Received: 13 February 2021 | Accepted: 16 December 2021

DOI: 10.1002/berj.3774

ORIGINAL ARTICLE

BERJ BERA

The impact of test language on PISA scores. New evidence from Wales

John Jerrim[1] | Luis Alejandro Lopez-Agudo[2] | Oscar David Marcenaro-Gutierrez[2]

[1]Department of Social Science, UCL Institute of Education, University College London, London, UK

[2]Departamento de Economía Aplicada (Estadística y Econometría), Facultad de Ciencias Económicas y Empresariales, Universidad de Málaga, Málaga, Spain

Correspondence
Luis Alejandro Lopez-Agudo, Departamento de Economía Aplicada (Estadística y Econometría), Facultad de Ciencias Económicas y Empresariales, Universidad de Málaga, Plaza de El Ejido s/n, 29013, Málaga, Spain.
Email: lopezagudo@uma.es

Funding information
FEDER, Grant/Award Number: PY20-00228; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2020-119471RB-I00; Andalusian Regional Government, Grant/Award Number: SEJ-645

Abstract

In this paper we investigate the link between the language in which pupils take the Programme for International Student Assessment (PISA) test and the scores they achieve in this assessment, focusing upon the case of Wales. Using five rounds of PISA data and an instrumental variable approach, we show how pupils who took the test in Welsh score around 0.3 standard deviations (30 PISA test points) lower in reading, mathematics and science than their peers who took the test in English. This finding is robust to different model specifications and statistical approaches. We argue that this may indicate that the academic performance of teenagers in Wales may be underestimated in PISA – particularly amongst those who take this test in Welsh.

KEYWORDS

home language, PISA, test language, Wales

INTRODUCTION

The Programme for International Student Assessment (PISA) is a study designed to measure 15-year-olds' ability in reading, mathematics and science. According to the PISA technical reports (PISA technical standard 2.1[1]), students should take the test in the language in which they are most comfortable. This is so that their scores on the test reflect their actual skill in the subject(s) being assessed and are not unduly influenced by pupils having limited language skills[2] (OECD, 2012, pp. 369–370). In doing so, the Organisation for Economic Co-operation and Development (OECD) implicitly acknowledges how a

Key insights

What is the main issue that the paper addresses?

We investigate the link between the language in which pupils take the PISA test and the scores they achieve in this assessment, focusing upon the case of Wales.

What are the main insights that the paper provides?

We find that pupils in Wales who took the test in Welsh score around 0.3 standard deviations lower in reading, mathematics and science than their peers who took the test in English.

difference between pupil's "home language" (i.e. the language they speak outside of school, such as with their friends and family) and their "test language" may have an impact upon the results.

Kennedy and Park (1994) have studied the link between test language, home language and academic achievement in the context of middle-school Asian-American and Mexican pupils in the United States. They found that those students who did not speak English at home obtained lower reading test scores. Similarly, in an analysis of PISA 2000 data for Australia, De Bortoli and Cresswell (2004) found that pupils who took the test in a language they did not regularly speak at home achieved lower scores overall. The work by Mancilla-Martinez and Lesaux (2011) in the United States found that students whose home language was Spanish – but who took a test in English – tended to achieve lower test scores than native English speakers.

This issue is also prominent in international public policy and political debates. In many countries the existence of multilingualism has been used as a symbol of nationalism by political parties. That is the case for French in Canada, Catalan in Spain and Welsh in Wales. In the present study, we focus on the latter.

In this context, a distinctive characteristic of the Welsh education system is the existence of schools that use Welsh as the primary language of instruction (Johnes, 2020). Concretely, in Wales, there are different types of school, which vary in their use of English and/or Welsh in the classroom. According to the Welsh Assembly Government (2007): "a school is Welsh-speaking if more than one half of the following subjects are taught (wholly or partly) in Welsh: (a) religious education, and (b) the subjects other than English and Welsh which are foundation subjects in relation to pupils at the school" (Parliament of the United Kingdom, 2002, 105(7)). In particular, there are four categories of secondary school (Jones, 2016; Welsh Assembly Government, 2007):

a. Welsh-medium secondary schools. Schools where all subjects (apart from English) are taught in Welsh, with this language used for everything throughout the school.
b. Bilingual secondary schools. These schools use a combination of Welsh and English. There are four sub-groups within this category, which differ in the percentage of subjects taught in Welsh and whether they are also offered in English at the same time. These sub-groups are: 2A (at least 80% of subjects, apart from English and Welsh, are taught in Welsh), 2B (at least 80% of subjects, except English and Welsh, are taught in Welsh, but also in English), 2C (50–79% of subjects, excluding English and Welsh, are taught in Welsh, but also in English) and 2CH (all subjects, apart from English and Welsh, are taught in both languages). Both languages are used for communication in these schools, but priority is given to Welsh.

c. Predominantly English-medium secondary schools with significant use of Welsh. In these schools English and Welsh are used in teaching (with 20–40% of subjects in Welsh) and both languages are used for communication in the school, giving priority to Welsh.
d. Predominantly English-medium secondary schools. In these schools most subjects are in English, and only one or two subjects are optionally offered in Welsh. English is the predominant language for communication, but some Welsh is also used.

The distribution of these schools across different areas within Wales can be found in Table 1. This highlights how around 72% of secondary schools in Wales are English-medium only, although there is significant regional variation. For instance, whereas 90% of schools are English medium in south-east Wales, this falls to around 52% in the north, where Welsh-medium (or Bilingual with a strong emphasis on Welsh) education is much more prevalent.

Interestingly, previous research has suggested that the socio-economic composition of English-medium, Welsh-medium and Bilingual schools may also differ. For instance, Van den Brande et al. (2019) illustrate how the average rate of disadvantaged pupils in English-medium schools is 21%, which is notably higher than in Bilingual (14%) and Welsh-medium (10%) schools. According to these authors, “Welsh medium schools have frequently been supposed to be ‘better’ than English-medium schools, and thus may attract parents with cultural capital and particular aspirations for their children, regardless of their own linguistic background” (p. 49). Jones (2017b) also noted how Welsh-medium schools have a reputation of being highly successful, which is likely to be attractive for upper- and middle-class parents. In addition, Welsh-medium schools attract a profile of higher socio-economic status students also owing to the bilingual education that they offer, a feature which has been found to be beneficial to students’ learning (Edwards & Newcombe, 2006; Jones, 2017a). Yet the existing literature (e.g. Jerrim & Shure, 2016) has also found average reading and science scores to be lower amongst pupils attending Welsh-medium schools than amongst pupils attending English-medium schools.

This has led most existing work in this area to focus upon differences in academic performance between Welsh-medium and English-medium schools. For instance, Gorard (1998) shows that, once differences in local-area characteristics are taken into account, there is no significant difference between the performance of Welsh-medium and English-medium schools in Wales. However, to our knowledge, there is no evidence as to whether taking the PISA test in Welsh might be detrimental for the scores obtained by Welsh pupils (compared with the alternative of taking the test in English). As we will see, in spite of the PISA technical standard 2.1, Welsh students may not be freely choosing the language in which they take the PISA test (according to Sizmur et al., 2019, p. 199, students take the PISA test in the school’s language of instruction in Wales). Specifically, we aim to address the following research question:

TABLE 1 The distribution of secondary school types across Wales by area in 2019/2020

	Welsh medium	Bilingual – 2A	Bilingual – 2B/2C	English with significant Welsh	English medium	Total
North Wales	5	10	7	2	26	50
South-west and mid Wales	4	4	6	5	32	51
Central south Wales	6	—	—	—	44	50
South-east Wales	3	—	—	—	29	32
All schools	18	14	13	7	131	183

Notes: There were no bilingual 2CH schools in Wales in 2019/2020.

Source: Retrieved from Welsh Government StatsWales, https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Schools-Census/Pupil-Level-Annual-School-Census/Schools/schools-by-localauthorityregion-welshmediumtype

Does taking the test in Welsh reduce students' PISA test scores?

To do this, we use five cycles (2006–2018) of PISA data for Wales. These data allow us to identify both the language spoken by pupils at home and the language in which they took the PISA test (students in Wales took the test in either English or Welsh). One novel aspect is that we attempt to move a step closer towards estimating a causal effect by implementing an instrumental variable approach.

The paper now proceeds as follows: Section 2 provides a description of the data, followed by an overview of the methodological approach in Section 3. Results are reported in Section 4, followed by discussion and conclusions in Section 5.

