0.95), as well as the moderate effect on self-efficacy (− 0.37), suggest that retained students are at a pronounced disadvantage relative to their non-retained peers. These results are consistent with prior research, such as that by Jerrim et al. (2022), López-Rupérez et al. (2021), and Rodriguez-Rodriguez (2022), which have highlighted the detrimental academic outcomes associated with grade retention. Our study expands on these findings by providing evidence via propensity scores, suggesting the need to reconsider the efficacy of grade retention as a policy intervention. This consistency across different methodological approaches strengthens the importance of exploring alternative strategies to support students at risk of academic underperformance.

Limitations and future perspective

Several limitations and future perspectives can be highlighted. One key limitation relates to the age of students in our analysis. Our research focuses on same-age comparisons, which provide a more appropriate counterfactual for assessing the efficacy of retention (Valbuena et al., 2021). By comparing retained students to their same-age peers in higher grades, we observed that retained students may be relatively disadvantaged, as they lack exposure to the same material covered by their non-retained counterparts. This misalignment raises concerns about whether retention truly provides students with the foundation they need to meet required competencies, as intended. Furthermore, the diminished self-efficacy observed in retained students suggests that the experience of retention may exacerbate existing challenges rather than resolve them. This issue may be particularly relevant for students who have been retained in multiple grades, whereas the present study has treated grade retention as a dichotomous variable. While this methodological decision has allowed for a more concise and interpretable analysis, potential biases associated with it should be further examined in future research.

Finally, it is important to note that this study is based on a secondary analysis of PISA 2022 data. This large-scale assessment has faced significant criticism, particularly regarding its non-curricular approach to measuring performance and the quality of the scales used to construct composite variables (e.g., Math Self-Efficacy, ESCS). Therefore, we should be careful with the results obtained, seeking in future studies to replicate them from other databases.

Future research could explore how teachers and schools make decisions about grade retention, particularly whether these choices are driven by academic performance or influenced by factors such as behavior, health, or subjective perceptions. Understanding how school resources, teacher training, or regional policies impact retention practices could reveal patterns or biases in how retention is applied. Combining this analysis with studies on the long-term academic and non-academic outcomes of retention would provide a more comprehensive view of its effects and help identify ways to make policies more consistent, equitable, and effective.

In addition, when using a propensity score model researchers need to control for aspects that had occurred at the moment of the treatment, but not the outcome or factors that have been affected by the treatment. In our case, it would be the number of school changes, scientific experience outside the school the time spent in class, the time spent in homework, or the reason to miss the class, which are variables that we did not have access to, and which could be affected by retention.

One limitation, not only of our study, but of propensity score matching is the lack of testing for indicators of endogeneity, such as self-selection or reversed causality (Cordero et al., 2018). A consequence of this is an overestimation of the negative effect of grade retention on academic performance. For example, our results might be biased if a large proportion of parents choose schools for their children attending to PISA scores. Fortunately, this is not the case in Spain, where most children go to a public school.

Conclusion

The findings of this study carry significant implications for educational policy and pedagogical practice in Spain and potentially in similar contexts with high rates of grade retention. Firstly, given the evidence suggesting the negative effects of repeating a grade on academic performance, policymakers should reconsider the use of grade retention as the main strategy to remedy poor academic achievement. Alternatives such as intensive academic support programs, personalized tutoring, and differentiated teaching strategies might offer more effective means to address the needs of at-risk students without incurring the social and emotional costs associated with grade retention.

In the classroom, educators and school administrators could use these findings to develop early interventions targeted at students exhibiting signs of academic or social difficulties. Such interventions could include closer monitoring of student progress, promoting parental involvement, and applying inclusive teaching techniques that address the diverse learning needs of students.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-025-00243-0.

Supplementary Material 1.

Author contributions

J.L: Conceptualization, methodology, supervision, project administration, writing–original draft. F.M.A: Software, formal analysis, funding acquisition, visualization, writing–review and editing.

Funding

Grant PID2021-125775NB-I00 funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A Way of Making Europe.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 26 January 2025 Accepted: 4 March 2025
Published online: 19 March 2025

References

Agasisti, T., Gil-Izquierdo, M., & Han, S. W. (2020). ICT Use at home for school-related tasks: what is the effect on a student's achievement? Empirical evidence from OECD PISA data. *Education Economics, 28*(6), 601–620. https://doi.org/10.1080/09645292.2020.1822787

Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S., Roes, V., Hoes, K. C. B., de Boer, A. W., & Klungel, A. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology Drug Safety, 23*(8), 802–811. https://doi.org/10.1002/pds.3574

Allen, J. D. (2001). Grades as valid veasures of academic achievement of classroom learning. *The Clearing House: A Journal of Educational Strategies, 78*(5), 218–223.

