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

Abstract

In this paper 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. Using five rounds of PISA data, 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 abilities of teenagers in Wales may be underestimated in PISA – particularly amongst those who take this test in Welsh.

Key insights:

-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.

-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.

Keywords: test language; home language; PISA; instrumental variables; Wales.

JEL Codes: I20, I21, I28, C10.

Data Availability statement: The data employed in the present research is publicly available in http://www.oecd.org/pisa/data/

Ethical Guidelines: Ethical guidelines have been followed in the present research study.

Conflict of Interest: The authors do not have any conflict of interests.

1. 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 2.1 technical standard[1]), students should take the test in the language which they are most comfortable. This is so that their scores on the test reflects their actual skill in the subject(s) being assessed, and is not unduly influenced by pupils having limited language skills[2] (OECD, 2012, pp. 369-370). In doing so, the OECD implicitly acknowledges how a 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.

Authors such as 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. 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 of French in Canada, Catalan in Spain or 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 (Welsh Assembly Government, 2007; Jones, 2016):

> (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

[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).

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 school 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.

(d) Predominantly English medium secondary school. 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 72% of secondary schools in Wales are English-medium only, though there is significant regional variation. For instance, whereas 90% of schools are English medium in South East Wales, this falls to around 50% in the North, where Welsh-medium (or Bilingual with a strong emphasis on Welsh) education is much more prevalent.

-Insert Table 1 here-

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, Hillary and Cullinane (2019) illustrate how the average rate of Free School Meal eligibility in English-medium schools is 21%, which is notably higher than in Bilingual (14%) and Welsh-medium (10%) schools. This might suggest that Welsh-medium schools attract this particular profile of students due to the bilingual education they offer, a feature which has been found to be beneficial to students' learning (Edwards & Newcombe, 2006; Jones, 2017). 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 to 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). Thus, the aim of this paper is to quantifying the gap in achievement between students who take the PISA test in English and Welsh. Specifically, we aim to address the following research question:

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

To do so, 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 as well as 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, in addition to ordinary least squares regression models.

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.

2. Data

2.1. 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 5 options: "English", "Welsh", "Irish", "Ulster Scots" and "Other languages". Other relevant variables for our analysis are pupil, father and mother regions of birth, which include Germany, India, Ireland, Pakistan, Poland, United Kingdom (England), and "other".

2.2. 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).

-Insert Table 2 here-

Table 3 illustrates how almost all pupils who took the PISA test in English also speak 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, 2017). 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 (to strongly encouraged) to take the test in Welsh. Yet this would be a violation of PISA technical standard 2.1 – which stipulates how pupils should take the test in the language that they are most comfortable with. Indeed, from a total of 607 school-by-PISA-cycle observations in our data, all pupils took the test in

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

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 the pupils took the test in Welsh suggests that students may not have had complete freedom in choosing the language to take the test.

-Insert Table 3 here-

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. For instance, Welsh-language test takers were significantly more likely to be in the top socio-economic status quartile, and have a mother and father who hold a degree level (ISCED 5A) qualification, than their English-language peers. Table 4 hence clearly illustrates how there are some important observable differences between students who took the test in English and students who took the test in Welsh. In the following section (Methodology) we discuss how we use regression analyses and an instrumental variable approach to attempt to control for such differences between these groups.

-Insert Table 4 here-

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 in reading than those who took the test in English. The difference is smaller, though still non-trivial, in mathematics (14 points) and science (27 points).

-Insert Table 5 here-

3. Methodology

3.1. Ordinary Least Squares

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

$$C_{ijc} = \alpha + \beta WT_{ijc} + \gamma X_{ijc} + \delta F_{ijc} + \vartheta SCH_{jc} + \rho PISA_c + \varepsilon_{ijc} \qquad (1)$$

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 United Kingdom 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 (school funding); $PISA_c$ controls for PISA cycle; 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,

[4] This variable controls the potential differences in language skills between those students who arrived at the United Kingdom 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.

controlling for a wide 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 due to potential unobservables included in ε_{ijc} which have been omitted from our model (for instance, students' ability). This omitted variable problem has been highlighted by many authors when dealing with observational cross-sectional data such as PISA (Hanchane & Mostafa, 2010; Micklewright, Schnepf, & Silva, 2012; Lounkaew, 2013; Cordero and Pedraja, 2018). As an illustrative example of this problem, one could imagine that some schools might force pupils to take the test in Welsh, regardless of the language they most regularly speak at home (which seems to happens in practice given how "*pupils in Wales were assigned assessments and questionnaires according to the relevant language of instruction* Sizmur et al 2019, p. 199). Focusing on students who speak English at home in these schools, there may be: (a) those who have strong Welsh skills (who might be more generally more able students) and thus perform better on the PISA test and (b) those who do not have strong Welsh skills (who might be less able students) and thus perform lower. Therefore, the OLS estimate of β might be biased due to us not being able to control for this student ability variable. The direction of this bias could either be positive or negative, depending on the proportion of pupils with strong and weak Welsh skills in these schools. We hence employ an instrumental variable approach to try and overcome this problem, which is implemented via two-stage least squares (2SLS).