DATA

Data description

PISA is a test taken by 15-year-old pupils from 80 countries. It aims to assess their skills in reading, mathematics and science and has been conducted every three years since 2000. Participating students also complete a background questionnaire, while headteachers complete a school questionnaire. We analyse those PISA cycles in which both test and home languages are known for Wales: 2006, 2009, 2012, 2015 and 2018. These cycles have been chosen because (a) they are the most recent ones; (b) an oversample was drawn for Wales to facilitate national reporting; and (c) information on test and home languages has been collected.

The test language variable has two options: "English" and "Welsh". The home language variable indicates the language spoken most often at home by the pupil, with five options: "English", "Welsh", "Irish", "Ulster Scots" and "Other languages". Other relevant variables for our analysis are pupil's, father's and mother's region of birth, which includes "Germany", "India", "Ireland", "Pakistan", "Poland", "United Kingdom (England)" and "Other countries".

Descriptive analyses

Table 2 illustrates the percentage of students who took the PISA test in each language (English or Welsh) and the percentage of students who spoke each language at home. The percentage of students who took the test in English is stable across PISA cycles (87%), along with the percentage who speak each language at home (91% for English, 6% for Welsh, 0.1% for Irish, 0.1% for Ulster Scots and 3% for other languages). These figures are broadly similar to those reported by StatsWales (2021).

Table 3 illustrates how almost all pupils who took the PISA test in English also spoke English regularly at home. The situation is rather different for those who took the PISA test in Welsh, of whom more than half spoke English regularly at home. In other words, many pupils who usually speak English at home end up taking the test in Welsh. This might be due to English-speaking parents enrolling their children in Welsh or bilingual schools because of the bilingual education that they offer – a feature which may have benefits for students' learning (Edwards & Newcombe, 2006; Jones, 2017a). This would also be consistent with standard policy in many Welsh-medium schools (as previously described), where Welsh language is preferred. Indeed, Sizmur et al. (2019, p. 199) note in the official Welsh Government PISA 2018 report how "pupils in Wales were assigned assessments and questionnaires according to the relevant language of instruction".[3] This suggests that students in Welsh-medium schools were not offered a choice of test language, but were forced (or

TABLE 2 Percentage of students who took the test in each language and who spoke each language at home in Wales

	Language				
	English	Welsh	Irish	Ulster Scots	Other language
Language of the test					
2006	87	13	—	—	—
2009	87	13	—	—	—
2012	87	13	—	—	—
2015	90	10	—	—	—
2018	86	14	—	—	—
All cycles	87	13	—	—	—
Language at home					
2006	92	7	<0.1	—	1
2009	91	6	<0.1	<0.1	2
2012	92	6	<0.1	<0.1	2
2015	90	6	0.2	0.2	4
2018	89	6	<0.1	0.1	5
All cycles	91	6	0.1	<0.1	3

Notes: All OECD recommended practices (final student weights and balanced repeated replication (BRR) weights) have been employed. The "—" indicates that there are no data for that region in the PISA cycle.

Source: authors' own calculations.

strongly encouraged) to take the test in Welsh. Yet this would be a violation of PISA technical standard 2.1 – which stipulates that pupils should take the test in the language that they are most comfortable with. In fact, from a total of 607 school-by-PISA-cycle observations in our data, all pupils took the test in English in 503, in 44 all pupils took the test in Welsh and in 60 there was a mix of English and Welsh test takers. The fact that in some schools all of the pupils took the test in Welsh suggests that students may not have had complete freedom in choosing the language to take the test.

Table 4 compares the background characteristics of pupils who took the PISA test in English and Welsh. Those who took the test in Welsh tend to come from more advantaged socio-economic backgrounds than those who took it in English (a difference which is statistically significant at convention levels). For instance, Welsh-language test takers were significantly more likely to be in the top socio-economic status quartile and to have a mother and father who hold a degree-level ISCED 5A or 6 qualification than their English-language peers. Table 4 hence clearly illustrates how there are some observable differences between students who took the test in English and those who took the test in Welsh. In the following Methodology section we discuss how we use regression analyses and an instrumental variable approach to attempt to control for such differences between these groups.

Finally, Table 5 presents the raw, unconditional differences in PISA scores between students who took the test in English and Welsh. These estimates are reported on the PISA scale, with a mean of approximately 500 and standard deviation of approximately 100 across OECD countries. When no other factors are controlled, pupils who took the test in Welsh score 41 points lower on average in reading than those who took the test in English. The difference is smaller, although still non-trivial, in mathematics (14 points) and science (27 points).

TABLE 3 Percentage of students who took the test in English or Welsh by language spoken at home in Wales

	Language spoken at home					
	English	Welsh	Irish	Ulster Scots	Other language	Total
Language of the test						
2006						
English	97	2	0.1	—	1	100
Welsh	58	42	0	—	0	100
2009						
English	94	3	<0.1	<0.1	3	100
Welsh	67	32	0	0	0.3	100
2012						
English	95	2	<0.1	<0.1	3	100
Welsh	67	33	0	0	0	100
2015						
English	93	2	0.2	0.2	4	100
Welsh	59	40	0.3	0	1	100
2018						
English	92	2	<0.1	0.1	6	100
Welsh	69	29	0	0.3	1	100
All cycles						
English	94	2	0.1	<0.1	3	100
Welsh	64	35	0.05	<0.1	0.4	100

Notes: All OECD recommended practices (final student weights and BRR weights) have been employed.

Source: authors' own calculations.

METHODOLOGY

Ordinary least squares

To begin, we analyse the influence of taking the test in Welsh using ordinary least squares (OLS). The model is specified in the following way:

$$C_{ijc} = \alpha + \beta \mathrm{WT}_{ijc} + \gamma X_{ijc} + \delta F_{ijc} + \vartheta \mathrm{SCH}_{jc} + \rho \mathrm{PISA}_{c} + \varepsilon_{ijc} \tag{1}$$

where i is the individual, j the school and c the PISA cycle; C_{ijc} are students' standardised scores in reading, mathematics and science (alternatively); WT_{ijc} is a dummy variable which indicates whether the pupil took the PISA test in Welsh (1) or English (0); X_{ijc} are students' background characteristics (i.e. sex, grade retention, student's region of birth, if the student has lived in the UK since age 6 or older or not[4] and month of birth); F_{ijc} are family characteristics (socio-economic status, father's region of birth and mother's region of birth)[5]; SCH_{jc} are school characteristics (private or public funding); PISA_c controls for PISA cycle; and ε_{ijc} is the idiosyncratic error term.

The estimated β coefficient will illustrate whether taking the PISA test in Welsh continues to be associated with lower PISA scores than taking the test in English, controlling for a wide

TABLE 4 Comparison of the demographic characteristics of pupils who took the PISA test in English and Welsh and test of mean differences

	Test language: English		Test language: Welsh	
Variables	**Observations**	**Mean**	**Observations**	**Mean**
Sex of the student				
Male	13,467	0.50	1942	0.50
Female	13,467	0.50	1942	0.50
Socio-economic status quartile				
Fourth quartile (top)	12,936	0.21^{D}	1824	0.31^{D}
Third quartile	12,936	0.24	1824	0.27
Second quartile	12,936	0.27^{D}	1824	0.24^{D}
First quartile (bottom)	12,936	0.28^{D}	1824	0.18^{D}
Father's level of education				
None	11,391	0.01	1547	0.02
ISCED 1	11,391	0.01	1547	0.01
ISCED 2	11,391	0.09^{D}	1547	0.06^{D}
ISCED 3b, c	11,391	0.28^{D}	1547	0.23^{D}
ISCED 3a, 4	11,391	0.19^{D}	1547	0.13^{D}
ISCED 5b	11,391	0.18	1547	0.18
ISCED 5a, 6	11,391	0.24^{D}	1547	0.37^{D}
Mother's level of education				
None	12,164	0.01	1650	0.01
ISCED 1	12,164	0.01	1650	0.01
ISCED 2	12,164	0.03^{D}	1650	0.02^{D}
ISCED 3b, c	12,164	0.27^{D}	1650	0.21^{D}
ISCED 3a, 4	12,164	0.21^{D}	1650	0.13^{D}
ISCED 5b	12,164	0.23	1650	0.22
ISCED 5a, 6	12,164	0.24^{D}	1650	0.40^{D}
Number of books at home				
0–10 books	13,081	0.17^{D}	1849	0.13^{D}
11–25 books	13,081	0.18^{D}	1849	0.15^{D}
26–100 books	13,081	0.30	1849	0.31
101–200 books	13,081	0.16	1849	0.18
201–500 books	13,081	0.12^{D}	1849	0.15^{D}
More than 500 books	13,081	0.07	1849	0.08
Term of birth				
First term	13,467	0.25	1942	0.24
Second term	13,467	0.25	1942	0.25
Third term	13,467	0.25	1942	0.24
Fourth term	13,467	0.25	1942	0.26

Notes: All OECD recommended practices (final student weights and BRR weights) have been employed. The "D" indicates that there are significant differences (at 5% or lower) between "Test language: English" and "Test language: Welsh" columns.

