PISA trends, social changes and education reforms

Cesare Aloisi[a] and Peter Tymms[b]

[a]*Centre for Education Research and Practice, AQA, Manchester, UK. Formerly at Planning and Strategy Office, University of Reading, Reading, UK. ORCiD: https://orcid.org/0000-0002-7151-7379*

caloisi@aqa.org.uk

[b]*School of Education, Durham University, Durham, UK.* https://orcid.org/0000-0002-7170-2566

PISA trends, social changes and education reforms

The stability of educational test results from PISA over 15 years was examined and the influence of demographics and social capital was assessed, as was the impact of educational reforms. The test results were remarkably stable, with correlations up to 0.99 for country-level results over two cycles. Despite this stability, trends were observed with scores generally rising year on year, but so too were the socio-economic indicators. Together with measures of gender, immigration, OECD membership and first language, these variables went a long way to account for the rising scores. Case studies suggest that the clearest impact of reforms on test scores amount to an annual Effect Size of around 0.02. The paper argues for the greater prominence of fairer adjusted PISA league tables and suggests that multi-disciplinary approaches to educational data analyses and policy advice are needed.

Keywords: PISA, socio-economic status, standards, international, OECD, literacy

Introduction

This study seeks to contrast the ability of policy-malleable variables to affect PISA scores with that of non-policy-malleable variables. Recent research has challenged the optimistic view, promoted by international organisations such as the OECD, whereby all countries can improve their educational outcomes by borrowing "successful" education policies from top-scoring jurisdictions. Scheerens, Luyten, van den Berg, & Glas (2015) showed that student socio-economic status and school composition were more strongly related to PISA 2009 outcomes than several system-level variables. The authors framed their findings within Weick's (1976) conceptualisation of education systems being 'loosely coupled'. Loose coupling 'suggests that shared expectations do not automatically mean shared action or implementation of these expectations at either the macro or micro levels' (Wiseman & Chase-Mayoral, 2014, p. 107). There is general agreement among school effectiveness and econometric studies that both

student-level factors and some structural features of education systems (such as streaming pupils by perceived ability or curriculum) matter for educational outcomes (Creemers & Kyriakides, 2015; Hanushek & Woessmann, 2011). However, there is more disagreement when it comes to composite indicators such as school socio-economic composition (Harker & Tymms, 2004; Timmermans & Thomas, 2015).

This article tackles the issue of system coupling and policy-malleable versus non-malleable determinants of educational achievement starting with the relationship between global trends in PISA and two other trends: the change in the socio-economic and demographic characteristics of PISA students, and the changes in country school curricula. It is theoretically situated within educational policy effectiveness research and it aims to provide additional empirical evidence to the field.

The article is structured as follows. The literature on PISA is reviewed along the two dimensions of the PISA population composition and curricular effectiveness. The first dimension is analysed in terms of the variables the OECD uses to account for societal changes: socio-economic status, immigration status and first language. School grade and student sex are also known to correlate with achievement, but the country-level gender balance tends to be rather stable over time; therefore, the statistical analyses below control for student sex but the literature review focuses on grade. After defining the research questions, the methodology is presented in detail. Particular attention is given to the measurement of grade and curricular changes, and to the method of separating the specific effect of the variables of interest from general performance growth. The main analysis, which employs multilevel growth models, follows. More qualitative evidence and case studies are used to study curricular effects. After a critical discussion of the findings, some general conclusions are drawn.

Literature review

The relationship between socio-economic and demographic factors and achievement

Fifty years have passed since the Coleman Report (Coleman, 1966), which provided a snapshot of the (in)equalities of opportunities in the US and showed that there was a clear connection between the socio-economic and racial characteristics of students and their achievement. This is now an uncontroversial statement; therefore, this section focuses only on the link between certain student characteristics and PISA outcomes, and why it is thought that changes in student demographics might affect country performance.

The factors under consideration are primarily socio-economic status (SES), cultural capital, immigration status and first language, as they are used by the OECD to adjust PISA scores. The possible impact of student grade on PISA is also reviewed.

Socio-economic status

> Across OECD countries, a student from a more socio-economically advantaged background (among the top one seventh) outperforms a student from an average background by 38 score points, or about one year's worth of education, in reading. (OECD, 2010, p. 13)

Socio-economic status is broadly an index of access or control over some economic or social assets. It is generally measured by parental income, parental education, and parental occupation at the student, school or neighbourhood level (Sirin, 2005), but the OECD, and other international large-scale assessment (LSA) providers, also include proxies to capture the social capital dimension (Bourdieu, 1986) such as the number of books available in the household. The SES effect sizes are constantly large across countries and LSAs, and over time (Hanushek & Woessmann, 2011; Scheerens et al., 2015). There is, however, an interplay between SES and the

structure of education systems, and many studies (Hopfenbeck et al., 2017; Marks, 2006) find that SES is more strongly correlated with PISA scores in countries with early streaming of students into fixed educational pathways.

There are two main reasons why SES affects educational performance. Firstly, a high SES can open opportunities early in the child's cognitive development. The fundamental importance of early childhood and pre-school education was established by interdisciplinary studies (Ramey & Ramey, 1998; Ramey, Ramey, & Lanzi, 2006) and confirmed by economists who reframed it in terms of skill formation: early education has a higher rate of return than later education because 'skill begets skill and motivation begets motivation' (Cunha & Heckman, 2010; Heckman, 2008; Kautz, Heckman, Diris, ter Weel, & Borghans, 2014). While early childhood education is beneficial to children from all backgrounds (Link, 2012; Salverda, 2011), those from high-SES families are more likely to participate in it (OECD, 2011a). Especially in inequitable countries, low-SES families have also less access to credit, which limits their capacity for investing in education from the earliest stages (Lochner & Monge-Naranjo, 2012). Children in large low-SES families face competition (from their siblings) for few resources (Wolter & Vellacott, 2003), which becomes an issue where there are no compensatory social policies or free pre-school education (Park, 2008). Marks (2014) used data from the Australian Year 3 to show that prior achievement is a stronger predictor variable for later achievement that SES, and that the correlation between SES and achievement decreases even more in later Years. There is no contradiction in this. In Year 3, children are already aged 8 or 9. By then, family SES has had a chance to alter the children's future trajectories and to contribute to entry achievement (Duncan et al., 2007; Heckman & Mosso, 2014; Sylva, 2014).

Secondly, SES is related to cultural capital, which also correlates with PISA outcomes (Caro, Sandoval-Hernández, & Lüdtke, 2014; Pokropek, Borgonovi, & Jakubowski, 2015). Cultural capital can be expressed as a range of cultural activities, including cultural conversations in the family, as well as in the possession of cultural and educational resources (Xu & Hampden-Thompson, 2012). The relationship between cultural capital and achievement changes by welfare regime and it is usually stronger in more unequal societies (Tan, 2015; Xu & Hampden-Thompson, 2012), where parental financial and cultural support has greater implications. Within countries, however, there is mixed evidence on whether cultural capital reduces (Andersen & Jæger, 2015) or exacerbates structural inequalities (Marteleto & Andrade, 2014).

While earlier approaches to the measurement of SES and cultural capital were hampered by the choice of variables, recent research and the OECD tend to treat economic, social and cultural capital as complementary dimensions for a child's access and exposure to education (Nonoyama-Tarumi, 2008). Because of this, the present study considers changes in the OECD Economic, Social and Cultural Status (ESCS) as a proxy for changes in the socio-economic and cultural capital composition of PISA cohorts.

Immigration status and first language

> Across OECD countries, first-generation students – those who were born outside the country of assessment and who also have foreign-born parents – score, on average, 52 score points below students without an immigrant background. (OECD, 2010, p. 14)

Immigrant students are generally behind in terms of access to education, participation and academic achievement. Across countries and in a range of international LSAs, the effect sizes of immigration status are about 0.38 in reading and mathematics, and 0.43 in science, and they are

overwhelmingly in favour of non-immigrant students (Andon, Thompson, & Becker, 2014). Since the proportion of immigrant students in secondary schools is weakly but negatively related to the scores of non-immigrant students (Brunello & Rocco, 2013), it may be posited that an increase in the number of immigrant students in the PISA sample might have a small impact on overall country outcomes. Despite this, immigration does appear to have had a positive impact on the educational success of schooling in London (Burgess, 2014).

Much of the educational shortfall exhibited by immigrants can be explained by socio-economic disadvantage (Marks, 2005), and in fact countries with more pronounced inequalities among non-immigrants also show the wider gaps among immigrants (Schnepf, 2008). Sometimes, socio-economic effects can be striking; for example, Cattaneo & Wolter (2015) used changes in the Swiss immigration policy to show that 75% of the variance in the 40-point increase in the PISA score of immigrant students between 2000 and 2009 was due to changes in family background and improved school composition.

Immigrant students are at risk of reduced confidence because of fear of being judged inadequate due to their identity and teacher expectations, and they are often grouped with other immigrant students (Schofield, Alexander, Bangs, & Schauenburg, 2006). Because of this, policies like early integration, language interventions and desegregation tends to be successful in many cases; however, there is also a strong interplay between immigrant student characteristics and some features of the education system on arrival, which means that comparable policies may not work in every country or with all groups (Cobb-Clark, Sinning, & Stillman, 2012; Dronkers, Levels, & de Heus, 2014; Nusche, 2009). For example, using data from PISA 2006, Shapira (2012) found that while family SES and misalignment between migrant family and host-country-specific cultural capital explain the largest share of these differences, school-level and country-

level features can reinforce or attenuate the achievement gap. For instance, first-generation immigrants seem to be better off in countries with a liberal welfare regime, more standardised educational systems and more selective immigration policies. Indeed, it seems that the structure of the education system, rather than specific targeted interventions, is a key determinant in the achievement gap of immigrant across countries (Nolan, 2010).

A further factor affecting the performance of immigrant students might be the linguistic barrier, though findings in this case are more contested. Immigrant students tend to perform better in assessments requiring less reading and writing (Cheng, Wang, Hao, & Shi, 2014; Schnepf, 2008). Those who speak the language of assessment at home do better in PISA that those who do not (Christensen & Stanat, 2007), but in many countries the correlation between language spoken at home and PISA outcomes disappears when SES is controlled (Cummins, 2012). The most likely explanation is that, while the language skills of immigrant students are key to their educational success, the language they speak at home is an imperfect proxy for these skills. Modelling this variable does reduce the "unexplained" variance (Schneeweis, 2011), but speaking a language different from that of the assessment in the house does not always mean that a student is unable to speak the host country language. Therefore, this study controls for both immigration status and first language, along with SES.

Grade

PISA students are sampled by age, not by grade. Student in higher grades have higher outcomes, even after accounting for student age (Luyten, Peschar, & Coe, 2008) or external factors such as grade retention or advancement (Luyten & Veldkamp, 2011). The difference in the average OECD score by grade in PISA reading 2012 is shown in Table 1.