3.2. 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 effect 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}) and uncorrelated with the error term (ε_{ijc}). 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:

$$WT_{ijc} = \pi_0 + \pi_1 Z_{ijc} + \pi_2 X_{ijc} + \pi_3 F_{ijc} + \pi_4 SCH_{jc} + \pi_5 PISA_c + \omega_{ijc} \qquad (2)$$

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

$$C_{ijc} = \alpha + \beta \widehat{WT}_{ijc} + \gamma X_{ijc} + \delta F_{ijc} + \vartheta SCH_{jc} + \rho PISA_c + \varepsilon_{ijc} \qquad (3)$$

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[6] 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 to ensure standard errors are estimated correctly (as suggested by authors such as Andrews, Stock, & Sun, 2019; Angrist & Psichke, 2008). Note that these standard errors are also "robust" to correct for potential heteroskedasticity.

4. Results

Table 5 presents three sets of results. First, an OLS model where no other variables are controlled. Second, estimates from an OLS model including controls (with the full set of parameter estimates presented in Appendix B – Table B1). 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 to 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* (further details are again provided in Appendix A). Furthermore, the Wooldridge (1995) endogeneity 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.

The second stage of our instrumental variable estimates produce 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. 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 socio-economic status achievement gap. This is hence clearly a very sizeable effect.

-Insert Table 5-

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 6, with the full set of parameter estimates provided in Appendix B (Table B2).

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

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 7 we replicate the analysis again, but now also additionally including school fixed-effects (see Appendix B – Table B3 – for the full set of parameter estimates). 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 parental selection of secondary schools (and, in particular, parental selection of English/Welsh-medium education). 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).

-Insert Table 6-

-Insert Table 7-

5. 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 variables approach applied to 5 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[7]) lower in reading, 0.26 SDs (26 PISA points) lower in mathematics and 0.33 SDs (33 PISA points) lower in science, compared to 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, though 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 conditional associations. Second, 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 affect 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, Goldstein, and Guérin-Pace (2001) noting how such problems have affected the validity of other cross-national studies (e.g. the International Adult Literacy Survey - IALS). Indeed, we note how the PISA 2018 technical report (OECD 2019: 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

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

population of the UK. We therefore recommend that, in the future, there is greater independent verification of the Welsh versions of the PISA test.

Yet we also do not believe that translation issues are likely to be the major driving force behind out results. Minor differences between source versions and national translations have been found elsewhere in the literature (e.g. Murat and Rocher (2004) – for PISA 2000 data – and Grisay, de Jong, Gebhardt, Berezner, and Halleux-Monseur (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 that for the other languages in this country (Galician, Catalan and Basque) were equivalent and would have not had a substantial impact upon the results.

Hence we believe the most likely explanation for our findings 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 being fully applied. This could, in turn, mean that the academic abilities of Welsh 15-year-olds are being underestimated in PISA, due 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. This means that all children who take the PISA test in Wales should get a 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 .

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. doi: 10.1146/annurev-economics-080218-025643

Angrist, J. D., & Psichke, J.-S. (2008). *Mostly Harmless Econometrics. An Empiricist's Companion*. New Jersey: Princeton University Press.

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., & Yu, G. (2011). *Policy Effects of PISA*. Oxford: Oxford University Centre for Educational Assessment, 1–38.

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. doi: 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. doi: 10.1080/09695940123977

Bradshaw, J., Ager, R., Burge, B., & Wheater, R. (2010). *PISA 2009: Achievement of 15-year-olds in Wales*. Slough: National Foundation for Educational Research.

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance*. OECD Education Working Papers, No. 71, 1-31. doi: 10.1787/5k9fdfqffr28-en

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. doi: 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. doi: 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). Clevedon: Multilingual Matters.

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.

Froese-Germain, B. (2010). *The OECD, PISA and the Impacts on Educational Policy*. Canadian Teacher's Federation: Virtual Research Centre, 1–31.

Gorard, S. (1998). Four errors ... and a conspiracy? The effectiveness of schools in Wales. *Oxford Review of Education, 24*(4), 459–472. doi: 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*. London: 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, M. (2016). *Research Briefing. Welsh-medium education and Welsh as a subject*. Wales: National Assembly for Wales. Research Service.