Source: authors' own calculations.

TABLE 5 Average scores and standard errors for pupils taking the PISA test in English and in Welsh

	English		Welsh	
	Mean	Standard error	Mean	Standard error
Reading	485***	2.9	444***	10.5
Mathematics	482***	4.6	468***	5.9
Science	493***	3.9	466***	7.2

Notes: All OECD recommended practices (final student weights, BRR weights and plausible values) have been employed (OECD, 2020) and standard errors are robust. The asterisks indicate significant differences between those students who took the PISA test in English (the "English" column) and those who took it in Welsh (the "Welsh" column): ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations based upon PISA data for Wales pooled between 2006 and 2018.

array of observable characteristics. Note that PISA scores have been standardised within Wales for each cycle in each subject. Results are hence presented in terms of effect sizes.

However, this β coefficient may be biased owing to potential unobservables that are not included in our model (and hence form part of the error term – ε_{ijc}) and which are also correlated with the language of the test variable. An example of these unobservables could be students' prior achievement. For instance, higher-achieving pupils may be more likely to take the PISA test in a particular language (English or Welsh), owing to for instance parental school selection,[6] but we are unable to control for that factor within our model. This omitted variable problem has been highlighted by many authors when dealing with observational cross-sectional data such as PISA (Cordero & Pedraja, 2018; Hanchane & Mostafa, 2010; Lounkaew, 2013; Micklewright et al., 2012). The direction of this bias could either be positive or negative, depending on the relationship between the omitted variable(s) (e.g. prior achievement), test language and PISA scores as the outcome. We hence employ an instrumental variable approach to try and overcome this problem, which is implemented via two-stage least squares (2SLS).

Two-stage least squares

Our instrumental variable approach needs the identification of an instrument (Z_{ijc}), and also the use of control variables (X_{ijc}, F_{ijc}, SCH_{jc}) to try and reduce the influence of any potential confounding from unobservable characteristics. The instrument we use is language spoken at home, denoted as Z_{ijc}. This is a categorical variable which can be decomposed into a set of binary variables, each one representing a different language spoken at home. This methodology requires that the instrument is correlated with the endogenous variable (the "treatment", i.e. taking the test in Welsh, WT_{ijc})[7] and uncorrelated with the error term (ε_{ijc}). This means that the instrument needs to have no independent effect on PISA scores, i.e. its sole effect is assumed to come through its influence on the language of test variable. Unfortunately, this is an untestable assumption; although certain checks can be performed, a potential link between the instrument (language spoken at home) and the outcome (PISA scores) cannot be completely ruled out. Concretely, there are four assumptions that the instrument has to meet, which we describe (and discuss how this applies within our context) in detail in Appendix A. We therefore attempt to address the potential endogeneity of taking the PISA test in Welsh (with respect to students' test scores – C_{ijc}) by using language spoken at home as an instrumental variable (IV).

This IV approach is implemented via two-stage least squares (2SLS). The *first stage* is specified as:

$$\mathrm{WT}_{ijc} = \pi_0 + \pi_1 Z_{ijc} + \pi_2 X_{ijc} + \pi_3 F_{ijc} + \pi_4 \mathrm{SCH}_{jc} + \pi_5 \mathrm{PISA}_c + \omega_{ijc} \tag{2}$$

where ω_{ijc} is the idiosyncratic error term. After estimating this equation we obtain a prediction of the Welsh test language variable ($\widehat{\mathrm{WT}}_{ijc}$), which can be included in model (1) to define the following *reduced form*:

$$C_{ijc} = \alpha + \beta \widehat{\mathrm{WT}}_{ijc} + \gamma X_{ijc} + \delta F_{ijc} + \vartheta \mathrm{SCH}_{jc} + \rho \mathrm{PISA}_c + \varepsilon_{ijc} \tag{3}$$

where β indicates the influence of taking the test in Welsh on academic performance in reading, mathematics and science, respectively. The results obtained have been checked using the Stock and Yogo (2005) test of weak instruments and the Wooldridge (1995) endogeneity test (reported below). The recommended practices for analysing PISA data (final student weights, balanced repeated replication weights[8] and plausible values) have been used throughout our analysis (OECD, 2020). In addition, instead of using a “manual” procedure for the 2SLS estimations, i.e. estimating Equations (2 and 3) separately by OLS, these equations have been estimated using a canned routine in Stata (using the ‘ivregress’ command) to ensure that standard errors are estimated correctly (as suggested by authors such as Andrews et al., 2019 and Angrist & Psichke, 2008). Note that these standard errors are also “robust” to correct for potential heteroscedasticity.

RESULTS

Table 6 presents three sets of results: first, an OLS model where the test language variable is included and no other variables are controlled;[9] second, estimates from an OLS model including controls (with the full set of parameter estimates presented in Appendix B, Table B1); and finally, our instrumental variable results.

Starting with our conditional OLS model, one can see that taking the PISA test in Welsh is negatively associated with pupils’ test scores. Specifically, those who took the test in Welsh scored 0.42 standard deviations (SDs) lower in reading, 0.18 SDs lower in mathematics and 0.36 SDs lower in science, than pupils who took the test in English. Interestingly, the inclusion of controls has increased the magnitude of the effect sizes, compared with the OLS estimates without controls.

These OLS results may, however, omit certain (unobservable) variables, which may confound the relationship between students’ academic achievement and test language. We hence move on to results from our instrumental variable approach. The first-stage estimations from Equation (2) are reported in Table B1 (Appendix B). The results for the instrument (i.e. language at home) are significant in explaining the treatment variable (Welsh test language), which supports the *relevance condition* (see Appendix A for further details). Relatedly, the null hypothesis of the Stock and Yogo (2005) test (that the instrument is weak) can clearly and decisively be rejected, supporting the *relevance condition*.

The second stage of our instrumental variable estimates produces similar substantive results to those produced under OLS. Specifically, there continues to be a negative influence of taking the test in Welsh (relative to taking the test in English) upon pupils’ reading (0.39 SDs), mathematics (0.26 SDs) and science (0.33 SDs) PISA scores. To give readers a perspective of the magnitude of these effects, the difference in PISA scores between students from the most advantaged and least advantaged socio-economic status quartile is around 0.80 SDs (as reported in Table B1, Appendix B). In other words, in reading, the effect of taking the test in Welsh rather than English is equal to approximately half the size of the

TABLE 6 Influence of taking the test in Welsh on students' competences in Wales, effect sizes

	Effect of taking test in Welsh (compared with English)	Standard error	Stock and Yogo (2005) test of weak instruments	Wooldridge (1995) endogeneity test
Reading				
OLS (no controls)	−0.341***	0.042	—	—
OLS (with controls)	−0.421***	0.036	—	—
2SLS	−0.392***	0.074	63.774***	0.226
Mathematics				
OLS (no controls)	−0.106***	0.042	—	—
OLS (with controls)	−0.183***	0.034	—	—
2SLS	−0.262***	0.069	63.774***	1.880
Science				
OLS (no controls)	−0.282***	0.039	—	—
OLS (with controls)	−0.363***	0.034	—	—
2SLS	−0.328***	0.063	63.774***	0.380

Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2020) and standard errors are robust. The null hypothesis of the Stock and Yogo (2005) test of weak instruments is that the instrument is weak and the null hypothesis of the Wooldridge (1995) endogeneity test is that the endogenous variable is now exogenous. Complete estimations are presented in Table B1 (Appendix B). The sample size of these estimations is 14,951 students.

Estimation method: ordinary least squares (OLS) and two-stage least squares (2SLS). The instrument is student's language at home and the rest of variables in the estimation.

Dependent variable: student's standardised scores in reading, mathematics and science, using Welsh mean and standard deviations in each PISA cycle.

Coefficient: ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations.

socio-economic status achievement gap. This is hence clearly a very sizeable influence. Furthermore, the Wooldridge (1995) endogeneity[10] test cannot be rejected (in which the null hypothesis is that the endogenous variable is now exogenous), providing reassurance that our IV estimates are likely to move us a step closer to obtaining causal effects.