[TABLE 1 HERE]

The distribution of students in grades varies by country. For instance, while most 15-year-old students are in grades 9–11, their percentage in grade 7 ranges from 0 in many countries to 15.4% in Tunisia in 2003 and 16.3% in Brazil in 2000. Similarly, while in most cases 15-year-olds are not in grades 12 or 13, in New Zealand in 2012 5.4% were. In the same year, 15-year-olds in Argentina could be found in all grades between 7 and 13. Some have argued that this puts some countries at relative disadvantage (Doyle, 2008); but it should be recalled that the stated purpose of PISA is to compare the experiences of pupils of a certain age regardless of the grade or type of schools in which they are enrolled.

It is more interesting to note that the distribution of students in grades varies within countries over time. Consider, for example, the case of Germany (Table 2). While the modal grade has remained the same (Grade 9, shaded column), the "centre of mass" has moved progressively towards the higher grades, as evidenced by the increasing proportion of students in Grade 10 and 11. Even though the PISA scores for Germany improved between 2000 and 2015, the improvement would have been more modest if the distribution of students in grades had remained the same as in 2000, conditional on the scores in each grade also being stable, as they were in Grades 9 and 10. The performance of students in Grades 7 and 8 did increase (the former, from 349 to 365 points; the latter, from 395 to 422 points); however, the proportion of PISA participants in these grades dropped so substantively that the increase in score was compensated by the decrease in weight.

[TABLE 2 HERE]

Germany is not unique in this trend. In 2000, Brazil reportedly did not have any 15-year-old student enrolled in Grade 11, but in 2012 they represented 42.0% of the cohort. In general, if student results vary so much between grades, and if the proportions of students in grades also

changes over time, then it is reasonable to expect that fluctuations in the grade distribution might affect country outcomes.

To summarise, there is strong evidence showing that SES, immigration status and grade are related to PISA outcomes. Building on this literature, the first question explored in the article is whether changes in these variables can become so sizeable as to make substantial differences to trend analyses. The ability of PISA to measure country trends rests not just on the equivalence of different versions of the assessment and on the reliability of sampling procedures[1], but also on the comparability of student populations, which change because of socio-economic trends and migration. This is acknowledged by the OECD:

> Changes in a country's […] performance can have many sources. While improvements may result from improved education services, they can also result from demographic changes that have shifted the country's population profile. […] PISA ensures that all countries and economies are measuring the mathematics performance of their 15-year-olds enrolled in school; but because of migration or other demographic and social trends, the characteristics of this reference population may change. (OECD, 2014c, pp. 58–59)

Because of this, the OECD has recently started to provide "adjusted" trends that try to account

[1] There has been extensive critique regarding the ability of PISA to measure change in individual country performance reliably. Many items have been shown to exhibit differential item functioning (DIF; Grisay et al., 2007; Grisay, Gonzalez, & Monseur, 2009), which affected country-level estimations; a quarter of observed changes between 2000 and 2003 may have been due to DIF (Gebhardt & Adams, 2007), and even large drops, such as the 24 points lost by Japan between 2000 and 2003, were probably a statistical artefact (Monseur & Berezner, 2007). Other sources of error include that due to the linking of different versions of the assessment, which at some point may have accounted for score differences of up to 40 points (Wu, 2010). Large improvements (or decreases) from one cycle to the next are possible but unlikely, and in general, 95% of country score changes are well within the ±20 points range (Lenkeit & Caro, 2014).

for education-policy-independent changes in the PISA population. Specifically, the OECD centres five variables to the latest assessment: student age, gender, socio-economic status, immigration status and whether their first language is different from the language of the assessment. The centred variables are then used to calculate the adjusted country means for the previous cycles (OECD, 2014c). These are the same variables used in this study, with the exception that age was replaced with grade.

The picture that emerges from OECD analyses shows that the difference between the usual reported scores and the adjusted ones is often not negligible.

[TABLE 3 HERE]

Table 3 shows the difference between the annualised change (i.e., the average change across a country's PISA cycles) as reported in international reports and the same statistic adjusted by age, gender, SES, migration status and first language, for reading 2000–2012 in a selection of countries. While PISA publications return a global picture of year-on-year rates of change that is mostly positive, the picture changes substantively once variations in population characteristics are considered: almost half of all country trends change sign if changes in the population are accounted for.

If their demographics had not changed, countries like Chile, Colombia or Qatar would have been reported to have positive reading trends instead of negative, whereas the opposite would have happened for countries such as Argentina, France or Spain. The improvement of Poland between 2000 and 2012 would have halved and that of Germany would have become statistically non-significant. The difference between the adjusted and not adjusted rates of change

ranges from the −11 or −10 points of Romania, Russian Federation, Taipei, Bulgaria and Argentina to the +10 to +12 points of Qatar, Serbia and Colombia[2].

These results raise the question: At what stage can one claim that two PISA populations are no longer comparable? Clearly, it would not be fair to hold education stakeholders to account if student outcomes were influenced by factors outside their control.

The relationship between curriculum and performance in international LSAs

The considerations in the previous section can be contrasted with the view promoted by the OECD whereby improvement in PISA means that a country was more effective in teaching its students. This article explores this through the lens of changes in the school curriculum, a highly-malleable policy element. Before reviewing the relevant literature, it is important to address one obvious objection, that PISA tests "skills for life" and not country curricula (e.g., OECD, 2001).

Goldstein (2004; Goldstein & Thomas, 2008) have questioned whether it is possible for a test to provide a meaningful measure of educational achievement that is not influenced by curricula. In PISA, items that function very differently in some countries are eliminated. This may reduce the sensitivity of the assessment to specific instruction (Wiliam, 2008), but it does not change the fact that PISA measures *something* and that country curricula might be more or less aligned to it. McGaw (2008) pointed out that countries do not score better on items they rate

[2] Interestingly, the differences in the unadjusted and adjusted changes for mathematics 2003–2015 are much more modest, ranging from +2 points in the adjusted trends for Germany to the -5.2 points for Qatar, and the standard errors are almost identical (data not shown). This may be in part accounted for by the fact that reading trajectories are more unstable, and it might also be the case that the OECD has improved the consistency of its estimates.

as being more in line with their curriculum, but the issue might lie with the accuracy of the rating (Goldstein & Thomas, 2008).

Research shows that the construct and contents of PISA science have changed over time (Kind, 2013; Lau, 2009). This has had a washback effect on countries, which are increasingly aligning their curricular contents and assessment formats to PISA (Breakspear, 2012; Hopkins, Pennock, & Ritzen, 2008; Schleicher, 2009). The OECD is not a neutral spectator in this process, as it advises countries to introduce curricula that are 'better aligned with […] 21st century skills' (OECD, 2014c, p. 253) and praises them when they do so (see for example the case of a Canadian region and its new 21st century curriculum OECD, 2014b, p. 119).

All this suggests that curricular changes may play a role in PISA trends: countries amending their instruction, so that it is more in line with the constructs and contents agreed during the development of PISA, might be able to educate cohorts who are more attuned to the test requirements. It is the kind of teaching to the test that the OECD would support (see for example Kanes, Morgan, & Tsatsaroni, 2014, on the PISA mathematics "regime"). The question then becomes what evidence exists that school curricula can make a difference.

There is a scarcity of research from an international comparative perspective on this topic. A lively research strand tackles curriculum differentiation, but this is broadly defined as the impact of ability grouping and streaming on student achievement (Schofield, 2010), not specific curriculum effects. Many recent studies show that more stable factors such as the structure of education systems capture most of the variance in international assessments (Scheerens et al., 2015; Woessmann, 2016), and there is scepticism around the effectiveness of education policies (Coe, 2009). Nevertheless, there have been some attempts that are relevant for the current discussion. Motiejunaite, Noorani, & Monseur (2014) analysed national policies to

support low achievers in reading in 32 countries, including the availability of curriculum guidelines on reading comprehension strategies. They found that the presence of absence of such guidelines in national frameworks was not significantly associated to PISA 2009 outcomes. Scherer & Beckmann (2014), instead, focused on mathematics, science and problem solving. They found that a country's educational objectives (whether prominence is given to academic subjects or general applied skills) were unrelated to PISA 2003 outcomes. However, a statistically significant correlation could be detected between PISA science and whether the national curriculum made explicit connections between different areas of science. The International Instructional Systems Study examined the education systems and curricula of 11 high-achieving jurisdictions (Creese, Gonzalez, & Isaacs, 2016). In these jurisdictions, the goals of the education system are clear and explicit, and they emphasise literacy and numeracy, problem-solving, critical and creative thinking and citizenship. Their curricula are interdisciplinary and integrate so-called 21st century skills, but the authors of the study cautioned that not all policy changes had made it into the classroom, often because of lack of change in examinations.

In mathematics, the 21st century narrative has resulted in a general alignment towards the PISA objective of applying mathematics to solve problems in "real-world" scenarios (Merriman, Shiel, Cosgrove, & Perkins, 2014; Smith & Morgan, 2016). Despite a cross-curricular emphasis on mathematics as an applied tool, there are variations even within the same jurisdiction in the extent to which these goals have been adopted (Smith & Morgan, 2016). In science, however, there are more similarities among curricular contents than there are among the overall goals of high-achieving jurisdictions (Hollins & Reiss, 2016).

The present study and research questions

The literature reviewed in the previous section indicates that there is scope for further work. On the one hand, countries around the world are rising to the PISA challenge by amending their curricular offer to include 21st century skills or, at least, to be more in line with the OECD's reading, mathematics and science literacies. Curricular revisions are sometimes accompanied by changes in the examinations. Could this make a difference to their performance, or is PISA essentially "unteachable"?

On the other hand, there is no doubt that students with differing socio-economic and demographic characteristics reach different levels of performance. There is also evidence that these characteristics change over time, and that PISA cohorts may differ along one or more dimensions. Although the OECD does publish PISA data adjusted by socio-economic and demographic changes, these are not the scores that make the news and they do not take all available factors, such as Grade level, into account. Can "fairer" performance trends, adjusted for factors outside the control of education policymakers, be created to distil out the impact of policy?

This article seeks to answer the following research questions:

(1) What is the relationship between changes in the socio-economic and demographic characteristics of the PISA cohorts, and changes in country outcomes?
(2) What is the relationship between changes in the curricular provision of PISA-participating countries and their outcomes?
(3) Overall, what is the relative importance of non-policy-malleable factors (student SES and demographics), when compared with policy-malleable factors (curricular changes) with respect to PISA scores?

Methodology

Country Sample

More than 70 countries participated in PISA between 2000 and 2015, and since the focus of this article is to investigate trends over time, only countries with at least two data points were retained, bringing the total available number to 63 (Appendix

Table 14). Because each PISA cycle had at least one missing case, some analyses placed additional restrictions to the number of available countries. The choice of reading, mathematics or science as an outcome variable also influenced the number of countries available for analysis. For instance, the reading scales are broadly comparable—with some caveats (OECD, 2016b, Annex A5)—between 2000 and 2015; the mathematics scales between 2003 and 2015; and the science scales between 2006 and 2015. Because there were so few time points for science, it was not included in the analysis.