Arikan, S., van de Vijver, F. J. R., & Yagmur, K. (2020). Mainstream and immigrant students' primary school mathematics achievement differences in European countries. *European Journal of Psychology of Education, 35*(4), 819–837. https://doi.org/10.1007/s10212-019-00456-2

Asensio, I., Carpintero, E., Expósito, E., & López, E. (2018). Cuánto oro hay entre la arena? Minería de datos con los resultados de España en PISA 2015. *Revista Española de Pedagogía, 76*(270), 225–245. https://doi.org/10.2550/REP76-2-2018-02

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Belitser, S., Martens, E. P., Pestman, W. R., Groenwold, R. H. H., de Boer, A., & Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety, 20*(11), 1115–1129. https://doi.org/10.1002/pds.2188

Choi, A., & Calero, J. (2013). Determinants of the risk of school failure in Spain in PISA-2009 and proposals for reform. *Revista de Educación, 362*, 562-593. https://doi.org/10.4438/1988-592X-RE-2013-362-242

Choi, A., Gil, M., Mediavilla, M., & Valbuena, J. (2018). Predictors and effects of grade repetition. *Revista de Economía Mundial, 48*, 21–42.

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys, 32*(3), 878–915. https://doi.org/10.1111/JOES.12217

Crespo-Cebada, E., Pedraja-Chaparro, F., & Santín, D. (2014). Does school ownership matter? An unbiased efficiency comparison for regions of Spain. *Journal of Productivity Analysis, 41*(1), 153–172. https://doi.org/10.1007/s11123-013-0338-y

Ehmke, T., Sälzer, C., Pietsch, M., Drechsel, B., & Müller, K. (2017). Competence development in the school year after PISA 2012: Effects of grade retention. *Zeitschrift Für Erziehungswissenschaft, 2017*(20), 2. https://doi.org/10.1007/S11618-017-0752-4

Eide, E. R., & Showalter, M. H. (2001). The effect of grade retention on educational and labor market outcomes. *Economics of Education Review*. https://doi.org/10.1016/S0272-7757(00)00041-8

Gamazo, A., Martínez-Abad, F., Olmos-Migueláñez, S., & Rodríguez-Conde, M. J. (2018). Assessment of factors related to school effectiveness in PISA 2015. A multilevel analysis. *Revista de Educación, 379*, 56-84. https://doi.org/10.4438/1988-592X-RE-2017-379-369

González-Betancor, S. M., & López-Puig, A. J. (2016). Grade retention in primary education is associated with quarter of birth and socioeconomic status. *PLoS ONE, 11*(11), e0166431. https://doi.org/10.1371/journal.pone.0166431

Goos, M., Pipa, J., & Peixoto, F. (2021). Effectiveness of grade retention: A systematic review and meta-analysis. *Educational Research Review*. https://doi.org/10.1016/j.edurev.2021.100401

Goos, M., van Damme, J., Onghena, P., Petry, K., & de Bilde, J. (2013). First-grade retention in the Flemish educational context: effects on children's academic growth, psychosocial growth, and school career throughout primary education. *Journal of School Psychology, 51*(3), 323–347. https://doi.org/10.1016/J.JSP.2013.03.002

Guo, S., Fraser, M., & Chen, Q. (2020). Propensity score analysis: Recent debate and discussion. *Journal of the Society for Social Work and Research, 11*(3), 463-482. https://doi.org/10.1086/711393

Hernán, M. A., & Robins, J. M. (2022). *Causal inference: what if chapman*. Hall/CRC.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal onference. *Journal of Statistical Software, 42*(8), 1–28. https://doi.org/10.1637/jss.v042.i08

Imbens, G. W., & Rubin, D. B. (2010). Rubin causal model. In S. N. Durlauf & L. E. Blume (Eds.), *Microeconometrics*. Palgrave Macmillan.