In order to check the robustness of our results, we have replicated the analysis using the subsample of schools where there was a mix of pupils taking the test in English and Welsh. This is to check whether our results are being driven by the particular characteristics of Welsh-medium schools or not. These alternative results are presented in Table 7, with the full set of parameter estimates provided in Appendix B (Table B2). This leads to little change to our substantive results; we continue to observe a 0.38 SDs difference in reading, 0.32 SDs in mathematics and 0.34 SDs in science. Similarly, in Table 8 we replicate the analysis from Table 7 again, but now also additionally including school fixed effects (see Appendix B, Table B3 for the full set of parameter estimates). Including school fixed effects means that a dummy variable for each school (omitting one as reference) has been included in the regression. The aim of this methodology is to control by variables that are fixed within the school and hence remove all between-school differences (such as differences in language of instruction used by different schools) that may impact upon the results. This further confirms that our findings are driven by within-school – and not between-school – differences. In particular, the inclusion of the school fixed effects allows us to rule out that the results are being driven by school-level factors (most notably secondary school language of instruction). Again, there is little change to our substantive conclusions, with a sizeable difference continuing to be observed in reading (0.41 SDs), mathematics (0.22 SDs) and science (0.29 SDs).[11]

TABLE 7 Influence of taking the test in Welsh on students' competences in Wales. Sample restricted to schools with a mix of English and Welsh test-takers, effect sizes

	Effect of taking test in Welsh (compared with English)	Standard error	Stock and Yogo (2005) test of weak instruments	Wooldridge (1995) endogeneity test
Reading				
OLS	−0.417***	0.068	—	—
2SLS	−0.383***	0.109	36.525***	0.149
Mathematics				
OLS	−0.256***	0.06	—	—
2SLS	−0.321***	0.105	36.525***	0.516
Science				
OLS	−0.360***	0.057	—	—
2SLS	−0.340***	0.094	36.525***	0.061

Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2020) and standard errors are robust. The null hypothesis of the Stock and Yogo (2005) test of weak instruments is that the instrument is weak and the null hypothesis of the Wooldridge (1995) endogeneity test is that the endogenous variable is now exogenous. Complete estimations are presented in Table B2 (Appendix B). The sample size of these estimations is 1502 students.

Estimation method: OLS and 2SLS. The instrument is student's language at home and the other variables in the estimation.

Dependent variable: student's standardised scores in reading, mathematics and science, using Welsh mean and standard deviations in each PISA cycle.

Coefficient: ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations.

CONCLUSION

This paper has investigated the influence of PISA test language on students' academic performance in Wales, an issue that is at the core of this country's education debate. It is the first time that a study has attempted to produce quasi-experimental evidence on this issue, using an instrumental variable approach applied to five cycles of PISA data. Our results show that students who took the test in Welsh performed around 0.39 SDs (39 points on the PISA scale[12]) lower in reading, 0.26 SDs (26 PISA points) lower in mathematics and 0.33 SDs (33 PISA points) lower in science, compared with their peers who took the test in English. Taking into account that 25–30 points in the PISA scale is equivalent to one year of schooling (OECD, 2019), this is clearly a sizable effect. As one would anticipate, reading is the subject most affected by this problem, although with non-trivial differences in achievement between English and Welsh test-takers also observed in science and mathematics.

These findings should of course be taken in light of the limitations of this study. First, the use of observational and cross-sectional data means that, in spite of using an instrumental variable approach, it may be prudent to continue to interpret our estimates as correlations. Second, although this research has internal validity for Wales, results may or may not generalise to other national settings. They nevertheless raise some important questions about how the PISA test has been conducted in Wales, and changes that may need to be made to the data collection in the future.

In particular, it is important to consider what may be driving our results. As our analysis demonstrated, the negative influence of taking the PISA test in Welsh does not seem to be due to studying in Welsh-medium schools *per se*. We continue to find a sizeable difference in PISA scores even in bilingual schools, where some students took the test in English and others took the test in Welsh. One plausible alternative explanation is that our findings may reflect a problem with translation, with authors such as Blum et al. (2001) noting how such

TABLE 8 Influence of taking the test in Welsh on students' competences in Wales. Sample restricted to schools with a mix of English and Welsh test-takers and school fixed effects included, effect sizes

	Effect of taking test in Welsh (compared with English)	Standard error	Stock and Yogo (2005) test of weak instruments	Wooldridge (1995) endogeneity test
Reading				
OLS	−0.395***	0.068	—	—
2SLS	−0.405***	0.116	37.715***	0.009
Mathematics				
OLS	−0.190***	0.062	—	—
2SLS	−0.216*	0.121	37.715***	0.046
Science				
OLS	−0.316***	0.065	—	—
2SLS	−0.293***	0.111	37.715***	0.046

Notes: PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2020) and standard errors are robust. The null hypothesis of the Stock and Yogo (2005) test of weak instruments is that the instrument is weak and the null hypothesis of the Wooldridge (1995) endogeneity test is that the endogenous variable is now exogenous. Complete estimations are presented in Table B3 (Appendix B). The sample size of these estimations is 1502 students.

Estimation method: OLS and 2SLS. The instrument is student's language at home and the other variables in the estimation.

Dependent variable: student's standardised scores in reading, mathematics and science, using Welsh mean and standard deviations in each PISA cycle.

Coefficient: ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations.

problems have affected the validity of other cross-national studies (e.g. the International Adult Literacy Survey). Indeed, we note how the PISA 2018 technical report (OECD, 2020, Chapter 5) states that "international verification was carried out for all national versions in languages used in schools attended by more than 10% of the country's target population". Importantly, this would seem to suggest that there has not been independent verification of the translation of the Welsh PISA survey instruments – given how Wales makes up around 5% of the population of the UK.[13] This is also confirmed by the PISA 2012 technical report (OECD, 2014, p. 94), which notes how the Welsh translation only went through national – rather than international – verification. We therefore recommend that, in the future, there is greater independent verification of the Welsh versions of the PISA test.

The challenges with translation of the PISA test in Wales are likely to be exacerbated by the important disparity between spoken and literary Welsh (see Fife, 1986 for a discussion of this issue). The difference between literary and colloquial Welsh means that many people who grew up speaking mainly in Welsh find the written form difficult to understand. Thus, while the translation of the PISA questions into Welsh may be technically correct, they may also not be how people would say them, thus making them harder to understand (than the English versions).

Yet translation issues are only one potential explanation for our findings. Minor differences between source versions and national translations have been found elsewhere in the literature (e.g. Murat & Rocher, 2004, for PISA 2000 data, and Grisay et al., 2007, for PISA 2006 data), but the impact of this was thought to be minimal. Similarly, Oliden and Lizaso (2013) analysed PISA 2009 data for Spain and found that the Spanish translation and those for the other languages in this country (Galician, Catalan and Basque) were equivalent and would not have had a substantial impact upon the results.

An alternative explanation is that, as previously indicated, there may be issues with how the PISA test language is chosen in Wales. Specifically, some students may be forced – or strongly encouraged – to take the test in Welsh if that is the most commonly used medium of instruction in their school, when English (the language they most often speak at home) would actually be a more appropriate choice. This is important, as PISA is meant to capture pupils' skills in each subject area and not their level of understanding of the test language *per se*. Thus, in Wales, PISA technical standard 2.1 (pupils should take the test in the language they are most comfortable with) might not be fully applied. This could, in turn, mean that the academic abilities of Welsh 15-year-olds are being underestimated in PISA, owing to the inappropriate allocation of test language for some children. We hence encourage those conducting the PISA test in Wales to provide greater reassurance that this technical standard is being properly applied in the future.

How should such issues be resolved?

One option could be that all children who take the PISA test in Wales get a genuinely free choice of whether to take the test in English or Welsh – regardless of the medium of instruction most frequently used within their school (even if this means some teenagers taking PISA in English within Welsh-medium schools). An alternative could be testing whether students have sufficient Welsh language skills before they are asked to take the Welsh version of the PISA test. However, perhaps the optimum solution would be for a very simple change to be made to how the PISA assessment is delivered. Currently, when sitting the PISA test, young people can only see the test questions in a single language (e.g. Welsh). This is at-odds with what happens in GCSE and A-Level examinations in Wales, where young people are provided with the test questions in both English and Welsh. Now that PISA is a computer-based assessment, it should be relatively simple to include a "toggle" button on each question, allowing young people to see each question in the language that they prefer. This would, of course, not help to only improve the PISA test-taking experience (and data quality) in Wales, but also in other countries where similar issues arise as well.

ETHICAL GUIDELINES

BERA ethical guidelines have been followed in the present research study.

ACKNOWLEDGEMENTS

This work has been partly supported by FEDER funding (under Research Project PY20-00228), the Ministerio de Ciencia e Innovación (under Research Project PID2020-119471RB-I00) and the Andalusian Regional Government (SEJ-645). We would also like to thank Ceri Jones (Estyn) for helpful comments made on the paper.