Regarding reforms, information was first sought in the PIRLS and TIMSS encyclopaedias and country questionnaires available between 1999 and 2015 on the Boston College website (http://timssandpirls.bc.edu/). Follow-up databases included the sixth and seventh editions of the *World Data on Education* (UNESCO-IBE, 2007, 2012), the Eurydice Network for European countries (European Commission/EACEA/Eurydice, 2016), as well as specific documentation on mathematics curricula in Europe (Eurydice, 2011). This allowed us to retrieve information about reading curricula reform dates in 39 PISA countries, and mathematics dates in 59 countries.

To increase and cross-validate the data collected from documents, a questionnaire was sent to 166 experts in the sampled countries (Table 15). Expert answers provided an extremely valuable insight into specific country changes; however, the low response rate (about 17%)

meant that little information could be integrated within this study over what had already been collected from databases. Eventually, we decided to focus the quantitative analysis on curricular changes in mathematics (N = 59). Since only about 60% of the original sample had data on changes in reading curricula, these were only explored qualitatively.

Outcome variable

The outcome variables are the country mean estimates in PISA reading and mathematics in OECD international reports (e.g., OECD, 2016b). These estimates are measured on a scale with a mean of 500 and a standard deviation (SD) of 100, and they are calculated from the students' plausible values (OECD, 2014a)[3].

Input variables

Time

The calendar year was used to produce a coded variable for each assessment administration ($Time$). The $Time$ variable orders the time-points in which each assessment was held and, to measure growth, it is zeroed at the first longitudinally-comparable assessment. This is 2000 for PISA reading and 2003 for PISA mathematics. Therefore, $Time_{Read} = (0, 1, \dots, 5)$ for PISA 2000, 2003, up to 2015; whereas $Time_{Maths} = (0, 1, \dots, 4)$ for PISA 2003 to 2015.

[3] Plausible values are not scores, but a function representing 'a likely distribution of a student's proficiency' (von Davier, Gonzalez, & Mislevy, 2009, p. 11). Student-level analyses in this paper used a subset of five such values that can be found in the PISA database.

OECD membership

$OECD$ is a dummy variable that takes the value of 1 for OECD countries and 0 for partner countries. Three countries (Chile, Estonia and Israel) joined the OECD in 2010, but since the process to acquire membership began some years earlier and they were all high-income economies by 2012 (The World Bank, 2014), they were considered OECD members for the purpose of this article.

Socio-economic status

The OECD measures SES through a standardised index, the Economic, Social and Cultural Status (ESCS). The ESCS is obtained through principal component analysis of standardised indices of home possessions, parental occupation and parental education. Although the procedure is slightly different for OECD and partner countries and it was modified throughout the years, the factor loadings are comparable across countries (see OECD, 2014a, pp. 351–354, 372–375). The OECD provides student-level ESCS data for the first five cycles that were rescaled to be comparable to the 2015 values (OECD, 2016a).

For this article, the country-level ESCS values (i.e. the country average SES levels at each PISA administration, variable SES) were calculated in IBM® SPSS® Statistics 22 (SPSS 22) using the procedures detailed in the *PISA 2009 Technical report* (OECD, 2012). The country-level value for Albania in 2012 (whose student-level data were completely missing) was predicted using the country coefficients estimated by a multilevel growth model (see section below) of SES on $Time$, controlling for $OECD$.

Demographic characteristics

The variables FEM, IMM and FL are respectively the percentage of females, first-generation

immigrants and students whose first language is different from the language of the test expressed as a ratio to the student population (e.g., $0.5 = 50\%$). Values for all countries were available in the PISA databases and are included as an additional file to this article.

Grade

As noted earlier, PISA samples students by age, and the target population comprises students aged between 15 years 2–4 months, and 16 years 1–3 months (OECD, 2014a). However, the grade requirements are less stringent and it is sufficient that a student be enrolled in grade 7 or above. In fact, the OECD invites countries to include students from a range of grades to expand the sampling frame, provided they fall within the correct age category (OECD, 2014a, p. 83). This translates into cross-country variations of participating grades, and it also means that the distribution of PISA participants in grades in any one country could change over time for reasons that may or may not be related to policy changes.

One option for measuring the impact of the shifts in the student distribution in grades is to treat grade as a composite variable, where the sum of the percentages of students from Grade 7 to 13 is bound to 100 (for methods, refer to van den Boogaart & Tolosana-Delgado, 2013). However, as noted earlier, an analysis of the PISA compendia revealed that the mean PISA scores usually increased with grade. Only a minority of countries (6 out of 65 in 2012) displayed scores in the highest assessed grade that were equal to or lower than those in the lower grades. This makes intuitive sense; in general, students of a similar age but in a different grade from the other PISA participants were probably retained (or advanced) or had learning difficulties (or were particularly gifted). Therefore, it was decided to treat $GRADE$ as a ratio variable like FEM, IMM and FL. Its value is the percentage of students *at or above the modal grade* for the country. Since the modal grade varied with time in some countries (it usually shifted by 1, but changed

from grade 9 to grade 11 in Brazil 2009–2012, suggesting extreme changes in the student distribution), a decision had to be made on which grade to keep fixed to enable trend comparisons. The modal grade was chosen and that is listed by country in Table 16. Students with missing grade information were included among those below the modal grade, since it was noticed that they had, on average, lower scores than those at the modal grade.

Curricular changes

There is no reliable or established way to quantify curricular changes. Changes in the curriculum vary in size, scope and grade affected; they are often rolled out progressively; and, of course, there are variations in the extent to which they are taken up by schools and teachers and integrated into classroom practice. This latter point is what makes cross-country quantification difficult, because evaluations of successful curricular implementations are not carried out in all countries, and when they are, they often have to rely either on self-reported school questionnaires or on a small number of inspection visits.

The simplest modelling approach is to use a dummy variable taking the value of 1 whenever a reform is adopted *and* it could influence PISA outcomes. Identifying which students are affected by a reform is not straightforward, but one could try to use the modal grade as reference. For example, a reform of the mathematics curriculum introduced in 2008 in Grade 9, in a country where the PISA modal grade is 10, could be expected to affect the PISA 2009 cohort but not the 2006 cohort, so the country values for the curriculum variable C would be $c_{06} = 0$ and $c_{09} = 1$. However, if the reform had been introduced in 2008 in Grade 10, the majority of students would have moved to Grade 11 by 2009 and not participated in the assessment, and the variable would take values $c_{06} = 0$ and $c_{09} = 0$.

This approach is appealing but has some critical shortcomings. Firstly, it does not account for the fact that a reform might not only affect students who are in a certain grade when the switch takes place: younger students eventually enter a grade where the new curriculum is taught. Moreover, if there are no further reforms, each successive cohort uses the new curriculum. A variable constructed in such way is not very informative, because it can only capture the first curricular change, just one big fluctuation in PISA scores per country. If $\boldsymbol{c}_{0315} = (0, 0, 1, 1, 1)$, where the first zero is PISA mathematics 2003 and the ones represent the effect of the hypothetical 2008 reform mentioned above, how can one know whether a second curricular change happened just before PISA 2012 or 2015? Change is modelled as one sudden performance jump and then a flat trend, which is not realistic. A solution was proposed by Braga, Checchi, & Meschi (2013):

> Take the case of "Pre-primary expansion" in Finland: we find records of significant reforms over this dimension in 1973, in 1985 and in 1999. We therefore construct a variable, which is zero before 1973, 1/3 between 1973 and 1984, 2/3 between 1985 and 1998 and 1 afterward. We then match this variable to individuals. (Braga et al., 2013, p. 61, footnote 8).

This approach solves the issue of measuring the effect of successive reforms but is still imperfect. Its main limitation is modelling policy change as an incremental growth, whereas often one curriculum is introduced to replace another's aims and values but keeps most taught content unchanged.

To account for these issues, the following approach was adopted:

(1) In each country, every reform of the mathematics curriculum for lower-secondary and secondary education (not reading because of the sample size limitations) that could affect any PISA cohort was identified, and its adoption year was noted.

(2) It was assumed that the greatest effects would be detectable soon after the policy had been implemented, with some time allowed for classroom penetration.

(3) This window of effect was assumed to span over the five years preceding the test (ages 11–15), with a tolerance of 1 year at the beginning (i.e., policies adopted when the PISA cohort was 10 still qualified for the 11–15 period). Any policy introduced during this period made the variable take a nonzero value. Because of these constraints, only country policies between 1995 and 2014 were considered.

(4) Unlike the method chosen by Braga et al. (2013), the value of the variable did not depend on the position of the reform in the policy sequence, but on the number of years a PISA cohort had been taught under the new curriculum. For instance, a reform of the mathematics curriculum adopted in 2002 would have had a chance to affect the PISA 2003 cohort for two years and the PISA 2006 cohort for five years (in 2002, these students were 11 years old).

(5) Under this approach, each year counts as one-fifth of the total reform "effect". This weighting can be thought as varying degrees of curricular penetration. In the example above (new mathematics curriculum in 2002), the PISA 2003 cohort had only two years of instruction under the reformed curriculum and, since curriculum integration almost certainly happened at different paces, it is possible that only a minority of students (40%) were able to benefit from it, against the totality of students in PISA 2006.

(6) Bringing point (4) and (5) together, the values for the curriculum variable for this example in these two PISA cohorts are $c_{03} = 2 \times \frac{1}{5} = \frac{2}{5}$ (partial implementation) and $c_{06} = 5 \times \frac{1}{5} = 1$ (full implementation).

(7) To avoid incremental effects, it was further decided to zero any reform effects after five years. In other words, if a cohort was younger than 10 when a new mathematics curriculum was introduced, the corresponding value of the curriculum variable was set to zero. This solution was adopted to use all available information on the interaction cohort-policy, while minimising the effect of external influences that would build up over time and confound policy effects. Of course, this does not mean that later PISA cohorts were unaffected by reforms, only that the specific impact of policy change (over and above socio-demographic changes) would have been absorbed.

(8) Some countries, including Australia, Canada, Germany and Belgium (both Communities), do not have a centralised curriculum; therefore, reforms could affect different regions in the country at different times. The approach adopted in these cases was *ad hoc*. Where there was evidence that all or most of the country took a certain curricular turn at the same time, the method in (5) was followed. Otherwise, the basic weight was applied but not compounded, as if the probability of each cohort being taught under a modified curriculum had been the same. For instance, various curricular reforms in Australia took place between 2000 and 2009. To reflect this uncertainty, all PISA cohorts between 2003 and 2012 were associated with a value of one-fifth.

Table 17, in the Appendix, shows the reform years and the values of the curriculum variable.

Analytical Methods

Figure 1 and Figure 2 map the trends in PISA scores in the two domains for OECD and partner countries. To model these trends, multilevel/hierarchical modelling (Goldstein, 2011; Raudenbush & Bryk, 2002) was employed. This analysis takes the grouping of observations into

account by assigning a variance component to the different clusters, or levels, within which the observations are nested, in addition to the observation residual.

To a first approximation, one may think that the trends in Figure 1 and Figure 2 are statistically equivalent to horizontal, parallel lines, where only the intercept with the *y*-axis varies by country. The equation for this null model, which is a random-intercept model, would then be:

$$Scores_{ij} = \beta_{0j} + e_{ij} \quad (1)$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad (2)$$

The subscript i indexes country-year observations (the Level-1 unit) and the subscript j indexes countries—the cluster in which the observation took place (the Level-2 unit). The intercept β_{0j} takes the j-subscript, because it can vary by group (country), and it can be decomposed into a fixed part β_0, which is common to all groups, plus a "random" part u_{0j} that varies by group.