Jerrim, J., Lopez-Agudo, L. A., & Marcenaro-Gutierrez, O. D. (2022). Grade retention and school entry age in Spain: a structural problem. *Educational Assessment, Evaluation and Accountability, 34*(3), 331–359. https://doi.org/10.1007/S11092-021-09375-7

Jimerson, S. R. (2001). Meta-analysis of grade retention research: implications for practice in the 21st century. *School Psychology Review, 30*, 420–437.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: a review and synthesis. *Large-Scale Assessments in Education*. https://doi.org/10.1186/s40536-016-0022-6

Lopez-Agudo, L. A., Latorre, C. P., & Marcenaro-Gutierrez, O. D. (2023). Grade retention in Spain: the right way? *Education and Assessment Evaluation Accountability*. https://doi.org/10.1007/s11092-023-09421-6

López-Rupérez, F., García-García, I., & Expósito-Casas, E. (2021). La repetición de curso y la graduación en Educación Secundaria Obligatoria en España Análisis empíricos y recomendaciones políticas. *Revista de Educación, 394*, 18.

Lumley, T. (2004). Analysis of complex survey samples. *Journal Statistical Software, 9*(8), 1–19. https://doi.org/10.1637/jss.v009.i08

Martínez-Abad, F., & León, J. (2023). Causal inference in educational research: Causal analysis in cross-sectional observational studies. *RELIEVE - Electronic Journal of Educational Research and Evaluation,* 29(2). https://doi.org/10.30827/relieve.v29i2.26843

Ministerio de Educación y Formación Profesional. (2022). *Las cifras de la educación en España: estadísticas e indicadores*. Secretaría General Técnica.

Nieto-Isidro, S., & Martínez-Abad, F. (2023). Grade Retention and its relationship with socioeconomic andeducative variables in Spain. *Revista de Educación,* 402, 195–220. https://doi.org/10.4438/1988-592X-RE-2023-402-600

OECD. (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/19963777

OECD. (2023). *PISA 2022 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

Oyenubi, A. (2020). A note on covariate balancing propensity score and instrument-like variables. *Economics Bulletin, 40*(1).

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Range, B. G., Pijanowski, J., Holt, C. R., & Young, S. (2012). The perceptions of primary grade teachers and elementary principals about the effectiveness of grade-level retention. *Professional Educator, 36*(1), 1–16.

Resino, D. A., Amores, I. A. C., & Muñoz, I. A. (2019). La repetición de curso a debate: Un estudio empírico a partir de PISA 2015. *Educación XX1, 22*(2), 69–92. https://doi.org/10.5944/EDUCXX1.22479

Rizzotto, J. S., & França, M. T. A. (2022). Indiscipline: The school climate of Brazilian schools and the impact on student performance. *International Journal of Educational Development*. https://doi.org/10.1016/j.ijedudev.2022.102657

Rodriguez-Rodriguez, D. (2022). Grade retention, academic performance and motivational variables in compulsory secondary education: A longitudinal study. *Psicothema, 34*(3), 429–436.

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

Rosenbaum, P. R., & Rubin, D. P. (2022). Propensity scores in the design of observational studies for causal effects. *Biometrika*. https://doi.org/10.1093/biomet/asac054

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2*(3), 808–840. https://doi.org/10.1214/08-AOAS187

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education, 4*(1), 6. https://doi.org/10.1186/s40536-016-0019-1

Santos, N. N., & Monteiro, V. C. (2021). Crenças de professores e futuros professores portugueses sobre a reprovação no 2° ano de escolaridade. *Revista Brasileira de Educação*. https://doi.org/10.1590/s1413-24782021260068

Tingle, L. R., Schoeneberger, J., & Algozzine, B. (2012). Does grade retention make a difference? *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 85*(5), 179–185. https://doi.org/10.1080/00098655.2012.679325

Tübbicke, S. (2023). How sensitive are matching estimates of active labor market policy effects to typically unobserved confounders? *Journal for Labour Market Research, 57*(1). https://doi.org/10.1186/s12651-023-00352-9

Valbuena, J., Mediavilla, M., Choi, Á., & Gil, M. (2021). Effects of grade retention policies: A literature review of empirical studies applying causal inference. *Journal of Economic Surveys, 35*(2), 408–451. https://doi.org/10.1111/JOES.12406

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice multivariate imputation by chained equations in R. *Journal Statistical Software*. https://doi.org/10.1637/jss.v045.i03

Young, S., Trujillo, N. P., Bruce, M. A., Pollard, T., Jones, J., & Range, B. (2019). Preservice teachers' views about grade retention as an intervention for struggling students. *Preventing School Failure: Alternative Education for Children and Youth, 63*(2), 113–120. https://doi.org/10.1080/1045988X.2018.1523124

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*. https://doi.org/10.21037/atm.2018.12.10

Zinovyeva, N., Felgueroso, F., & Vazquez, P. (2014). Immigration and student achievement in Spain: Evidence from PISA. *Series, 5*(1), 25–60. https://doi.org/10.1007/s13209-013-0101-7

Publisher's Note