CONFLICT OF INTEREST

The authors do not have any conflict of interests.

DATA AVAILABILITY STATEMENT

The data employed in the present research is publicly available in https://www.oecd.org/pisa/data/ (OECD, 2021).

ORCID

John Jerrim https://orcid.org/0000-0001-5705-7954
Luis Alejandro Lopez-Agudo https://orcid.org/0000-0002-0906-3206
Oscar David Marcenaro-Gutierrez https://orcid.org/0000-0003-0939-5064

ENDNOTES

[1] Following this standard, students with insufficient experience in the language of assessment are excluded from PISA. In particular, these students are those who: (a) are not native speakers of the assessment language; (b) have limited proficiency in the assessment language; and (c) have received less than one year of instruction in the assessment language. Furthermore, students are also excluded from PISA when there are no available materials in the language in which the student is taught.

[2] This is also indicated in the technical reports of all the PISA cycles under analysis in the present study in OECD (2009, 2014, 2017, 2020).

[3] More information on PISA administration in Wales can be found in Sizmur et al. (2019, pp. 199–200).

[4] This variable controls the potential differences in language skills between those students who arrived in the UK and started compulsory education at age 6 or before, and those who arrived and started after that age.

[5] The combination of the student's, father's and mother's region of birth variables also controls for student's immigrant status.

[6] In this sense, Freeman and Viarengo (2014, p. 405) indicated that "the assignment of students among schools is not random. Students of similar ability are likely to be sorted among schools".

[7] The case of using a binary endogenous variable with 2SLS was analysed by Angrist and Psichke (2008), who reached the conclusion that using the "garden-variety 2SLS" (i.e. common 2SLS, in which having a binary treatment means having a linear probability model in the first stage) would be the best alternative (Kyriopoulos et al., 2018; Carlin, Olafsson, & Pagel, 2019). Thus, this is the methodology that we employed in the present study. However, we replicate our 2SLS main results using a first-stage probit and, as we will see, the results changed only slightly.

[8] Balanced repeated replication (BRR) weights control the multi-level structure of the data, producing unbiased standard errors, also clustering at school level (OECD, 2020).

[9] The complete estimations with the test language variable and no additional controls, for both OLS and 2SLS, are presented in Table S1 (Online Data S1).

[10] The Wooldridge (1995) endogeneity test is a standard for testing endogeneity in estimations with a robust variance–covariance matrix (Stata, 2021).

[11] The estimations of Tables 6–8 were replicated using probit estimations in the first stage of the two-stage least squares models and the results changed only slightly. The coefficients of our variable of interest, “Effect of taking test in Welsh (compared to English)”, and their standard errors are presented in Table S2 (Online Data S1), replicating in specification I the coefficients of Table 6, in specification II the coefficients of Table 7 and in specification III the coefficients of Table 8.

[12] This scale presents a mean of 500 and standard deviation of 100.

[13] The UK was treated as a single country in PISA and students’ results were reported for it as a whole, rather than for each of the four countries within. This is because Wales is not an “adjudicated sub-region”, so the 10% rule will not be applied to it separately; rather, it will be applied to the UK as a whole, so Wales was considered as part of the UK for sampling purposes and hence represents 5% of the UK population. This was the case for all the PISA cycles under analysis (OECD, 2009, 2012, 2014, 2017).

REFERENCES

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*, 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angrist, J. D., & Psichke, J.-S. (2008). *Mostly harmless econometrics. An Empiricist's companion*. Princeton University Press.

Barua, R., & Lang, K. (2016). School entry, educational attainment and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital*, *10*(3), 347–376. https://doi.org/10.1086/687599

Blum, A., Goldstein, H., & Guérin-Pace, F. (2001). International Adult Literacy Survey (IALS): An analysis of international comparisons of adult literacy. *Assessment in Education: Principles, Policy & Practice*, *8*(2), 225–246. https://doi.org/10.1080/09695940123977

Carlin, B., Olafsson, A., & Pagel, M. (2019). *FinTech and Consumer Financial Well-Being in the Information Age*. Federal Deposit Insurance Corporation.

Cordero, J. M., & Pedraja, F. (2018). The effect of financial education training on the financial literacy of Spanish students in PISA. *Applied Economics*, *51*(16), 1679–1693. https://doi.org/10.1080/00036846.2018.1528336

De Bortoli, L., & Cresswell, J. (2004). *Australia’s Indigenous Students in PISA 2000: Results from an International Study*. ACER Research Monograph No 59, 1–42. Australia: Australian Council for Education Research.

Dhuey, E., Figlio, D., Karbownik, K., & Roth, J. (2019). School starting age and cognitive development. *Journal of Policy Analysis and Management*, *38*(9), 538–578. https://doi.org/10.1002/pam.22135

Edwards, V., & Newcombe, L. P. (2006). Back to basics: Marketing the benefits of Bilingualism to parents. In O. García, T. Skutnabb-Kangas, & M. E. Torres-Guzmán (Eds.), *Imagining multilingual schools: Languages in education and glocalization* (pp. 137–149). Multilingual Matters.

Fife, J. (1986). Literary vs. colloquial Welsh: Problems of definition. *Word*, *37*(3), 141–151.

Fiorini, M., & Stevens, K. (2014). *Assessing the monoticity assumption in IV and fuzzy RD designs*. The University of Sydney Economics Working Paper Series, 13, 1–52.

Freeman, R. B., & Viarengo, M. (2014). School and family effects on educational outcomes across countries. *Economic Policy*, *29*(79), 395–446. https://doi.org/10.1111/1468-0327.12033

Gorard, S. (1998). Four errors … and a conspiracy? The effectiveness of schools in Wales. *Oxford Review of Education*, *24*(4), 459–472. https://doi.org/10.1080/0305498980240403

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249–266.

Hanchane, S., & Mostafa, T. (2010). *Endogeneity Problems in Multilevel Estimation of Education Production Functions: an Analysis Using PISA Data*. LLAKES Research Paper 14, 1–45. London: Centre for Learning and Life Chances in Knowledge Economies and Societies.

Jerrim, J., & Shure, N. (2016). *Achievement of 15-year- olds in Wales: PISA 2015 national report*. UCL Institute of Education.

Johnes, G. (2020). Medium efficiency: Comparing Inputs and outputs by language of instruction in secondary schools in Wales. *Wales Journal of Education*, *22*(2), 52–66.

Jones, B. (2017a). Translanguaging in Bilingual schools in Wales. *Journal of Language, Identity & Education*, *16*(4), 199–215. https://doi.org/10.1080/15348458.2017.1328282

Jones, M. (2016). *Research Briefing. Welsh-medium education and Welsh as a subject*. National Assembly for Wales. Research Service.

Jones, S. L. (2017b). What do we know and not know about choice of medium of education in South-East Wales? *Wales Journal of Education*, *19*(2), 143–162. https://doi.org/10.16922/wje.19.2.8

Kennedy, E., & Park, H.-S. (1994). Home language as a predictor of academic achievement: A comparative study of Mexican- and Asian-American youth. *Journal of Research & Development in Education*, *27*(3), 188–194.

Kyriopoulos, I., Athanasakis, K., & Kyriopoulos, J. (2018). Are happy people healthier? An instrumental variable approach using data from Greece. *Journal of Epidemiology and Community Health*, *72*, 1153–1161. https://doi.org/10.1136/jech-2018-210568

Lounkaew, K. (2013). Explaining urban–rural differences in educational achievement in Thailand: Evidence from PISA literacy data. *Economics of Education Review*, *37*, 213–225. https://doi.org/10.1016/j.econedurev.2013.09.003

Mancilla-Martinez, J., & Lesaux, N. K. (2011). Early home language use and later vocabulary development. *Journal of Educational Psychology*, *103*(3), 535–546. https://doi.org/10.1037/a0023655

Micklewright, J., Schnepf, S. V., & Silva, P. N. (2012). Peer effects and measurement error: The impact of sampling variation in school survey data (evidence from PISA). *Economics of Education Review*, *31*(6), 1136–1142. https://doi.org/10.1016/j.econedurev.2012.07.015

Murat, F., & Rocher, T. (2004). The methods used for international assessments of educational competencies. In J. H. Moskowitz, & M. Stephens (Eds.), *Comparing learning outcomes. International assessment and educational policy* (pp. 190–214). Routledge Farmer.

OECD. (2009). *PISA 2006 Technical Report*. OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. OECD Publishing. https://doi.org/10.1787/9789264167872-en

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

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.