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

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.

Lounkaew, K. (2013). Explaining urban–rural differences in educational achievement in Thailand: Evidence from PISA literacy data. *Economics of Education Review, 37*, 213–225. doi: 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. doi: 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. doi: 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). London: Routledge Farmer.

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

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

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

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

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

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

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. doi: 10.7334/psicothema2013.46

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

Rees, G., & Taylor, C. M. (2015). *Is there a 'crisis' in Welsh education?* Transactions of the Honourable Society of Cymmrodorion, 97–113.

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*. Slough: NFER.

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. Retrieved from 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. Last time accessed: July 2021.

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

Welsh Assembly Government (2007). *Defining schools according to Welsh medium provision*. Wales: 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, and P. C. B. Phillips (Eds.), *Advances*

in Econometrics and Quantitative Economics: Essays in Honor of Profesor C. R. Rao (pp. 66–87). Oxford: Blackwell.

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 2, 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 stablishes that the instrument is randomly assigned or "as good as randomly assigned", meaning that it is uncorrelated to 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; (b) in plurilingual households, once all 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* (Fiorini & Stevens, 2014; Barua & Lang, 2016; Dhuey, Figlio, Karbownik, & Roth, 2019). 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

-Insert Table B1-

-Insert Table B2-

-Insert Table B3-

Table 1. The distribution of secondary school types across Wales by area

	Welsh medium	Bilingual - AB	Bilingual - BB / CB	English with significant Welsh	English medium	Total
North Wales	**5**	**10**	**7**	**2**	**26**	**50**
Isle of Anglesey	.	.	4	1	.	5
Gwynedd	1	10	.	1	.	12
Conwy	1	.	1	.	5	7
Denbighshire	1	.	2	.	3	6
Flintshire	1	.	.	.	10	11
Wrexham	1	.	.	.	8	9
South West and Mid Wales	**4**	**4**	**6**	**5**	**32**	**51**
Powys	.	1	2	2	5	10
Ceredigion	.	1	2	.	1	4
Pembrokeshire	1	.	.	1	4	6
Carmarthenshire	1	2	2	2	5	12
Swansea	2	.	.	.	12	14
Neath Port Talbot	.	.	.	.	5	5
Central South Wales	**6**	**.**	**.**	**.**	**44**	**50**
Bridgend	1	.	.	.	8	9
Vale of Glamorgan	.	.	.	.	7	7
Rhondda Cynon Taf	2	.	.	.	10	12
Merthyr Tydfil	.	.	.	.	4	4
Cardiff	3	.	.	.	15	18
South East Wales	**3**	**.**	**.**	**.**	**29**	**32**
Caerphilly	1	.	.	.	10	11
Blaenau Gwent	.	.	.	.	2	2
Torfaen	1	.	.	.	5	6
Monmouthshire	.	.	.	.	4	4
Newport	1	.	.	.	8	9
All schools	**18**	**14**	**13**	**7**	**131**	**183**

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

Table 2. Percentage of students who took the test in each language and who speak each language at home in Wales

	Language of the test						Language at home					
Language	**2006**	**2009**	**2012**	**2015**	**2018**	**All cycles**	**2006**	**2009**	**2012**	**2015**	**2018**	**All cycles**
English	87	87	87	90	86	87	92	91	92	90	89	91
Welsh	13	13	13	10	15	13	7	6	6	6	6	6
Irish	-	-	-	-	-	-	0.1<	0.1<	0.1<	0.2	0.1<	0.1
Ulster Scots	-	-	-	-	-	-	-	0.1<	0.1<	0.2	0.1	0.1<
Other languages	-	-	-	-	-	-	1.0	2	2	4	5	3

Notes: Notes: All OECD recommended practices (final student weights and BRR weights) have been employed. The "-" indicates that there are not data for that region in the PISA cycle.
Source: Authors' own calculations.

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

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

Notes: All OECD recommended practices (final student weights and BRR weights) have been employed.
Source: Authors' own calculations.