Figure 1. Trends in PISA reading 2000–2015

[FIGURE 1 HERE]

Figure 2. Trends in PISA mathematics 2003–2015

[FIGURE 2 HERE]

A better fit for the data shown in Figure 1 and Figure 2 allows countries not only to vary in their performance levels on the first measurement occasion, but also in their rate of change. This calls for a random intercept and random slope model, defined as follows:

$$Scores_{ij} = \beta_{0j} + \beta_{1j}Time_{ij} + e_{ij} \quad (3)$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad (4)$$

$$\beta_{1j} = \beta_1 + u_{1j} \quad (5)$$

Since the passing of time is explicitly modelled by the $Time$ variable, this kind of model is also called a growth (curve) model (Steele, 2008). Equations (3) to (5) can be expressed in long form while separating fixed from random (between brackets) components:

$$Scores_{ij} = \beta_0 + \beta_1 Time_{ij} + (u_{0j} + u_{1j} Time_{ij} + e_{ij}) \quad (6)$$

Equation (6) has six parameters to estimate: β_0 and β_1, the variances of u_0, $u_1$and e, as well as the covariance between u_0 and u_1, where it is assumed that:

$$e_{ij} \sim N(0, \sigma_e^2) \quad (7)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \quad (8)$$

This modelling returns the following information[4]:

[TABLE 4 HERE]

[TABLE 5 HERE]

[4] All models were fitted in the statistical environment `R` (R Core Team, 2016, version 3.3.0) using the `lme4` package, (Bates et al., 2015, v.1.1-12). The confidence intervals were based on the likelihood profiles of the model (see Bates et al., 2015, pp. 26–28), rather than the more common (and less accurate) standard error. The degrees of freedom were approximated using the Kenwald-Roger method (Kenward & Roger, 1997), which tends to be more conservative than traditional methods and result in fewer Type I error rates (Li & Redden, 2015; Spilke, Piepho, & Hu, 2005), with the package `pbkrtest` (Halekoh & Højsgaard, 2014, v. 0.4.6).

The tables are formed from three sections. The top section lists the estimated parameters (in this case the intercept, the slope, their variances and covariance and the residual variation), along with their standard errors (SE), 2.5% and 97.5% confidence intervals (CI low and CI high), *t* statistic, degrees of freedom (DF), *p* value and a visual indication of the significance level. The middle section provides three information criteria to evaluate model fit: the Akaike information criterion ([AIC], Akaike, 1973), the Bayesian information criterion ([BIC], Schwarz, 1978), and twice the negative logarithm of the likelihood function (sometimes called "deviance", see Bates, Mächler, Bolker, & Walker, 2015). Finally, the bottom section summarises the number of Level-1 and Level-2 data points used for estimating model parameters.

As a preliminary comment, it is interesting to note that while the global performance trend has been small and positive in reading, in mathematics this trend was not significantly different from zero at the 5% significance level (though individual country trends may well have been). Moreover, the trends for reading and mathematics appear to follow different directions between OECD and partner countries, and the variance of the latter's scores is certainly higher than that of OECD members as can be seen in Figures 1 and 2. Finally, the negative covariances between intercept and $Time$ suggests that countries that had a good initial PISA performance tend to increase less (or not increase at all) compared to countries that had lower scores at the beginning.

The growth models set out above are the reference models for this article: the additional variables of interest are entered individually or simultaneously and their impact on predictions is evaluated by comparing the deviance of the larger models with that of the smaller reference models.

One key question in this study was whether change in socio-economic, or demographic factors, could be correlated to change in scores. Following Fairbrother (2013) and others (Enders & Tofighi, 2007; Raudenbush, 1989; Raudenbush & Bryk, 2002), a solution was adopted by decomposing the indicators into two components: a time-invariant, group-centred mean at Level-2 ($\bar{x}_j$, or $X.mean$ in the model results), and its Level-1, time-varying difference (x^*_{ij}, or $X.change$), as follows:

$$x_{ij} = \bar{x}_j + x^*_{ij} \quad (9)$$

$$x^*_{ij} = x_{ij} - \bar{x}_j \quad (10)$$

Since $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = 0$ for all j, $\bar{x}_j$ and x^*_{ij} are orthogonal and the effect of x^*_{ij} is independent from the magnitude of $\bar{x}_j$. In other words, this decomposition allows us to explore changes in socio-economic and demographic conditions regardless of their average level in each country. A list of all models used for this research, including the null and reference growth models above, can be found in Table 18.

Analysis

The analysis starts with the multilevel modelling of PISA reading and mathematics scores with socio-economic and demographic variables taken individually: SES, immigration status, first language and grade. After a summary of the key findings from the individually-modelled variables, the analysis proceeds with multilevel multivariate models attempting to capture the combined effects of changes in the PISA cohort composition. In the second part of the analysis, mathematics scores are regressed on curricular changes. Because of the relative scarcity of quantitative data in this area, the evidence is complemented by a case-study based argument on

the plausibility of the claim that changes in school curricula might be behind PISA trends.

Non-policy-malleable variables

Socio-economic status

Figure 3. Deviation of SES from country average 2000–2015

[FIGURE 3 HERE]

Figure 3 shows the country-level changes in the socio-economic status of the PISA population, expressed in terms of deviations from the country mean as per Equation (10). Despite some fluctuations, there is a clear positive and linear trend, so that each successive cohort is becoming increasingly socio-economically advantaged, in OECD terminology (OECD, 2016b). The implications of this are discussed below.

[TABLE 6 HERE]

[TABLE 7 HERE]

A significant link between the average socio-economic status of students in a country and country scores was found (Table 6 and Table 7); this relationship persisted even after accounting for OECD membership. This was unexpected, because OECD countries tend to be richer and it was thought prior to the analysis that socio-economic advantage would be confounded with OECD membership. Even more interestingly, there was a statistically significant "effect" of changes in SES on changes in scores, both in reading and mathematics. Regardless of their initial average SES and their status as OECD members, countries whose students' average socio-economic status increased by one standard deviation could also see their scores improve by almost 22 points. This suggests that changes in the PISA sample or in the student population may indeed have had an impact on country scores.

It should be noted that the coefficient of $SES.change$ remains significant even when $Time$ is added to the equation, even though sometimes including the $Time$ variable (Table 7) provides a better fit. Formal tests such as the condition index (Belsley, Kuh, & Welsh, 1980) and the variance inflation factors ([VIF]: Zuur, Ieno, & Elphick, 2010; Zuur, Ieno, Walker, Saveliev, & Smith, 2009) did not provide any evidence of collinearity between $Time$ and $SES.change$.

Immigration status and first language

Unlike socio-economic status, the average proportion of pupils not born in the country was not related to differences in country scores. Deviations in the proportion of immigrant students over time (variable $IMM.change$) were also not significantly correlated to changes in scores on a global level, but this was because the error was large. Moreover, including $IMM.change$ as a random effect did improve the model fit in both reading and mathematics. This suggests that the relationship between immigration and performance trends could be substantive, if only in a minority of countries. However, neither the mean nor changes in the percentage of students whose first language is not the language of the test was significantly related to average level or change in the countries' PISA scores. First language was not statistically significant even when interacting with immigration status.

Grade

In general, the percentage of students at or above the country modal grade (or changes in this percentage) was not significantly different from zero. Statistical significance aside, these coefficients were small compared to SES or immigration status, and information criteria showed no improvements compared to the null model.

There was one instance when this was not the case: a random-intercept random-slope model of mathematics scores on $GRADE$. In this case, the cross-country change coefficient was almost 33 points and significant, and the fit was also better. This is in line with previous research and with the differences in PISA scores between the various grades (Table 1). However, because only one regression provided evidence of statistical effects related to departures in the percentage of students at or above the countries' modal grade, these results might have been caused by sampling variability.

Full models

The analyses above pointed towards the following findings:

- Country differences in the average level of SES correlated with differences in PISA outcomes. These relationships still held when countries were grouped by OECD membership.
- At the cross-country level, there was a strong correlation between changes in student SES and changes in PISA outcomes that went beyond general overall growth (i.e., the $Time$ effect).
- There might be a strong relationship between fluctuations in the percentage of immigrant students and PISA scores, as suggested by the fact that adding this variable improved the model fit. No statistical evidence could be found of correlations between the percentage of students whose first language is different from that of the test and country outcomes.
- Very little evidence could be found of correlations between the percentage of students at or above the country modal grade and country outcomes.

These findings where explored further by fitting increasingly complex models that included

gender, level and variation in socio-economic status, and a different slope for the immigration "effects" (Table 8 and Table 9).

[TABLE 8 HERE]

[TABLE 9 HERE]

Because of the large standard errors, some relationships can be identified with greater confidence than others, but there are general patterns.

(1) A change of 1 SD in SES between PISA cycles is associated with a change of 17.6 points (reading) and 31.7 points (mathematics), regardless of the average level of SES in a country ($SES.mean$) and OECD membership. Moreover, the magnitude of this change is greater than that of the passing of time and it is in addition to it. This model stretched the analyst's computational capacities, so it was not possible to calculate the confidence intervals of the covariance between SES and intercept. This means that it cannot be said for definite whether countries that had the highest PISA scores in 2000, or soon after, experienced the lowest increase in SES, as the estimate seems to suggest.

(2) One can tentatively claim that trends in the percentage of immigrant students in the 15-year-old population might be related to PISA trends, at least for reading. Again, this correlation seems to be independent from the average level of immigrant students in a country, which was found to have a non-significant coefficient in previous stages of the analysis.

(3) There is a negative relationship between the percentage of female students in a cohort and PISA scores. This seems to be due to specific sample characteristics rather than to some underlying causal mechanisms. Looking at the country sample, the average value for the gender value is 50.06%, with a standard deviation of only 1.58 percentage points. This

means that in most countries, the ratio of boys and girls is extremely balanced. There are, however, a few outliers with values that are up to ±5 percentage points away from the average, which have therefore a sizeable weight on the correlation. In particular, many low performing non-OECD countries such as Argentina, Costa Rica, Brazil, Tunisia or Thailand, have 52% or 53% female students in their population; conversely, OECD countries tend to have fewer girls in their cohorts, with high-scoring Korea having less than 46%. This is a reminder of the risk with extrapolating too much from the sample.

Policy-malleable variables

Since data on reading curricula was available for only 39 out of 63 countries (62%), it was decided to limit the multilevel analysis to mathematics, and to complement it with a more qualitative case studies review. The output of a model of PISA mathematics on curriculum changes is shown in Table 10.

[TABLE 10 HERE]

At the global level, there does not seem to be a "curriculum effect"; reforms are not significantly associated with changes in PISA mathematics. However, the model ran into convergence issues, because there is insufficient variability within countries in the relationship between curricular changes and PISA. In other words, curricular changes *as modelled in this study* were not related to country scores.