OECD. (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing.

OECD. (2020). *PISA 2018 Technical Report*. OECD Publishing.

OECD. (2021). PISA Data; OECD PISA web. https://www.oecd.org/pisa/data/

Oliden, P. E., & Lizaso, J. M. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema*, *25*(3), 390–395. https://doi.org/10.7334/psicothema2013.46

Parliament of the United Kingdom. (2002). *Education Act 2002*. https://www.legislation.gov.uk/ukpga/2002/32/section/105/enacted?view=plain

Sizmur, J., Ager, R., Bradshaw, J., Classick, R., Galvis, M., Packer, J., Thomas, D., & Wheater, R. (2019). *Achievement of 15-year-olds in Wales: PISA 2018 National report*. NFER.

Stata. (2021). *Stata 13: Ivregress postestimation*. https://www.stata.com/manuals13/rivregresspostestimation.pdf

StatsWales. (2021). Speaking Welsh at home, as assessed by parents, of pupils aged 5 and over in primary, middle and secondary schools by year, sector and category. https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Schools-Census/Pupil-Level-Annual-School-Census/Welsh-Language/speakingwelshhomepupils5andover-by-year-sector-category

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews, & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press.

Van den Brande, J., Hillary, J., & Cullinane, C. (2019). *Selective comprehensives: Great Britain*. National Foundation for Educational Research (NFER).

Welsh Assembly Government. (2007). *Defining schools according to Welsh medium provision*. Welsh Assembly Government.

Wooldridge, J. M. (1995). Score diagnostics for linear models estimated by two stage least squares. In G. S. Maddala, T. N. Srinivasan, & P. C. B. Phillips (Eds.), *Advances in Econometrics and Quantitative Economics: Essays in Honor of Profesor C. R. Rao* (pp. 66–87). Blackwell.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Jerrim, J., Lopez-Agudo, L. A., & Marcenaro-Gutierrez, O. D. (2022). The impact of test language on PISA scores. New evidence from Wales. *British Educational Research Journal*, 00, 1–26. https://doi.org/10.1002/berj.3774

APPENDIX A

To have a credible two-stage least squares estimation approach, our instrument has to fulfil the following assumptions:

a. The *relevance condition* or *first stage*. This means that the instrument should be strongly associated with the "treatment" variable (i.e. taking the test in Welsh). This is clearly the case in this study, as language at home (Z_{ijc}) is strongly linked to the language of the test (WT_{ijc}). We have already illustrated this point descriptively in Table 3, with more formal results from the Stock and Yogo (2005) test of weak instruments presented in the Results section.
b. The *independence/exogeneity assumption*. This condition means that the instrument should be randomly assigned or "as good as randomly assigned", meaning that it is uncorrelated with the omitted variables we might like to control for. In our study, the language at home instrument might be considered as good as randomly assigned after controlling by X_{ijc}, F_{ijc} and SCH_{jc} – specifically after controlling for pupils' socio-economic status, their country of birth, their parents' country of birth and school characteristics. This exogeneity of the language at home is due to: (a) there being no home language choice in monolingual households; and (b) in plurilingual households, once all of the control variables have been included, the choice between one language and another is assumed to be as good as random.
c. The *exclusion restriction*. This means that there is a sole channel (this is, through WT_{ijc}) for the effect of the instrument (Z_{ijc}, language at home) on the outcome (C_{ijc}, students' competences). This single channel requires the previous independence assumption, to the extent that the other potential channels of influence have been controlled (X_{ijc}, F_{ijc} and SCH_{jc}).
d. The *monotonicity property* (Barua & Lang, 2016; Dhuey et al., 2019; Fiorini & Stevens, 2014). As defined by Barua and Lang (2016) "while the instrument may have no effect on some individuals, all of those who are affected should be affected unidirectionally" (p. 348). This is also known as the *no defiers assumption*, i.e. there are no students who, if they are assigned to take the test in their home language, always choose to take it in another language. Likewise, if they are assigned to take the test in a different language than the one spoken at home, they always choose to take the test in their home language. In schools where all pupils were made to do the test in the same language, the monotonicity property is fulfilled, as students could not choose; as previously indicated, in mixed-language schools this test-language obligation may not happen. Hence, in the cases of schools in which pupils could choose the test language, it is assumed they always chose the language they were most comfortable with. In particular, as previously indicated, from a total of 607 school observations in our dataset, all pupils took the test in English in 503 schools, in 44 schools all pupils took the test in Welsh and in 60 schools there was a mix of English and Welsh test takers.

APPENDIX B

TABLE B1 Influence of taking the test in Welsh on students' competences in Wales, effect sizes. Full parameter estimates

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Test language: Welsh (ref.: English)	−0.421*** (0.036)	−0.183*** (0.034)	−0.363*** (0.034)	—	−0.392*** (0.074)	−0.262*** (0.069)	−0.328*** (0.063)
Female: yes (ref.: no)	0.244*** (0.015)	−0.165*** (0.015)	−0.088*** (0.016)	−0.004 (0.004)	0.244*** (0.015)	−0.165*** (0.015)	−0.088*** (0.016)
Socio-economic status quartile (ref.: first quartile)							
Fourth quartile	0.754*** (0.025)	0.807*** (0.024)	0.798*** (0.026)	−0.063*** (0.011)	0.751*** (0.026)	0.815*** (0.025)	0.794*** (0.027)
Third quartile	0.398*** (0.022)	0.432*** (0.021)	0.434*** (0.022)	−0.042*** (0.008)	0.397*** (0.022)	0.435*** (0.021)	0.432*** (0.022)
Second quartile	0.242*** (0.020)	0.245*** (0.020)	0.243*** (0.020)	−0.023*** (0.007)	0.241*** (0.020)	0.247*** (0.020)	0.242*** (0.020)
Socio-economic status quartile. Missing flag	−0.389*** (0.062)	−0.334*** (0.059)	−0.349*** (0.062)	−0.051*** (0.018)	−0.391*** (0.062)	−0.329*** (0.059)	−0.351*** (0.062)
Grade retention (ref.: no)							
Repeater	−0.674*** (0.056)	−0.706*** (0.051)	−0.627*** (0.054)	0.041*** (0.015)	−0.673*** (0.056)	−0.710*** (0.051)	−0.626*** (0.054)
Repeater. Missing flag	−0.555*** (0.125)	−0.451*** (0.121)	−0.576*** (0.128)	0.036 (0.033)	−0.555*** (0.124)	−0.452*** (0.122)	−0.575*** (0.127)
Country of birth (ref.: other countries)							
UK	0.089 (0.058)	0.089 (0.057)	0.030 (0.056)	−0.010 (0.013)	0.089 (0.057)	0.090 (0.057)	0.029 (0.056)
Country of birth. Missing flag	−0.302*** (0.083)	−0.312*** (0.081)	−0.380*** (0.082)	0.005 (0.023)	−0.302*** (0.083)	−0.312*** (0.081)	−0.379*** (0.082)
Father's country of birth (ref.: other countries)							
UK	0.016 (0.036)	−0.014 (0.038)	0.025 (0.038)	−0.026 (0.017)	0.014 (0.036)	−0.010 (0.039)	0.023 (0.038)

TABLE B1 (Continued)