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		**Obs.**	**%**	**Obs.**	**%**
Sex of the student	**Male**	13,467	50	1,942	50
	Female	13,467	50	1,942	50
Socio-economic status quartile	**Fourth quartile (Top)**	12,936	21[D]	1,824	31[D]
	Third quartile	12,936	24	1,824	27
	Second quartile	12,936	27[D]	1,824	24[D]
	First quartile (bottom)	12,936	28[D]	1,824	18[D]
Father level of education	**None**	11,391	1	1,547	2
	ISCED 1	11,391	1	1,547	1
	ISCED 2	11,391	9[D]	1,547	6[D]
	ISCED 3b, c	11,391	28[D]	1,547	23[D]
	ISCED 3a, 4	11,391	19[D]	1,547	13[D]
	ISCED 5b	11,391	18	1,547	18
	ISCED 5a, 6	11,391	24[D]	1,547	37[D]
Mother level of education	**None**	12,164	1	1,650	1
	ISCED 1	12,164	1	1,650	1
	ISCED 2	12,164	3[D]	1,650	2[D]
	ISCED 3b, c	12,164	27[D]	1,650	21[D]
	ISCED 3a, 4	12,164	21[D]	1,650	13[D]
	ISCED 5b	12,164	23	1,650	22
	ISCED 5a, 6	12,164	24[D]	1,650	40[D]
Number of books at home	**0 to 10 books**	13,081	17[D]	1,849	13[D]
	11 to 25 books	13,081	18[D]	1,849	15[D]
	26 to 100 books	13,081	30	1,849	31
	101 to 200 books	13,081	16	1,849	18
	201 to 500 books	13,081	12[D]	1,849	15[D]
	More than 500 books	13,081	7	1,849	8
Term of birth	**First term**	13,467	25	1,942	24
	Second term	13,467	25	1,942	25
	Third term	13,467	25	1,942	24
	Fourth term	13,467	25	1,942	26

Notes: All OECD recommended practices (final student weights and BRR weights) have been employed. "Obs." stands for "Observations". 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 if there are significant differences between those 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.

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

	Reading			Mathematics			Science		
	OLS (no controls)	**OLS (with controls)**	**2SLS**	**OLS (no controls)**	**OLS (with controls)**	**2SLS**	**OLS (no controls)**	**OLS (with controls)**	**2SLS**
Effect of taking test in Welsh (compared to English)	-0.341***	-0.421***	-0.392***	-0.106***	-0.183***	-0.262***	-0.282***	-0.363***	-0.328***
Standard error	0.042	0.036	0.074	0.042	0.034	0.069	0.039	0.034	0.063
Stock and Yogo (2005) test of weak instruments	-	-	63.774***	-	-	63.774***	-	-	63.774***
Wooldridge (1995) endogeneity test	-	-	0.226	-	-	1.88	-	-	0.38

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 students' language at home and the rest of variables in the estimation.
Dependent variable: Students' 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 6. . 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.

	Reading		Mathematics		Science	
	OLS	**2SLS**	**OLS**	**2SLS**	**OLS**	**2SLS**
Effect of taking test in Welsh (compared to English)	-0.417***	-0.383***	-0.256***	-0.321***	-0.360***	-0.340***
Standard error	0.068	0.109	0.060	0.105	0.057	0.094
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: 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 1,502 students.
Estimation method: Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS). The instrument is students' language at home and the other variables in the estimation.
Dependent variable: Students' 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 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 and school fixed-effectsincluded.