Another way to consider the issue is by comparing the significance of the $Time$ coefficient under different specifications. In practice, this can be done by looking at the *t* statistic of the reference growth model (Model 3), the best model fit accounting for demographic characteristics (Model 48b), and the curriculum model (Table 11). Adding non-policy-malleable

but time-varying variables decreases the probability of observing a time effect on scores in the case of reading, and slightly increases it in mathematics. Notice, instead, that adding the curriculum variable has practically no impact in the *t* statistic for mathematics (1.81 vs 1.80).

[TABLE 11 HERE]

Of course, the non-significant coefficient of the curriculum variable is not by itself sufficient evidence to dismiss the possibility that curricular reforms matter. In the following section, which draws substantively from Aloisi's (2016) doctoral thesis, the argument is developed further.

Further evidence on the impact of curricular reforms

The first source of evidence is provided by the correlations between country mean scores in reading and mathematics from PISA 2000 to PISA 2015 (Table 12). The correlations are very strong, ranging from .89 to .99. This suggests that participating countries experience very little variation in scores within and across domains over time. This is comparable to the visual impression of uniformity that one can get from Figure 1 and Figure 2. From these data alone, one could question the extent to which education policies can be expected to make a difference to international outcomes.

[TABLE 12 HERE]

Consider, for example, the .99 correlation between mathematics 2003 and 2006, which is plotted in Figure 4. The plot highlights the countries whose score changes between 2003 and 2006 were statistically significant along with the principal axis, which helps to visualise changes in scores from one PISA wave to the next. Very few countries experienced statistically significant changes in mathematics between these two PISA cycles. Could this mean that only these countries managed to introduce effective policies in those years?

Figure 4. The highest score correlation: mathematics 2003–2006

[FIGURE 4 HERE]

Figure 4 caption: 'Source: Adapted from Aloisi (2016, p. 142)'

This is not a rhetorical question. Indonesia, for example, introduced competency-based curricula alongside some systemic changes like a more stringent teacher certification system in 2003 (UNESCO-IBE, 2011); therefore, it could be that its score increase in mathematics was due to better instruction. This was the conclusion reached by Barrera-Osorio, Garcia-Moreno, Patrinos, & Porta (2011) after they showed that the greatest share of score variance between 2003 and 2006 was not captured by school and student characteristics. This seems to contrast with further results; if 'the 2006 score was partly the result of reforms, policies, strategies, and interventions that were put in place years ago, even a generation ago' (Barrera-Osorio et al., 2011, p. 11), as they claimed, the trend should have continued, but it did not. In mathematics, the 2003–2006 gain was followed by a 20-point decrease in 2009, and an overall flat trend emerges between 2003 and 2012 once socio-demographic changes are considered (OECD, 2014c). The country's performance was also stable in TIMSS between 1999 and 2007. TIMSS and PISA are not completely incomparable. For instance, Wu (2009) argued that after accounting for student age and content coverage, TIMSS and PISA 2003 share 93% of the score variance. This means that a slightly different sample and assessment content would cancel the score increase between PISA 2003 and 2006. Finally, Lenkeit & Caro (2014) showed that any long-term change in the reading and mathematics performance of Indonesia was driven by non-policy-malleable factors, such as changes in the students' socio-economic profile.

This raises a more general question on the extent to which one should take between-cycle fluctuations at face value. Because of changes in some technical properties of the assessment and

systematic error due to DIF and linking (see footnote 1 above), variations of up to 20 points may easily be within the error range. For example, the OECD noted that the new scaling approach introduced in 2015 may have affected the trend comparisons of several countries: many significant score changes may in fact be non-significant (or vice versa); Korea and Thailand "lost" 13 and 9 points respectively in reading between 2009 and 2015 because of the differences in scaling, whereas Estonia "gained" 8; Taipei's score change in mathematics between 2012 and 2015 was 15 points lower, and Albania's score change was 12 points higher (OECD, 2016b, Annex A5). These errors are only due to changes in scaling, so the total (but unknown) error is probably higher.

Given these issues, one could decide to focus only on extreme score changes. A set of outliers is shown in Table 13 and the list immediately raises questions: Why did most outliers appear in reading and why were they mostly among non-OECD countries? Did those countries implement particularly effective policies targeting reading literacy? And, if this is the case, what was the catastrophic reform adopted by Argentina in that period?

[TABLE 13 HERE]

In Argentina, new curricula were adopted between 1994 and 1998, and a framework to make the system more equitable was introduced in 2004 (UNESCO-IBE, 2007). At the same time, the country was hit by a severe economic crisis which resulted in job loss, social unrest and a default on foreign debt. If the PISA-measured change is accurate, it is not unreasonable to expect students who spent a large part of their education under great economic difficulties to do worse than those in 2000, in spite of any curricular reform.

Another telling case is that of Qatar. Its "Education for a New Era" reform of 2001 was based on OECD policy recommendations (Guarino & Tanner, 2012), with a major intervention

concerning the creation of publicly-funded independent schools (Zellman et al., 2009). The first evaluation by Zellman et al. (2009) was positive. After the reforms, Qatar improved its PISA performance by up to 75 points and by over 100 points in TIMSS 2007–2011. However, the following points should be noted.

Firstly, the performance gap between public schools (including the new independent schools) and private ones did not change: private schools did better even though they were not affected by the policy changes (Cheema, 2015). Secondly, PISA adjusted scores show that gains between 2009 and 2012 were due to socio-economic and demographic changes of the student population (OECD, 2014c). And thirdly, it is possible that the main driver behind PISA trends was the performance improvement of immigrant students which, in Qatar, are a highly-motivated majority that routinely outperforms native students (Areepattamannil, 2012; Areepattamannil, Melkonian, & Khine, 2015; Cheema, 2014). In PISA 2006, students taking the English version of the assessment displayed levels of achievement that were more closely comparable to those of pupils in English-speaking countries than to those of other Qatari students sitting the test in Arabic (Grisay, De Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007). Overall, taking the improvement of Qatar as a sign of policy effectiveness might be overly optimistic, since most variation could be due to changes in the socio-economic and demographic composition of test-takers, in line with the general findings from the multilevel models.

Of course, there are cases in which the impact of curricular reforms and interventions seems more plausible. In Ireland, for example, after a substantive score drop in PISA 2009 reading and mathematics—which, was mostly due to DIF, socio-demographic changes and changes in the engagement of test-takers (Cartwright, 2011; Cosgrove & Cartwright, 2014)—the Educational Research Centre in Dublin suggested that reforms of the primary and secondary

science curricula ‘may have mitigated the effects of changes in demography and sampling that might otherwise have lowered performance in science in PISA 2009’ (Perkins, Moran, Cosgrove, & Shiel, 2010, p. 58).

Even this explanation, however, leaves some outstanding questions. Why was the good performance in science driven by lower-performing students, if most policies targeting these students at the time did not focus on science, but on reading and mathematics? Why did the Educational Research Centre think that the curriculum had been beneficial, when evidence of its implementation was at best moderate (Eivers, Shiel, & Cheevers, 2006; Varley, Murphy, & Veale, 2008) and the Irish performance in science in previous international assessments had been stable? Notice also that mathematics performance in Ireland did not change when the curriculum was first revised in 2000, but it did when the examination changed and was brought more in line with the PISA format (Merriman et al., 2014; Perkins, Shiel, Merriman, Cosgrove, & Moran, 2013).

In another example, the French-speaking Community of Belgium improved in PISA reading 2009 and 2012. Some argue this was facilitated by a range of interventions on reading that had started in the early 2000s (Baye, Demonty, Lafontaine, Matoul, & Monseur, 2010). However, a series of important structural reforms also took place in those years. Specifically, the legislation regulating access and progression in the first year of lower-secondary education was changed, and there were changes in the examinations. This altered the PISA sample composition alongside the instruction (Lafontaine, 2014), which makes it difficult to disentangle the two components.

Discussion

This study has some limitations. Firstly, there is a comparatively large set of socio-economic and

demographic predictors for a relatively small number of countries and time points. Some jurisdictions have as few as two data points, which limits the complexity of the models that can be fit; ideally, one would like to explore how all variables interact at a global and country level. This may be why, when interaction between immigration status and first language was modelled, no statistically significant results were found. A future study might focus specifically on the interdependence of socio-economic and demographic variables, their change and the effect of this on PISA.

A second limitation is the sole focus on PISA, whereas it would be informative to carry out similar analyses with PIRLS and TIMSS. PIRLS results can be used, at least in principle, to track a population over time: since both the PIRLS and the PISA samples are meant to be representative of a student population, there is a theoretical expectation that the PIRLS population *will become* the PISA population after a few years. For TIMSS, there is a gap of only one or two years with the PISA sample, so this assumption is even more strongly justified, and the results more closely comparable. The analyses reported here could not include PIRLS and TIMSS data for reasons of time and space, but this would be a natural extension of the study, especially in light of the findings.

Finally, there are many difficulties with the quantification of curricular changes. As explained in the methodology section, there are outstanding issues surrounding the measurement of curriculum implementation or adoption, which makes it difficult to capture curricular effects with accuracy. In this study, we tried to improve the quality of subjective estimates using both documentary evidence and local experts. However, the response rate was low. This meant forgoing a quantitative analysis of changes in the reading curriculum. While we think the case

studies in the qualitative section support findings from the multilevel models, more research is needed to fine-tune the curriculum variable.

Despite these limitations, this study is unique in that it was able to show that (research question 1) changes in PISA demographics are correlated to improvement in PISA: countries that saw their socio-economic index grow compared to their own baseline also saw their PISA score increase. Conversely, decreases in scores are associated with increases in the percentage of immigrant students in the population and decreases in SES. This corroborates previous findings on the relationship between student characteristics and PISA outcomes whilst adopting a longitudinal perspective.

Of course, these are still correlational effects, but they can be compared to extant cross-sectional estimates. In PISA 2012, one standard deviation difference in ESCS was associated with a 36-point difference in mathematics score (own calculations based on OECD, 2013, Table II.A). In this study, an increase by one standard deviation in ESCS is related to a 31.7-point increase in the same domain. Similarly, in PISA 2009 when reading was the primary domain, immigrant students (both first and second generation) scored on average 42 point less than their non-immigrant peers in OECD countries, or 24 points considering all countries. In this study, the decrease was 49.6 points.

This means that our findings are not just in line with theoretical expectations and previous evidence (including some of the case studies presented here), they are of comparable magnitude. Of course, more research would improve the accuracy of these estimates, but the fact that population composition is consistently found to be associated with different outcomes both cross-sectionally and now longitudinally adds some weight to the argument that the observed (weak) trends might not be the result of education policies. This also raises a more general

question about the comparability of the PISA cohorts and how outcomes are presented. If sample composition substantively affects PISA scores, then the OECD adjusted scores should be given more prominence in the report or even replace the standard estimates.