	OLS			2SLS			
					Second stage		
Variables	**Reading**	**Mathematics**	**Science**	**First stage**	**Reading**	**Mathematics**	**Science**
Father's country of birth. Missing flag	−0.172** (0.069)	−0.218*** (0.067)	−0.159** (0.069)	−0.036 (0.025)	−0.174** (0.069)	−0.212*** (0.067)	−0.161** (0.069)
Mother's country of birth (ref.: other countries)							
UK	−0.113*** (0.037)	−0.133*** (0.039)	−0.109*** (0.039)	−0.035** (0.015)	−0.114*** (0.037)	−0.129*** (0.039)	−0.111*** (0.039)
Mother's country of birth. Missing flag	−0.256*** (0.093)	−0.330*** (0.093)	−0.284*** (0.092)	−0.023 (0.033)	−0.258*** (0.093)	−0.324*** (0.093)	−0.286*** (0.092)
Has lived in the UK since age 6 or older: yes (ref.: no)	−0.216*** (0.081)	−0.171** (0.074)	−0.229*** (0.081)	0.018 (0.016)	−0.216*** (0.081)	−0.172** (0.074)	−0.228*** (0.080)
Term of birth (ref.: fourth term)							
First term	−0.058*** (0.020)	−0.035* (0.020)	−0.040** (0.020)	0.009 (0.006)	−0.058*** (0.020)	−0.036* (0.020)	−0.040** (0.020)
Second term	−0.142*** (0.021)	−0.109*** (0.020)	−0.111*** (0.021)	0.007 (0.006)	−0.142*** (0.021)	−0.109*** (0.020)	−0.111*** (0.021)
Third term	−0.051** (0.021)	−0.024 (0.021)	−0.035 (0.022)	0.010 (0.006)	−0.051** (0.021)	−0.025 (0.021)	−0.034 (0.022)
School funding (ref.: public)							
Private	0.543*** (0.099)	0.614*** (0.096)	0.557*** (0.085)	0.110*** (0.016)	0.547*** (0.099)	0.602*** (0.096)	0.562*** (0.086)
School funding. Missing flag	−0.133** (0.060)	−0.097** (0.048)	−0.074 (0.053)	0.013 (0.035)	−0.133** (0.060)	−0.098** (0.048)	−0.073 (0.053)
PISA cycle (ref.: 2006)							
2018	0.043 (0.046)	0.015 (0.040)	0.025 (0.041)	−0.025 (0.036)	0.043 (0.046)	0.017 (0.040)	0.024 (0.041)
2015	0.050 (0.046)	0.042 (0.038)	0.029 (0.041)	0.010 (0.030)	0.050 (0.046)	0.041 (0.038)	0.030 (0.041)
2012	0.024 (0.048)	0.024 (0.039)	0.018 (0.044)	−0.009 (0.032)	0.024 (0.048)	0.024 (0.040)	0.018 (0.044)
2009	0.027 (0.045)	0.022 (0.041)	0.021 (0.043)	−0.030 (0.039)	0.027 (0.045)	0.024 (0.042)	0.020 (0.043)

TABLE B1 (Continued)

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Language at home (ref.: English)							
Welsh	—	—	—	0.591*** (0.037)	—	—	—
Irish	—	—	—	−0.010 (0.047)	—	—	—
Ulster Scots	—	—	—	0.033 (0.083)	—	—	—
Other languages	—	—	—	0.008 (0.015)	—	—	—
Constant	−0.295***	−0.116	−0.125*	−0.013	−0.295***	−0.118	−0.124*
	(0.074)	(0.074)	(0.073)	(0.028)	(0.074)	(0.074)	(0.073)
Observations	14,951	14,951	14,951	14,951	14,951	14,951	14,951
R^2	0.168	0.160	0.152	0.208	0.168	0.159	0.152
Stock and Yogo (2005) test of weak instruments	—	—	—	—	63.774***	63.774***	63.774***
Wooldridge (1995) endogeneity test	—	—	—	—	0.226	1.880	0.380

Notes: Standard errors in parentheses are robust. PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2020). The null hypothesis of the Stock and Yogo (2005) test of weak instruments is that the instrument is weak and the null hypothesis of the Wooldridge (1995) endogeneity test is that the endogenous variable is now exogenous.

Estimation method: OLS and 2SLS. The instrument is student's language at home and the rest of variables in the estimation.

Dependent variable: student's standardised scores in reading, mathematics and science, using Welsh mean and standard deviations in each PISA cycle.

Coefficient: ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations.

TABLE B2 Influence of taking the test in Welsh on students' competences in Wales. Sample restricted to schools with a mix of English and Welsh test-takers, effect sizes. Full estimates

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Test language: Welsh (ref.: English)	−0.417*** (0.068)	−0.256*** (0.060)	−0.360*** (0.057)	—	−0.383*** (0.109)	−0.321*** (0.105)	−0.340*** (0.094)
Female: yes (ref.: no)	0.226*** (0.048)	−0.192*** (0.051)	−0.085 (0.052)	−0.015 (0.024)	0.226*** (0.047)	−0.192*** (0.050)	−0.085* (0.051)
Socio-economic status quartile (ref.: first quartile, bottom)							
Fourth quartile (top)	0.689*** (0.062)	0.766*** (0.058)	0.703*** (0.060)	−0.107*** (0.040)	0.684*** (0.063)	0.775*** (0.059)	0.700*** (0.062)
Third quartile	0.360*** (0.064)	0.400*** (0.068)	0.387*** (0.073)	−0.066* (0.037)	0.359*** (0.063)	0.403*** (0.067)	0.386*** (0.072)
Second quartile	0.152** (0.059)	0.155** (0.064)	0.206*** (0.060)	−0.036 (0.040)	0.152*** (0.058)	0.156** (0.062)	0.205*** (0.059)
Socio-economic status quartile. Missing flag	−0.307** (0.137)	−0.490*** (0.133)	−0.407*** (0.129)	−0.036 (0.092)	−0.312** (0.134)	−0.481*** (0.132)	−0.410*** (0.128)
Grade retention (ref.: no)							
Repeater	−0.915*** (0.178)	−0.839*** (0.193)	−0.806*** (0.172)	0.057 (0.058)	−0.911*** (0.176)	−0.846*** (0.192)	−0.804*** (0.169)
Repeater. Missing flag	0.146 (0.520)	0.064 (0.422)	−0.114 (0.410)	0.012 (0.180)	0.148 (0.516)	0.059 (0.406)	−0.113 (0.406)
Country of birth (ref.: other countries)							
UK	0.312 (0.229)	0.259 (0.228)	0.214 (0.251)	−0.095 (0.119)	0.311 (0.225)	0.262 (0.225)	0.213 (0.246)
Country of birth. Missing flag	0.154 (0.305)	0.116 (0.288)	0.095 (0.308)	−0.263 (0.171)	0.150 (0.300)	0.124 (0.287)	0.092 (0.303)
Father's country of birth (ref.: other countries)							
UK	−0.253** (0.124)	−0.342*** (0.127)	−0.253* (0.136)	−0.099 (0.064)	−0.259** (0.123)	−0.331*** (0.126)	−0.256* (0.135)
Father's country of birth. Missing flag	−0.487*** (0.152)	−0.634*** (0.167)	−0.456*** (0.166)	−0.139 (0.087)	−0.495*** (0.147)	−0.619*** (0.159)	−0.461*** (0.160)
Mother's country of birth (ref.: other countries)							
UK	−0.163 (0.103)	−0.057 (0.111)	−0.085 (0.112)	−0.013 (0.078)	−0.163 (0.102)	−0.056 (0.108)	−0.085 (0.111)

(Continues)

TABLE B2 (Continued)

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Mother's country of birth. Missing flag	−0.139 (0.234)	0.038 (0.264)	−0.015 (0.243)	0.041 (0.122)	−0.138 (0.231)	0.036 (0.257)	−0.014 (0.239)
Has lived in the UK since age 6 or older: yes (ref.: no)	−0.005 (0.398)	−0.327 (0.419)	−0.198 (0.483)	0.183 (0.157)	0.003 (0.390)	−0.343 (0.413)	−0.194 (0.475)
Term of birth (ref.: fourth term)							
First term	−0.125* (0.066)	−0.060 (0.069)	−0.072 (0.068)	0.044 (0.038)	−0.124* (0.066)	−0.062 (0.069)	−0.071 (0.068)
Second term	−0.135** (0.053)	−0.062 (0.058)	−0.090 (0.059)	0.043 (0.032)	−0.135*** (0.052)	−0.064 (0.057)	−0.089 (0.057)
Third term	−0.026 (0.055)	0.015 (0.062)	0.006 (0.057)	0.033 (0.031)	−0.025 (0.056)	0.013 (0.061)	0.007 (0.057)
School funding (ref.: public)							
Private	—	—	—	—	—	—	—
School funding. Missing flag	0.006 (0.132)	0.001 (0.137)	0.102 (0.137)	−0.122 (0.097)	0.002 (0.135)	0.008 (0.130)	0.100 (0.139)
PISA cycle (ref.: 2006)							
2018	−0.256* (0.128)	−0.184 (0.118)	−0.079 (0.088)	−0.051 (0.093)	−0.257** (0.126)	−0.183 (0.115)	−0.080 (0.087)
2015	−0.039 (0.116)	−0.010 (0.107)	−0.001 (0.083)	−0.046 (0.086)	−0.040 (0.112)	−0.008 (0.106)	−0.002 (0.082)
2012	0.159 (0.114)	0.207* (0.107)	0.109 (0.098)	0.056 (0.080)	0.161 (0.111)	0.204* (0.105)	0.110 (0.097)
2009	−0.081 (0.111)	−0.051 (0.114)	−0.039 (0.080)	−0.014 (0.098)	−0.080 (0.108)	−0.053 (0.110)	−0.038 (0.079)
Language at home (ref.: English)							
Welsh	—	—	—	0.537*** (0.047)	—	—	—
Irish	—	—	—	0.351 (0.272)	—	—	—
Ulster Scots	—	—	—	0.135 (0.148)	—	—	—

TABLE B2 (Continued)

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Other languages	—	—	—	0.326** (0.154)	—	—	—
Constant	−0.052 (0.251)	0.109 (0.244)	0.018 (0.243)	0.031 (0.130)	−0.060 (0.247)	0.124 (0.238)	0.014 (0.238)
Observations	1502	1502	1502	1502	1502	1502	1502
R^2	0.202	0.186	0.156	0.289	0.202	0.185	0.156
Stock and Yogo (2005) test of weak instruments	—	—	—	—	36.525***	36.525***	36.525***
Wooldridge (1995) endogeneity test	—	—	—	—	0.149	0.516	0.061

Notes: Standard errors in parentheses are robust. PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2020). The null hypothesis of the Stock and Yogo (2005) test of weak instruments is that the instrument is weak and the null hypothesis of the Wooldridge (1995) endogeneity test is that the endogenous variable is now exogenous.