	Reading		Mathematics		Science	
	OLS	**2SLS**	**OLS**	**2SLS**	**OLS**	**2SLS**
Effect of taking test in Welsh (compared to English)	-0.395***	-0.405***	-0.190***	-0.216*	-0.316***	-0.293***
Standard error	0.068	0.116	0.062	0.121	0.065	0.111
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: 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 1,502 students.
Estimation method: Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS). The instrument is students' language at home and the other variables in the estimation.
Dependent variable: Students' 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 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***	-0.183***	-0.363***	-	-0.392***	-0.262***	-0.328***
	(0.036)	(0.034)	(0.034)		(0.074)	(0.069)	(0.063)
Female: Yes (Ref.: no)	0.244***	-0.165***	-0.088***	-0.004	0.244***	-0.165***	-0.088***
	(0.015)	(0.015)	(0.016)	(0.004)	(0.015)	(0.015)	(0.016)
Socio-economic status quartile (Ref.: first quartile)							
Fourth quartile	0.754***	0.807***	0.798***	-0.063***	0.751***	0.815***	0.794***
	(0.025)	(0.024)	(0.026)	(0.011)	(0.026)	(0.025)	(0.027)
Third quartile	0.398***	0.432***	0.434***	-0.042***	0.397***	0.435***	0.432***
	(0.022)	(0.021)	(0.022)	(0.008)	(0.022)	(0.021)	(0.022)
Second quartile	0.242***	0.245***	0.243***	-0.023***	0.241***	0.247***	0.242***
	(0.020)	(0.020)	(0.020)	(0.007)	(0.020)	(0.020)	(0.020)
Socio-economic status quartile. Missing flag	-0.389***	-0.334***	-0.349***	-0.051***	-0.391***	-0.329***	-0.351***
	(0.062)	(0.059)	(0.062)	(0.018)	(0.062)	(0.059)	(0.062)
Grade retention (Ref.: no)							
Repeater	-0.674***	-0.706***	-0.627***	0.041***	-0.673***	-0.710***	-0.626***
	(0.056)	(0.051)	(0.054)	(0.015)	(0.056)	(0.051)	(0.054)
Repeater. Missing flag	-0.555***	-0.451***	-0.576***	0.036	-0.555***	-0.452***	-0.575***
	(0.125)	(0.121)	(0.128)	(0.033)	(0.124)	(0.122)	(0.127)
Country of birth (Ref.: Other countries)							
United Kingdom	0.089	0.089	0.030	-0.010	0.089	0.090	0.029
	(0.058)	(0.057)	(0.056)	(0.013)	(0.057)	(0.057)	(0.056)
Country of birth. Missing flag	-0.302***	-0.312***	-0.380***	0.005	-0.302***	-0.312***	-0.379***
	(0.083)	(0.081)	(0.082)	(0.023)	(0.083)	(0.081)	(0.082)
Father's country of birth (Ref.: Other countries)							
United Kingdom	0.016	-0.014	0.025	-0.026	0.014	-0.010	0.023
	(0.036)	(0.038)	(0.038)	(0.017)	(0.036)	(0.039)	(0.038)
Father's country of birth. Missing flag	-0.172**	-0.218***	-0.159**	-0.036	-0.174**	-0.212***	-0.161**
	(0.069)	(0.067)	(0.069)	(0.025)	(0.069)	(0.067)	(0.069)
Mother's country of birth (Ref.: Other countries)							
United Kingdom	-0.113***	-0.133***	-0.109***	-0.035**	-0.114***	-0.129***	-0.111***
	(0.037)	(0.039)	(0.039)	(0.015)	(0.037)	(0.039)	(0.039)
Mother's country of birth. Missing flag	-0.256***	-0.330***	-0.284***	-0.023	-0.258***	-0.324***	-0.286***
	(0.093)	(0.093)	(0.092)	(0.033)	(0.093)	(0.093)	(0.092)

Lives in the United Kingdom since age 6 or older: yes (Ref.: no)	-0.216***	-0.171**	-0.229***	0.018	-0.216***	-0.172**	-0.228***
	(0.081)	(0.074)	(0.081)	(0.016)	(0.081)	(0.074)	(0.080)
Term of birth (Ref.: Fourth term)							
First term	-0.058***	-0.035*	-0.040**	0.009	-0.058***	-0.036*	-0.040**
	(0.020)	(0.020)	(0.020)	(0.006)	(0.020)	(0.020)	(0.020)
Second term	-0.142***	-0.109***	-0.111***	0.007	-0.142***	-0.109***	-0.111***
	(0.021)	(0.020)	(0.021)	(0.006)	(0.021)	(0.020)	(0.021)
Third term	-0.051**	-0.024	-0.035	0.010	-0.051**	-0.025	-0.034
	(0.021)	(0.021)	(0.022)	(0.006)	(0.021)	(0.021)	(0.022)
School funding (Ref.: public)							
Private	0.543***	0.614***	0.557***	0.110***	0.547***	0.602***	0.562***
	(0.099)	(0.096)	(0.085)	(0.016)	(0.099)	(0.096)	(0.086)
School funding. Missing flag	-0.133**	-0.097**	-0.074	0.013	-0.133**	-0.098**	-0.073
	(0.060)	(0.048)	(0.053)	(0.035)	(0.060)	(0.048)	(0.053)
PISA cycle (Ref.: 2006)							
2018	0.043	0.015	0.025	-0.025	0.043	0.017	0.024
	(0.046)	(0.040)	(0.041)	(0.036)	(0.046)	(0.040)	(0.041)
2015	0.050	0.042	0.029	0.010	0.050	0.041	0.030
	(0.046)	(0.038)	(0.041)	(0.030)	(0.046)	(0.038)	(0.041)
2012	0.024	0.024	0.018	-0.009	0.024	0.024	0.018
	(0.048)	(0.039)	(0.044)	(0.032)	(0.048)	(0.040)	(0.044)
2009	0.027	0.022	0.021	-0.030	0.027	0.024	0.020
	(0.045)	(0.041)	(0.043)	(0.039)	(0.045)	(0.042)	(0.043)
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-squared	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 parenthesis 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: Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS). The instrument is students' language at home and the rest of variables in the estimation.
Dependent variable: Students' 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, schools with English and Welsh languages, effect sizes. Full estimates