The second and third research questions could find only a partial answer, because of the limitations acknowledged above. Overall, no strong evidence for the effectiveness of curricular reforms emerged from the regression analysis or the case studies. Despite the optimistic narrative whereby education policies can make a difference when it comes to international rankings, PISA scores are exceptionally stable over time thanks to the high psychometric quality and reliability of the assessment. Because of the scale of the endeavour and the necessary changes that have been made to PISA over the years as a result of its uptake, some error has to be expected. Previous evidence suggests that systematic error, including DIF, could amount to a fifth of a PISA standard deviation (20 points), but the overall error is probably slightly larger. This implies that most score changes between consecutive years could be due to "random" fluctuation.

When the analysis focused just on outlying countries, it was difficult to find consistent evidence in support of the impact of policy changes, including curriculum changes. Often, socio-economic and demographic changes could explain the observed improvement equally well or better than policy factors. Even if some improvement was driven by education policy reforms, as suggested for example by Irish and Belgian research, their effect seems to be modest. Science scores in Ireland increased by 14 points between 2006 and 2012, about 2.3 points per year (Effect Size = 0.023), and then fell back to 2006 levels in 2015. The reading scores of the French-speaking Community of Belgium went from 476 in 2000 points to 496 points in 2012, 1.67 points per year (Effect Size = 0.017) and less than the expected international trend of 2.34 points per year (Table 4). This is, if *all* improvement came from policy.

Of course, the claim here is not that countries should not invest in curricular reforms or other policies, but there needs to be an expectation that the payoff, at least on a national scale, will be small and incremental, rather than large and frequent.

Conclusions

This article sought to contrast the ability of policy-malleable variables to affect PISA scores to that of non-policy-malleable variables. Currently, there seems to be more evidence pointing towards a strong relationship between the socio-economic and demographic characteristics of the PISA population and country outcomes than evidence in favour of the effectiveness of education policies such as reforms of the school curriculum. Measurement issues might mask some changes, but independent within-country studies do not paint a substantively different picture: curricular reforms often just change the goals or the "vision" for education; amendments to programmes of study remain at the level of intentions and do not translate into classroom instruction; teacher training and development is lacking (OECD, 2011b). Nevertheless, many countries want to affect change and some are adjusting their accountability and monitoring systems including, in some instances, their examinations (Breakspear, 2012). This might, in the end, achieve a comparable effect of non-policy-malleable changes but there is little evidence for its impact, so far.

There was remarkable stability in the PISA results over the years. As already mentioned, they are a tribute to the quality of the assessments: year-on-year errors might account for 20 or more points, but this is just a fifth of a standard deviation, which is impressive for such a large-scale operation over this timescale. At the same time, they are a pointer to country difficulties in influencing achievement. In the US, the performance of 9-year-olds increased by 13 points on a scale going from 0 to 500 (Effect Size = 0.13) in *forty years*, that of 13-year-olds by 8 points

(Effect Size = 0.08) and that of 17-year-olds did not change at all (National Center for Education Statistics, 2013).

This is surely because managing nationwide change is difficult: all stakeholders may claim they want better results, but society hardly works in unison. Efforts in one direction are met with resistance from another, new governments often want to make their mark during their time in power, organisations and people show inertia, and policies need to compromise. In the education sector, policymaking in the past two decades has often been influenced by an economic perspective; the OECD and World Bank experts are for the great part econometricians, and the think tank sponsored by the Directorate General for Education and Culture is the European Expert Network on Economics of Education. Education production functions (Hanushek, 1979) have many theoretical merits, but sometimes rely on mathematical assumptions that cannot be met in practice, and may lead to inaccurate policy inferences. Without interdisciplinary efforts to inform policy, and whilst the fickle political influence on educational policy is strong, it is likely that flat trends will be observed for many years.

Acknowledgements

The authors would like to thank all those who willingly gave their time during interviews, particularly in Ireland and Belgium, and in responding to questionnaires.

Conflict of interest

The authors declare no conflicting interests.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory,*

Tsahkadsor, Armenia, USSR, September 2-8, 1971 (pp. 267–281). Budapest: Akadémiai Kiadó.

Aloisi, C. (2016). *The theoretical and practical value of the OECD policy advice for education*. University of Durham.

Andersen, I. G., & Jæger, M. M. (2015). Cultural capital in context: Heterogeneous returns to cultural capital across schooling environments. *Social Science Research*, *50*, 177–188. https://doi.org/10.1016/j.ssresearch.2014.11.015

Andon, A., Thompson, C. G., & Becker, B. J. (2014). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large-Scale Assessments in Education*, *2*(1), 7. https://doi.org/10.1186/s40536-014-0007-2

Areepattamannil, S. (2012). Science self-beliefs and science achievement of adolescents in Gulf Cooperation Council countries. *Educational Studies*, *38*(1), 13–17. https://doi.org/http://dx.doi.org/10.1080/03055698.2011.567058

Areepattamannil, S., Melkonian, M., & Khine, M. S. (2015). International note: Exploring differences in native and immigrant adolescents' mathematics achievement and dispositions towards mathematics in Qatar. *Journal of Adolescence*, *40*, 11–13. https://doi.org/10.1016/j.adolescence.2014.12.010

Barrera-Osorio, F., Garcia-Moreno, V., Patrinos, H. A., & Porta, E. (2011). *Using the Oaxaca-Blinder Decomposition Technique to Analyze Learning Outcomes Changes over Time: An Application to Indonesia's Results in PISA Mathematics* (No. 5584). *Policy Research Working Paper*. The World Bank. Retrieved from http://elibrary.worldbank.org/content/workingpaper/10.1596/1813-9450-5584

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Baye, A., Demonty, I., Lafontaine, D., Matoul, A., & Monseur, C. (2010). La lecture à 15 ans. Premiers résultats de PISA 2009. *Les Cahiers Des Sciences de l'Éducation*, *31*.

Belsley, D. A., Kuh, E., & Welsh, R. E. (1980). *Regression Diagnostics*. New York, NY: John Wiley & Sons, Inc.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). Westport, CT: Greenwood Press.

Braga, M., Checchi, D., & Meschi, E. (2013). Educational policies in a long-run perspective. *Economic Policy*, *28*, 45–100. https://doi.org/10.1111/1468-0327.12002

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance* (No. 71). *OECD Education Working Papers*. Paris: OECD Publishing. https://doi.org/10.1787/19939019

Brunello, G., & Rocco, L. (2013). The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores. *Economics of Education Review*, *32*, 234–246. https://doi.org/10.1016/j.econedurev.2012.10.006

Burgess, S. (2014). *Understanding the success of London's schools* (Working Paper No. 14/333). Bristol: Centre for Market and Public Organisation, University of Bristol.

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: an evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433–450. https://doi.org/10.1080/09243453.2013.812568

Cartwright, F. (2011). *PISA in Ireland, 2000-2009: Factors affecting inferences about changes in student proficiency over time*. Educational Research Centre. Retrieved from http://www.erc.ie/?p=229

Cattaneo, M. A., & Wolter, S. C. (2015). Better migrants, better PISA results: Findings from a natural experiment. *IZA Journal of Migration*, *4*(1), 18. https://doi.org/10.1186/s40176-015-0042-y

Cheema, J. R. (2014). The migrant effect: An evaluation of native academic performance in Qatar. *Research in Education*, *91*(1), 65–77. https://doi.org/10.7227/RIE.91.1.6

Cheema, J. R. (2015). The private–public literacy divide amid educational reform in Qatar: What does PISA tell us? *International Review of Education*, *61*(2), 173–189. https://doi.org/10.1007/s11159-015-9479-8

Cheng, Q., Wang, J., Hao, S., & Shi, Q. (2014). Mathematics Performance of Immigrant Students Across Different Racial Groups: An Indirect Examination of the Influence of Culture and Schooling. *Journal of International Migration and Integration*, *15*(4), 589–607. https://doi.org/10.1007/s12134-013-0300-x

Christensen, G., & Stanat, P. (2007). *Language Policies and Practices for Helping Immigrants and Second-Generation Students Succeed*. The Transatlantic Task Force on Immigration and Integration.

Cobb-Clark, D. A., Sinning, M., & Stillman, S. (2012). Migrant Youths' Educational Achievement: The Role of Institutions. *The ANNALS of the American Academy of Political and Social Science*, *643*(1), 18–45. https://doi.org/10.1177/0002716212440786

Coe, R. (2009). School improvement: Reality and illusion. *British Journal of Educational Studies*, *57*(4), 363–379. https://doi.org/10.1111/j.1467-8527.2009.00444.x

Coleman, J. S. (1966). *Equality of Educational Opportunity: Summary report*. Washington, DC: National Center for Educational Statistics, U.S. Office of Education.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA : the case of Ireland and implications for international assessment practice. *Large-Scale Assessments in Education*, *2*(2), 1–17. https://doi.org/10.1186/2196-0739-2-2

Creemers, B. P. M., & Kyriakides, L. (2015). Developing, testing, and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement*, *26*(1), 102–119. https://doi.org/10.1080/09243453.2013.869233

Creese, B., Gonzalez, A., & Isaacs, T. (2016). Comparing international curriculum systems: the international instructional systems study. *The Curriculum Journal*, *27*(1), 5–23. https://doi.org/10.1080/09585176.2015.1128346

Cummins, J. (2012). The intersection of cognitive and sociocultural factors in the development of reading comprehension among immigrant students. *Reading and Writing*, *25*(8), 1973–1990. https://doi.org/10.1007/s11145-010-9290-7

Cunha, F., & Heckman, J. J. (2010). *Investing in Our Young People*. Retrieved from http://ideas.repec.org/p/nbr/nberwo/16201.html

Doyle, A. (2008). Educational performance or educational inequality: what can we learn from PISA about France and England? *Compare: A Journal of Comparative and International Education*, *38*(2), 205–217. https://doi.org/10.1080/03057920701542057

Dronkers, J., Levels, M., & de Heus, M. (2014). Migrant pupils' scientific performance: the influence of educational system features of origin and destination countries. *Large-Scale Assessments in Education*, *2*(3), 1–28. https://doi.org/10.1186/2196-0739-2-3

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–46. https://doi.org/10.1037/0012-1649.43.6.1428

Eivers, E., Shiel, G., & Cheevers, C. (2006). *Implementing the Revised Junior Certificate Science Syllabus*. Dublin: The Stationery Office.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*(2), 121–38. https://doi.org/10.1037/1082-989X.12.2.121

European Commission/EACEA/Eurydice. (2016). The Eurydice Network. Retrieved March 14, 2017, from https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php?title=Home

Eurydice. (2011). *Mathematics Education in Europe: Common Challenges and National Policies*. Brussels: EACEA P9 Eurydice. https://doi.org/10.2797/72660

Fairbrother, M. (2013). *Two Multilevel Modeling Techniques for Analyzing Comparative Longitudinal Survey Datasets* (Working Paper). Retrieved from http://seis.bris.ac.uk/~ggmhf/MHF.MLM-longit.2013.pdf

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*(3), 305–322. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17804896

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330. https://doi.org/10.1080/0969594042000304618

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester: John Wiley & Sons, Ltd.

Goldstein, H., & Thomas, S. M. (2008). Reflections on the international comparative surveys debate. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 215–222. https://doi.org/10.1080/09695940802417368

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.

Grisay, A., Gonzalez, E. J., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. In M. von Davier & D.

Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 63–84). IEA-ETS Research Institute (IERI).