Estimation method: OLS and 2SLS. The instrument is student's language at home and the other variables in the estimation.

Dependent variable: student's standardised scores in reading, mathematics and science, using Welsh mean and standard deviations in each PISA cycle.

Coefficient: ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations.

TABLE B3 Influence of taking the test in Welsh on students' competences in Wales. Sample restricted to schools with a mix of English and Welsh test-takers and school fixed effects included, effect sizes. Full estimates

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Test language: Welsh (ref.: English)	−0.395*** (0.068)	−0.190*** (0.062)	−0.316*** (0.065)	—	−0.405*** (0.116)	−0.216* (0.121)	−0.293*** (0.111)
Female: yes (ref.: no)	0.219*** (0.048)	−0.192*** (0.052)	−0.091* (0.053)	−0.016 (0.028)	0.219*** (0.047)	−0.192*** (0.050)	−0.091* (0.051)
Socio-economic status quartile (ref.: first quartile, bottom)							
Fourth quartile (top)	0.630*** (0.069)	0.675*** (0.059)	0.639*** (0.066)	−0.104** (0.042)	0.631*** (0.071)	0.679*** (0.061)	0.635*** (0.068)
Third quartile	0.332*** (0.066)	0.338*** (0.065)	0.339*** (0.073)	−0.089* (0.047)	0.333*** (0.066)	0.341*** (0.064)	0.337*** (0.072)
Second quartile	0.157** (0.060)	0.131** (0.061)	0.187*** (0.058)	−0.050 (0.045)	0.157*** (0.057)	0.132** (0.059)	0.186*** (0.056)
Socio-economic status quartile. Missing flag	−0.326** (0.153)	−0.439*** (0.146)	−0.420*** (0.139)	−0.110 (0.094)	−0.324** (0.147)	−0.434*** (0.142)	−0.424*** (0.136)
Grade retention (ref.: no)							
Repeater	−0.839*** (0.191)	−0.826*** (0.202)	−0.787*** (0.180)	0.108 (0.095)	−0.840*** (0.186)	−0.830*** (0.196)	−0.784*** (0.175)
Repeater. Missing flag	0.085 (0.480)	−0.046 (0.387)	−0.162 (0.391)	0.013 (0.167)	0.084 (0.462)	−0.048 (0.372)	−0.160 (0.380)
Country of birth (ref.: other countries)							
UK	0.247 (0.236)	0.253 (0.245)	0.191 (0.265)	−0.129 (0.173)	0.248 (0.229)	0.254 (0.237)	0.189 (0.257)
Country of birth. Missing flag	0.091 (0.294)	0.056 (0.290)	0.027 (0.317)	−0.292 (0.217)	0.093 (0.288)	0.061 (0.285)	0.023 (0.310)
Father's country of birth (ref.: other countries)							
UK	−0.264** (0.131)	−0.334** (0.135)	−0.254* (0.140)	−0.073 (0.072)	−0.262** (0.126)	−0.330** (0.130)	−0.258* (0.135)

TABLE B3 (Continued)

	OLS			2SLS			
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Father's country of birth. Missing flag	−0.511*** (0.169)	−0.630*** (0.168)	−0.455** (0.174)	−0.123 (0.106)	−0.509*** (0.160)	−0.625*** (0.157)	−0.460*** (0.164)
Mother's country of birth (ref.: other countries)							
UK	−0.137 (0.111)	−0.054 (0.121)	−0.078 (0.121)	0.038 (0.079)	−0.137 (0.107)	−0.055 (0.117)	−0.077 (0.117)
Mother's country of birth. Missing flag	−0.149 (0.257)	0.018 (0.269)	−0.074 (0.258)	0.107 (0.135)	−0.150 (0.247)	0.015 (0.258)	−0.072 (0.247)
Has lived in the UK since age 6 or older: yes (ref.: no)	−0.155 (0.432)	−0.388 (0.413)	−0.181 (0.483)	0.130 (0.237)	−0.158 (0.419)	−0.394 (0.403)	−0.175 (0.467)
Term of birth (ref.: fourth term)							
First term	−0.109* (0.063)	−0.068 (0.068)	−0.085 (0.068)	0.041 (0.044)	−0.109* (0.062)	−0.069 (0.067)	−0.085 (0.066)
Second term	−0.141** (0.054)	−0.083 (0.061)	−0.105* (0.062)	0.030 (0.037)	−0.141*** (0.052)	−0.084 (0.059)	−0.105* (0.060)
Third term	−0.041 (0.059)	−0.013 (0.064)	−0.021 (0.060)	0.029 (0.039)	−0.041 (0.058)	−0.014 (0.062)	−0.020 (0.059)
School funding (ref.: public)							
Private	—	—	—	—	—	—	—
School funding. Missing flag	0.418*** (0.037)	0.450*** (0.034)	0.441*** (0.034)	−0.119*** (0.022)	−0.221*** (0.038)	−0.344*** (0.038)	−0.236*** (0.039)
PISA cycle (ref.: 2006)							
2018	0.145*** (0.031)	0.800*** (0.031)	0.675*** (0.033)	−0.298*** (0.023)	−0.128** (0.056)	−0.147** (0.061)	−0.114** (0.057)
2015	−0.443*** (0.024)	−0.141*** (0.023)	−0.196*** (0.024)	−0.404*** (0.020)	0.112*** (0.036)	0.004 (0.040)	−0.024 (0.037)
2012	0.055* (0.029)	0.352*** (0.027)	0.164*** (0.029)	−0.245*** (0.029)	0.638*** (0.054)	0.371*** (0.057)	0.319*** (0.053)

TABLE B3 (Continued)

	OLS			2SLS			
					Second stage		
Variables	**Reading**	**Mathematics**	**Science**	**First stage**	**Reading**	**Mathematics**	**Science**
2009	−0.067*** (0.016)	0.598*** (0.017)	0.364*** (0.017)	−0.743*** (0.022)	−0.168** (0.078)	−0.260*** (0.079)	−0.276*** (0.073)
Language at home (ref.: English)							
Welsh	—	—	—	0.483*** (0.059)	—	—	—
Irish	—	—	—	0.545*** (0.145)	—	—	—
Ulster Scots	—	—	—	−0.002 (0.269)	—	—	—
Other languages	—	—	—	0.315* (0.173)	—	—	—
School fixed effects	✓	✓	✓	✓	✓	✓	✓
Constant	0.020 (0.220)	−0.209 (0.226)	−0.162 (0.250)	0.014 (0.178)	−0.190 (0.212)	0.130 (0.219)	0.116 (0.241)
Observations	1502	1502	1502	1502	1502	1502	1502
R^2	0.271	0.262	0.218	0.509	0.271	0.261	0.218
Stock and Yogo (2005) test of weak instruments	—	—	—	—	37.715***	37.715***	37.715***
Wooldridge (1995) endogeneity test	—	—	—	—	0.009	0.046	0.046

Notes: Standard errors in parentheses are robust. PISA recommended practices (final student weights, balanced repeated replication weights and plausible values) have been employed (OECD, 2020). The null hypothesis of the Stock and Yogo (2005) test of weak instruments is that the instrument is weak and the null hypothesis of the Wooldridge (1995) endogeneity test is that the endogenous variable is now exogenous. The "✓" indicates that it has been controlled by school fixed effects.

Estimation method: OLS and 2SLS. The instrument is student's language at home and the other variables in the estimation.

Dependent variable: student's standardised scores in reading, mathematics and science, using Welsh mean and standard deviations in each PISA cycle.

Coefficient: ***significant at 1%, **significant at 5%, *significant at 10%.

Source: authors' own calculations.