	OLS				2SLS		
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Test language: Welsh (Ref.: English)	-0.417***	-0.256***	-0.360***	-	-0.383***	-0.321***	-0.340***
	(0.068)	(0.060)	(0.057)		(0.109)	(0.105)	(0.094)
Female: Yes (Ref.: no)	0.226***	-0.192***	-0.085	-0.015	0.226***	-0.192***	-0.085*
	(0.048)	(0.051)	(0.052)	(0.024)	(0.047)	(0.050)	(0.051)
Socio-economic status quartile (Ref.: first quartile, bottom)							
Fourth quartile (top)	0.689***	0.766***	0.703***	-0.107***	0.684***	0.775***	0.700***
	(0.062)	(0.058)	(0.060)	(0.040)	(0.063)	(0.059)	(0.062)
Third quartile	0.360***	0.400***	0.387***	-0.066*	0.359***	0.403***	0.386***
	(0.064)	(0.068)	(0.073)	(0.037)	(0.063)	(0.067)	(0.072)
Second quartile	0.152**	0.155**	0.206***	-0.036	0.152***	0.156**	0.205***
	(0.059)	(0.064)	(0.060)	(0.040)	(0.058)	(0.062)	(0.059)
Socio-economic status quartile. Missing flag	-0.307**	-0.490***	-0.407***	-0.036	-0.312**	-0.481***	-0.410***
	(0.137)	(0.133)	(0.129)	(0.092)	(0.134)	(0.132)	(0.128)
Grade retention (Ref.: no)							
Repeater	-0.915***	-0.839***	-0.806***	0.057	-0.911***	-0.846***	-0.804***
	(0.178)	(0.193)	(0.172)	(0.058)	(0.176)	(0.192)	(0.169)
Repeater. Missing flag	0.146	0.064	-0.114	0.012	0.148	0.059	-0.113
	(0.520)	(0.422)	(0.410)	(0.180)	(0.516)	(0.406)	(0.406)
Country of birth (Ref.: Other countries)							
United Kingdom	0.312	0.259	0.214	-0.095	0.311	0.262	0.213
	(0.229)	(0.228)	(0.251)	(0.119)	(0.225)	(0.225)	(0.246)
Country of birth. Missing flag	0.154	0.116	0.095	-0.263	0.150	0.124	0.092
	(0.305)	(0.288)	(0.308)	(0.171)	(0.300)	(0.287)	(0.303)
Father's country of birth (Ref.: Other countries)							
United Kingdom	-0.253**	-0.342***	-0.253*	-0.099	-0.259**	-0.331***	-0.256*
	(0.124)	(0.127)	(0.136)	(0.064)	(0.123)	(0.126)	(0.135)
Father's country of birth. Missing flag	-0.487***	-0.634***	-0.456***	-0.139	-0.495***	-0.619***	-0.461***
	(0.152)	(0.167)	(0.166)	(0.087)	(0.147)	(0.159)	(0.160)
Mother's country of birth (Ref.: Other countries)							
United Kingdom	-0.163	-0.057	-0.085	-0.013	-0.163	-0.056	-0.085
	(0.103)	(0.111)	(0.112)	(0.078)	(0.102)	(0.108)	(0.111)
Mother's country of birth. Missing flag	-0.139	0.038	-0.015	0.041	-0.138	0.036	-0.014
	(0.234)	(0.264)	(0.243)	(0.122)	(0.231)	(0.257)	(0.239)