Guarino, C. M., & Tanner, J. C. (2012). Adequacy, accountability, autonomy and equity in a Middle Eastern school reform: The case of Qatar. *International Review of Education*, *58*(2), 221–245. https://doi.org/10.1007/s11159-012-9286-4

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–32. https://doi.org/10.18637/jss.v059.i09

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources*, *14*(3), 351–388.

Hanushek, E. A., & Woessmann, L. (2011). The Economics of International Differences in Educational Achievement. In E. A. Hanushek, S. Machin, & L. Wößmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 89–200). Amsterdam: North-Holland. https://doi.org/10.1016/B978-0-444-53429-3.00002-8

Harker, R., & Tymms, P. (2004). The Effects of Student Composition on School Outcomes. *School Effectiveness and School Improvement*, *15*(2), 177–199. https://doi.org/10.1076/sesi.15.2.177.30432

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic Inquiry*, *46*(3), 289–324. https://doi.org/10.1111/j.1465-7295.2008.00163.x

Heckman, J. J., & Mosso, S. (2014). The Economics of Human Development and Social Mobility. *Annual Review of Economics*, *6*(1), 689–733. https://doi.org/10.1146/annurev-economics-080213-040753

Hollins, M., & Reiss, M. J. (2016). A review of the school science curricula in eleven high achieving jurisdictions. *The Curriculum Journal*, *27*(1), 80–94. https://doi.org/10.1080/09585176.2016.1147968

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2017). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, 1–21. https://doi.org/10.1080/00313831.2016.1258726

Hopkins, D., Pennock, D., & Ritzen, J. (2008). *External evaluation of the policy impact of PISA* (26th meeting of the PISA Governing Board, 3-5 November). The Hague: OECD

Publishing. Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB/M(2008)2/REV1&docLanguage=En

Kanes, C., Morgan, C., & Tsatsaroni, A. (2014). The PISA mathematics regime: knowledge structures and practices of the self. *Educational Studies in Mathematics*, *87*(2), 145–165. https://doi.org/10.1007/s10649-014-9542-6

Kautz, T., Heckman, J. J., Diris, R., ter Weel, B., & Borghans, L. (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. Paris: OECD Publishing.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, *53*(3), 983. https://doi.org/10.2307/2533558

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, *50*(5), 530–560. https://doi.org/10.1002/tea.21086

Lafontaine, D. (2014). À petits pas dans la bonne direction. *TRACeS de ChanGements*, *215*(March-April), 4–5.

Lau, K.-C. (2009). A Critical Examination of PISA's Assessment on Scientific Literacy. *International Journal of Science and Mathematics Education*, *7*(6), 1061–1088. https://doi.org/10.1007/s10763-009-9154-2

Lenkeit, J., & Caro, D. H. (2014). Performance status and change – measuring education system effectiveness with data from PISA 2000–2009. *Educational Research and Evaluation*, *20*(2), 146–174. https://doi.org/10.1080/13803611.2014.891462

Li, P., & Redden, D. T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, *15*(1), 38. https://doi.org/10.1186/s12874-015-0026-x

Link, S. (2012). *Developing key skills: What can we learn from various national approaches?* (EENEE Analytical Report No. 14). European Commission. Retrieved from http://www.eenee.de/portal/page/portal/EENEEContent/_IMPORT_TELECENTRUM/DOCS/EENEE_AR14.pdf

Lochner, L., & Monge-Naranjo, A. (2012). Credit Constraints in Education. *Annual Review of Economics*, *4*(1), 225–256. https://doi.org/10.1146/annurev-economics-080511-110920

Luyten, H., Peschar, J., & Coe, R. (2008). Effects of Schooling on Reading Performance, Reading Engagement, and Reading Activities of 15-Year-Olds in England. *American Educational Research Journal*, *45*(2), 319–342. https://doi.org/10.3102/0002831207313345

Luyten, H., & Veldkamp, B. (2011). Assessing Effects of Schooling With Cross-Sectional Data: Between-Grades Differences Addressed as a Selection-Bias Problem. *Journal of Research on Educational Effectiveness*, *4*(3), 264–288. https://doi.org/10.1080/19345747.2010.519825

Marks, G. N. (2005). Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries. *Ethnic and Racial Studies*, *28*(5), 925–946. https://doi.org/10.1080/01419870500158943

Marks, G. N. (2006). Are between- and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research*, *48*(1), 21–40. https://doi.org/10.1080/00131880500498396

Marks, G. N. (2014). Demographic and socioeconomic inequalities in student achievement over the school career. *Australian Journal of Education*, *58*(3), 223–247. https://doi.org/10.1177/0004944114537052

Marteleto, L., & Andrade, F. (2014). The Educational Achievement of Brazilian Adolescents. *Sociology of Education*, *87*(1), 16–35. https://doi.org/10.1177/0038040713494223

McGaw, B. (2008). The Role of the OECD in International Comparative Studies of Achievement. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 223–243.

Merriman, B., Shiel, G., Cosgrove, J., & Perkins, R. (2014). *Project Maths and PISA 2012*. Dublin: Educational Research Centre.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323–35.

Motiejunaite, A., Noorani, S., & Monseur, C. (2014). Patterns in national policies for support of low achievers in reading across Europe. *British Educational Research Journal*, *40*(6), 970–985. https://doi.org/10.1002/berj.3125

National Center for Education Statistics. (2013). *The Nation's Report Card: Trends in Academic Progress 2012* (NCES 2013-456). *National Center for Education Statistics*. Washington, DC: Institute of Education Sciences, U.S. Department of Education. https://doi.org/10.1002/yd.20075

Nolan, B. (2010). *Promoting the well-being of immigrant youth*. Retrieved from http://researchrepository.ucd.ie/handle/10197/2689

Nonoyama-Tarumi, Y. (2008). Cross-National Estimates of the Effects of Family Background on Student Achievement: A Sensitivity Analysis. *International Review of Education*, *54*(1), 57–82. https://doi.org/10.1007/s11159-007-9069-5

Nusche, D. (2009). *What works in migrant education? A review of evidence and policy options* (OECD Education Working Paper No. 22). Paris: OECD Publishing.

OECD. (2001). *Knowledge and Skills for Life: first results from the OECD Programme for International Student Assessment (PISA) 2000*. Paris: OECD Publishing. https://doi.org/10.1787/9789264195905-en

OECD. (2010). *PISA 2009 Results: Overcoming Social Background* (PISA) (Vol. II). Paris: OECD Publishing. https://doi.org/10.1787/9789264091504-en

OECD. (2011a). Does participation in pre-primary education translate into better learning outcomes at school? *PISA in Focus*, *1*. https://doi.org/10.1787/5k9h362tpvxp-en

OECD. (2011b). *Teachers Matter: Attracting, Developing and Retaining Effective Teachers* (Pointers for policy development). Retrieved from https://www.oecd.org/edu/school/48627229.pdf

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

OECD. (2013). *PISA 2012 Results: Excellence through Equity* (Vol. II). Paris: OECD Publishing. https://doi.org/10.1787/9789264201132-en

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

OECD. (2014b). *PISA 2012 Results: Creative Problem Solving* (PISA) (Vol. V). Paris: OECD Publishing. https://doi.org/10.1787/9789264208070-en

OECD. (2014c). *PISA 2012 Results: What Students Know and Can Do* (Revised, Vol. 1). OECD Publishing. https://doi.org/10.1787/9789264201118-en

OECD. (2016a). PISA 2015 Database. Retrieved February 2, 2017, from http://www.oecd.org/pisa/data/2015database/

OECD. (2016b). *PISA 2015 Results: Excellence and Equity in Education* (PISA) (Vol. I). Paris: OECD Publishing. https://doi.org/10.1787/9789264266490-en

Park, H. (2008). Public policy and the effect of sibship size on educational achievement: A comparative study of 20 countries. *Social Science Research*, *37*(3), 874–887. https://doi.org/10.1016/j.ssresearch.2008.03.002

Perkins, R., Moran, G., Cosgrove, J., & Shiel, G. (2010). *PISA 2009: The Performance and Progress of 15-year-olds in Ireland* (Summary Report). Dublin: Educational Research Centre.

Perkins, R., Shiel, G., Merriman, B., Cosgrove, J., & Moran, G. (2013). *Learning for Life: The Achievements of 15-year-olds in Ireland on Mathematics, Reading Literacy and Science in PISA 2012*. Dublin: Educational Research Centre.

Pokropek, A., Borgonovi, F., & Jakubowski, M. (2015). Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learning and Individual Differences*, *42*(2015), 10–18. https://doi.org/10.1016/j.lindif.2015.07.011

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *The American Psychologist*, *53*(2), 109–120. https://doi.org/10.1037/0003-066X.53.2.109

Ramey, C. T., Ramey, S. L., & Lanzi, R. G. (2006). Children's health and education. In I. Sigel & A. Renninger (Eds.), *The handbook of child psychology* (pp. 864–892). Hoboken, NJ: John Wiley & Sons, Inc.

Raudenbush, S. W. (1989). "Centering" predictors in multilevel analysis: Choices and consequences. *Multilevel Modelling Newsletter*, *1*(2), 10–12.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Salverda, W. (Ed.). (2011). *Inequalities' Impacts* (GINI State of the Art Review No. 1). GINI Project. Retrieved from http://www.gini-research.org/articles/papers

Scheerens, J., Luyten, H., van den Berg, S. M., & Glas, C. A. W. (2015). Exploration of direct and indirect associations of system-level policy-amenable variables with reading literacy

performance. *Educational Research and Evaluation*, *21*(1), 15–39. https://doi.org/10.1080/13803611.2015.1008520

Scherer, R., & Beckmann, J. F. (2014). The acquisition of problem solving competence: evidence from 41 countries that math and science education matters. *Large-Scale Assessments in Education*, *2*(1), 10. https://doi.org/10.1186/s40536-014-0010-7

Schleicher, A. (2009). International Comparisons of Student Learning Outcomes. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second International Handbook of Educational Change* (pp. 485–504). Dordrecht: Springer Netherlands.