Lives in the United Kingdom since age 6 or older: yes (Ref.: no)	-0.005	-0.327	-0.198	0.183	0.003	-0.343	-0.194
	(0.398)	(0.419)	(0.483)	(0.157)	(0.390)	(0.413)	(0.475)
Term of birth (Ref.: Fourth term)							
First term	-0.125*	-0.060	-0.072	0.044	-0.124*	-0.062	-0.071
	(0.066)	(0.069)	(0.068)	(0.038)	(0.066)	(0.069)	(0.068)
Second term	-0.135**	-0.062	-0.090	0.043	-0.135***	-0.064	-0.089
	(0.053)	(0.058)	(0.059)	(0.032)	(0.052)	(0.057)	(0.057)
Third term	-0.026	0.015	0.006	0.033	-0.025	0.013	0.007
	(0.055)	(0.062)	(0.057)	(0.031)	(0.056)	(0.061)	(0.057)
School funding (Ref.: public)							
Private	-	-	-	-	-	-	-
School funding. Missing flag	0.006	0.001	0.102	-0.122	0.002	0.008	0.100
	(0.132)	(0.137)	(0.137)	(0.097)	(0.135)	(0.130)	(0.139)
PISA cycle (Ref.: 2006)							
2018	-0.256*	-0.184	-0.079	-0.051	-0.257**	-0.183	-0.080
	(0.128)	(0.118)	(0.088)	(0.093)	(0.126)	(0.115)	(0.087)
2015	-0.039	-0.010	-0.001	-0.046	-0.040	-0.008	-0.002
	(0.116)	(0.107)	(0.083)	(0.086)	(0.112)	(0.106)	(0.082)
2012	0.159	0.207*	0.109	0.056	0.161	0.204*	0.110
	(0.114)	(0.107)	(0.098)	(0.080)	(0.111)	(0.105)	(0.097)
2009	-0.081	-0.051	-0.039	-0.014	-0.080	-0.053	-0.038
	(0.111)	(0.114)	(0.080)	(0.098)	(0.108)	(0.110)	(0.079)
Language at home (Ref.: English)							
Welsh	-	-	-	0.537***	-	-	-
				(0.047)			
Irish	-	-	-	0.351	-	-	-
				(0.272)			
Ulster Scots	-	-	-	0.135	-	-	-
				(0.148)			
Other languages	-	-	-	0.326**	-	-	-
				(0.154)			
Constant	-0.052	0.109	0.018	0.031	-0.060	0.124	0.014
	(0.251)	(0.244)	(0.243)	(0.130)	(0.247)	(0.238)	(0.238)
Observations	1,502	1,502	1,502	1,502	1,502	1,502	1,502
R-squared	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 parenthesis 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: Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS). The instrument is students' language at home and the other variables in the estimation.
Dependent variable: Students' 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, schools with English and Welsh language with school fixed-effects, effect sizes. Full estimates

	OLS				2SLS		
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Test language: Welsh (Ref.: English)	-0.395***	-0.190***	-0.316***	-	-0.405***	-0.216*	-0.293***
	(0.068)	(0.062)	(0.065)		(0.116)	(0.121)	(0.111)
Female: Yes (Ref.: no)	0.219***	-0.192***	-0.091*	-0.016	0.219***	-0.192***	-0.091*
	(0.048)	(0.052)	(0.053)	(0.028)	(0.047)	(0.050)	(0.051)
Socio-economic status quartile (Ref.: first quartile, bottom)							
Fourth quartile (top)	0.630***	0.675***	0.639***	-0.104**	0.631***	0.679***	0.635***
	(0.069)	(0.059)	(0.066)	(0.042)	(0.071)	(0.061)	(0.068)
Third quartile	0.332***	0.338***	0.339***	-0.089*	0.333***	0.341***	0.337***
	(0.066)	(0.065)	(0.073)	(0.047)	(0.066)	(0.064)	(0.072)
Second quartile	0.157**	0.131**	0.187***	-0.050	0.157***	0.132**	0.186***
	(0.060)	(0.061)	(0.058)	(0.045)	(0.057)	(0.059)	(0.056)
Socio-economic status quartile. Missing flag	-0.326**	-0.439***	-0.420***	-0.110	-0.324**	-0.434***	-0.424***
	(0.153)	(0.146)	(0.139)	(0.094)	(0.147)	(0.142)	(0.136)
Grade retention (Ref.: no)							
Repeater	-0.839***	-0.826***	-0.787***	0.108	-0.840***	-0.830***	-0.784***
	(0.191)	(0.202)	(0.180)	(0.095)	(0.186)	(0.196)	(0.175)
Repeater. Missing flag	0.085	-0.046	-0.162	0.013	0.084	-0.048	-0.160
	(0.480)	(0.387)	(0.391)	(0.167)	(0.462)	(0.372)	(0.380)
Country of birth (Ref.: Other countries)							
United Kingdom	0.247	0.253	0.191	-0.129	0.248	0.254	0.189
	(0.236)	(0.245)	(0.265)	(0.173)	(0.229)	(0.237)	(0.257)
Country of birth. Missing flag	0.091	0.056	0.027	-0.292	0.093	0.061	0.023
	(0.294)	(0.290)	(0.317)	(0.217)	(0.288)	(0.285)	(0.310)
Father's country of birth (Ref.: Other countries)							
United Kingdom	-0.264**	-0.334**	-0.254*	-0.073	-0.262**	-0.330**	-0.258*
	(0.131)	(0.135)	(0.140)	(0.072)	(0.126)	(0.130)	(0.135)
Father's country of birth. Missing flag	-0.511***	-0.630***	-0.455**	-0.123	-0.509***	-0.625***	-0.460***
	(0.169)	(0.168)	(0.174)	(0.106)	(0.160)	(0.157)	(0.164)
Mother's country of birth (Ref.: Other countries)							
United Kingdom	-0.137	-0.054	-0.078	0.038	-0.137	-0.055	-0.077
	(0.111)	(0.121)	(0.121)	(0.079)	(0.107)	(0.117)	(0.117)
Mother's country of birth. Missing flag	-0.149	0.018	-0.074	0.107	-0.150	0.015	-0.072
	(0.257)	(0.269)	(0.258)	(0.135)	(0.247)	(0.258)	(0.247)

Lives in the United Kingdom since age 6 or older: yes (Ref.: no)	-0.155	-0.388	-0.181	0.130	-0.158	-0.394	-0.175
	(0.432)	(0.413)	(0.483)	(0.237)	(0.419)	(0.403)	(0.467)
Term of birth (Ref.: Fourth term)							
First term	-0.109*	-0.068	-0.085	0.041	-0.109*	-0.069	-0.085
	(0.063)	(0.068)	(0.068)	(0.044)	(0.062)	(0.067)	(0.066)
Second term	-0.141**	-0.083	-0.105*	0.030	-0.141***	-0.084	-0.105*
	(0.054)	(0.061)	(0.062)	(0.037)	(0.052)	(0.059)	(0.060)
Third term	-0.041	-0.013	-0.021	0.029	-0.041	-0.014	-0.020
	(0.059)	(0.064)	(0.060)	(0.039)	(0.058)	(0.062)	(0.059)
School funding (Ref.: public)							
Private	-	-	-	-	-	-	-
School funding. Missing flag	0.418***	0.450***	0.441***	-0.119***	-0.221***	-0.344***	-0.236***
	(0.037)	(0.034)	(0.034)	(0.022)	(0.038)	(0.038)	(0.039)
PISA cycle (Ref.: 2006)							
2018	0.145***	0.800***	0.675***	-0.298***	-0.128**	-0.147**	-0.114**
	(0.031)	(0.031)	(0.033)	(0.023)	(0.056)	(0.061)	(0.057)
2015	-0.443***	-0.141***	-0.196***	-0.404***	0.112***	0.004	-0.024
	(0.024)	(0.023)	(0.024)	(0.020)	(0.036)	(0.040)	(0.037)
2012	0.055*	0.352***	0.164***	-0.245***	0.638***	0.371***	0.319***
	(0.029)	(0.027)	(0.029)	(0.029)	(0.054)	(0.057)	(0.053)
2009	-0.067***	0.598***	0.364***	-0.743***	-0.168**	-0.260***	-0.276***
	(0.016)	(0.017)	(0.017)	(0.022)	(0.078)	(0.079)	(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	-0.209	-0.162	0.014	-0.190	0.130	0.116
	(0.220)	(0.226)	(0.250)	(0.178)	(0.212)	(0.219)	(0.241)
Observations	1,502	1,502	1,502	1,502	1,502	1,502	1,502
R-squared	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 parenthesis 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: Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS). The instrument is students' language at home and the other variables in the estimation.
Dependent variable: Students' 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.

Online Supplementary Material

Table S1. Influence of taking the test in Welsh on students' competences in Wales, effect sizes. Test language control

	OLS				2SLS		
					Second stage		
Variables	Reading	Mathematics	Science	First stage	Reading	Mathematics	Science
Test language: Welsh (Ref.: English)	-0.341***	-0.106**	-0.282***	-	-0.241***	-0.118	-0.176**
	(0.042)	(0.042)	(0.039)		(0.088)	(0.082)	(0.076)
Language at home (Ref.: English)							
Welsh	-	-	-	0.603***	-	-	-
				(0.037)			
Irish	-	-	-	-0.044	-	-	-
				(0.045)			
Ulster Scots	-	-	-	0.006	-	-	-
				(0.088)			
Other languages	-	-	-	-0.071***	-	-	-
				(0.012)			
Constant	0.058***	0.026	0.049***	0.088***	0.046**	0.028	0.036*
	(0.018)	(0.017)	(0.018)	(0.011)	(0.020)	(0.019)	(0.020)
Observations	14,951	14,951	14,951	14,951	14,951	14,951	14,951
R-squared	0.014	0.001	0.009	0.197	0.013	0.001	0.008
Stock and Yogo (2005) test of weak instruments	-	-	-	-	79.316***	79.316***	79.316***
Wooldridge (1995) endogeneity test	-	-	-	-	2.061	0.034	2.892

Notes: Standard errors in parenthesis 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: Ordinary Least Squares (OLS) and Two-Stage Least Squares (2SLS). The instrument is students' language at home and the rest of variables in the estimation.
Dependent variable: Students' 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.