Schneeweis, N. (2011). Educational institutions and the integration of migrants. *Journal of Population Economics*, *24*(4), 1281–1308. https://doi.org/10.1007/s00148-009-0271-6

Schnepf, S. V. (2008). *Inequality of Learning Amongst Immigrant Children in Industrialised Countries* (IZA working paper No. 3337). Institute for the Study of Labor (IZA). https://doi.org/10.1111/j.0042-7092.2007.00700.x

Schofield, J. W. (2010). International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools. *The Teachers College Record*, *112*(5), 8–9. Retrieved from http://www.tcrecord.org/Content.asp?contentid=15684

Schofield, J. W., Alexander, K., Bangs, R., & Schauenburg, B. (2006). *Migration Background, Minority-Group Membership and Academic Achievement*. Retrieved from http://193.174.6.11/alt/aki/files/aki_research_review_5.pdf

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shapira, M. (2012). An Exploration of Differences in Mathematics Attainment among Immigrant Pupils in 18 OECD Countries. *European Educational Research Journal*, *11*(1), 68–95. https://doi.org/10.2304/eerj.2012.11.1.68

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Smith, C., & Morgan, C. (2016). Curricular orientations to real-world contexts in mathematics. *The Curriculum Journal*, *27*(1), 24–45. https://doi.org/10.1080/09585176.2016.1139498

Spilke, J., Piepho, H.-P., & Hu, X. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics*, *10*(3), 374–389. https://doi.org/10.1198/108571105X58199

Steele, F. (2008). Multilevel models for longitudinal data. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *171*(1), 5–19. https://doi.org/10.1111/j.1467-985X.2007.00509.x

Sylva, K. (2014). The role of families and pre-school in educational disadvantage. *Oxford Review of Education*, *40*(6), 680–695. https://doi.org/10.1080/03054985.2014.979581

Tan, C. Y. (2015). The contribution of cultural capital to students' mathematics achievement in medium and high socioeconomic gradient economies. *British Educational Research Journal*, *41*(6), 1050–1067. https://doi.org/10.1002/berj.3187

The World Bank. (2014). *World Bank GNI per capita Operational Guidelines & Analytical Classifications* (Database). Retrieved from http://siteresources.worldbank.org/DATASTATISTICS/Resources/OGHIST.xls

Timmermans, A. C., & Thomas, S. M. (2015). The impact of student composition on schools' value-added performance: a comparison of seven empirical studies. *School Effectiveness and School Improvement*, *26*(3), 487–498. https://doi.org/10.1080/09243453.2014.957328

UNESCO-IBE. (2007). World Data on Education: Sixth edition 2006/07. Retrieved from http://www.ibe.unesco.org/en/resources/world-data-education

UNESCO-IBE. (2011). *Indonesia* (7th ed.). *World Data on Education*. Retrieved from http://www.ibe.unesco.org/

UNESCO-IBE. (2012). World Data on Education: Seventh edition 2010/11. Retrieved March 14, 2007, from http://www.ibe.unesco.org/en/resources/world-data-education

van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36809-7

Varley, J., Murphy, C., & Veale, Ó. (2008). *Science in Primary Schools, Phase 2* (Research Report No. 11). Dublin: NCCA.

von Davier, M., Gonzalez, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 9–36). Hamburg and Princeton, NJ: IEA-ETS Research Institute (IERI).

Weick, K. E. (1976). Educational Organizations as Loosely Coupled Systems. *Administrative Science Quarterly*, *21*(1), 1. https://doi.org/10.2307/2391875

Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy & Practice*. https://doi.org/10.1080/09695940802417426

Wiseman, A. W., & Chase-Mayoral, A. (2014). Shifting the discourse on neo-institutional theory in comparative and international education. In *Annual Review of Comparative and International Education 2013* (pp. 99–126). https://doi.org/10.1108/S1479-3679(2013)0000020014

Woessmann, L. (2016). The Importance of School Systems: Evidence from International Differences in Student Achievement. *Journal of Economic Perspectives, 30*(3), 3–32. https://doi.org/10.1257/jep.30.3.3

Wolter, S. C., & Vellacott, M. C. (2003). Sibling Rivalry for Parental Resources: A Problem for Equity in Education? A Six-Country Comparison with PISA Data. *Swiss Journal of Sociology*, *29*(3), 377–398.

Wu, M. L. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *PROSPECTS*, *39*(1), 33–46. https://doi.org/10.1007/s11125-009-9109-y

Wu, M. L. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x

Xu, J., & Hampden-Thompson, G. (2012). Cultural Reproduction, Cultural Mobility, Cultural Resources, or Trivial Effect? A Comparative Approach to Cultural Capital and Educational Performance. *Comparative Education Review*, *56*(1), 98–124. https://doi.org/10.1086/661289

Zellman, G. L., Ryan, G. W., Karam, R., Constant, L., Salem, H., Gonzalez, G., … Al-Obaidli, K. (2009). *RAND report: Implementation of the K-12 Education Reform in Qatar's Schools*. Santa Monica, CA: RAND Corporation.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-87458-6

Tables

Table 1. Average reading performance across grades in 2012, with standard error and difference between two consecutive grades

Grade	**OECD average**	**SE**	**Difference(from previous Grade)**
7	341	1.7	
8	396	1.46	+55
9	466	0.79	+70
10	517	1.09	+51
11	532	1.85	+15
12/13	501*	1.4	-31*

Source: PISA 2012 compendium. *The average was calculated with just three OECD countries.

Table 2. Percentage of students in grades in Germany and PISA reading scores over the years

Year	**Grade**									**Reading score**		
	NA	*7*	*8*	*9*	*10*	*11*	*12+*	*< 9*	*≥ 9*	*Observed*	*Adjust*	*Difference (Adjust-Observed)*
2000	1.2	1.3	14.5	59.8	23.2	0.1	0.0	17.0	83.0	479	479	0
2003	0.3	1.7	15.0	59.8	23.2	0.1	0.0	16.9	83.1	491	488	-3
2006	3.6	1.5	11.9	54.5	28.2	0.3	0.0	17.0	83.0	477	484	7

2009	4.0	1.2	10.5	52.6	31.3	0.4	0.2	15.7	84.3	478	486	8
2012	0.0	0.6	10.0	51.9	36.7	0.8	0.0	10.6	89.4	508	485	-23
2015	0.0	0.5	7.7	47.3	43.1	1.5	0.1	8.2	91.8	509	486	-23

Source: PISA compendia. The numbers in the "Grade" section represent the percentage of students enrolled in each grade. The shaded column is the modal grade. The two columns "<9" and "≥9" indicate the total percentage of students enrolled below or at/above the modal grade. NA is Not Available. The "Observed" reading score is the PISA outcome for Germany reported by the OECD, whereas the "Adjust" score is the score Germany would have achieved if the proportion of students in each grade in subsequent PISA cycles had been the same as it was for PISA 2000

Table 3. Annualised change in PISA reading for a selection of countries with and without an adjustment for socio-economic and demographic changes in the student population

Country	**Base**	**SE**	**Adj**	**SE**	**Diff**	**Country**	**Base**	**SE**	**Adj**	**SE**	**Diff**
Australia	-0.1	(0.86)	-2.8	(0.3)	-2.7	Portugal	1.9	(1.47)	0.6	(0.4)	-1.3
Austria	0.0	(1.40)	-1.6	(0.4)	-1.6	Slovak Rep.	-2.9	(1.79)	-0.5	(0.5)	2.4
Belgium	1.8	(0.95)	-1.0	(0.3)	-2.8	Slovenia	1.0	(0.99)	-4.0	(0.5)	-5.0
Canada	-0.1	(0.82)	-1.5	(0.3)	-1.4	Sweden	-5.3	(1.13)	-3.2	(0.3)	2.2
Switzerland	1.8	(1.05)	0.0	(0.3)	-1.9	Turkey	5.1	(2.23)	5.0	(0.6)	-0.1
Chile	-4.0	(1.66)	1.1	(0.4)	5.1	United States	1.2	(2.01)	-1.2	(0.4)	-2.4
Czech Rep.	2.9	(1.24)	-1.7	(0.4)	-4.6	*Argentina*	7.8	(2.70)	-2.3	(0.7)	-10.1
Germany	2.2	(1.26)	0.5	(0.4)	-1.7	*Bulgaria*	10.3	(2.77)	-0.2	(0.5)	-10.4
Denmark	0.9	(0.97)	-1.0	(0.3)	-1.9	*Brazil*	2.2	(1.11)	0.1	(0.3)	-2.1
Spain	7.2	(0.97)	-1.3	(0.3)	-8.5	*Colombia*	-9.5	(3.38)	2.7	(1.0)	12.3
Estonia	7.6	(2.03)	0.7	(0.7)	-6.9	*Hong Kong*	3.7	(1.05)	1.4	(0.4)	-2.3
Finland	-4.8	(0.83)	-2.9	(0.3)	1.9	*Croatia*	4.7	(2.85)	0.8	(0.8)	-3.9
France	5.1	(1.23)	-1.3	(0.3)	-6.4	*Indonesia*	-2.1	(1.98)	2.0	(0.5)	4.1

UK	2.7	(2.63)	-0.3	(0.6)	-3.0	*Jordan*	-3.7	(2.94)	-0.7	(0.8)	3.1
Greece	3.2	(1.42)	-0.4	(0.4)	-3.6	*Liechtenstein*	-3.4	(1.38)	-0.3	(0.6)	3.1
Hungary	0.7	(1.18)	0.5	(0.4)	-0.2	*Lithuania*	4.7	(1.94)	0.3	(0.8)	-4.4
Ireland	4.2	(1.19)	-2.2	(0.3)	-6.4	*Latvia*	-2.9	(1.22)	1.0	(0.5)	4.0
Iceland	0.4	(0.64)	-2.3	(0.3)	-2.7	*Macao-China*	7.7	(0.63)	-0.4	(0.4)	-8.1
Israel	10.8	(2.19)	2.4	(0.7)	-8.4	*Montenegro*	4.7	(1.45)	2.6	(0.6)	-2.1
Italy	5.8	(0.98)	0.0	(0.3)	-5.8	*Peru*	4.7	(2.85)	4.7	(0.4)	0.0
Japan	12.1	(1.39)	3.3	(0.5)	-8.8	*Qatar*	-2.0	(0.82)	8.6	(0.5)	10.6
Korea	-5.2	(1.29)	-1.9	(0.5)	3.3	*Romania*	13.8	(2.05)	2.1	(0.7)	-11.7
Luxembourg	4.7	(0.66)	1.7	(0.3)	-3.0	*Russian Fed.*	10.7	(1.40)	-0.5	(0.4)	-11.2
Mexico	4.8	(1.04)	0.6	(0.3)	-4.2	*Serbia*	-4.8	(2.48)	6.6	(0.9)	11.4
Netherlands	2.2	(1.83)	-1.2	(0.5)	-3.4	*Taipei*	14.1	(2.44)	3.4	(0.7)	-10.7
Norway	4.6	(1.27)	-0.8	(0.3)	-5.4	*Thailand*	8.2	(1.18)	-0.3	(0.4)	-8.5
New Zealand	-1.4	(0.96)	-1.5	(0.3)	0.0	*Tunisia*	2.6	(2.20)	2.6	(0.6)	0.1
Poland	0.6	(1.29)	1.5	(0.4)	0.8	*Uruguay*	-0.1	(1.74)	-1.3	(0.5)	-1.2

Source: OECD (2014c, Table I.4.3b and I.4.4). Non-OECD countries (partner countries) are in *italic*. Column headers: "Base" = Annualised change (average change across assessments); "Adj" = Annualised change adjusted by age, gender, SES, migration status and first language; "SE" = Standard error on the base/adjusted annualised change; "Diff" = Difference between the adjusted and the non-adjusted change.

Table 4. Comparison of the null (Model 1 in Table 18) and the reference growth model (Model 3 in Table 18) for reading literacy

	NULL	**REFERENCE GROWTH**							
READING	*Estimate*	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
Intercept	468.3	461.1	(7.3)	446.6	475.5	63.48	107.2	<0.001	***

Time		2.34	(0.63)	1.09	3.64	3.70	107.2	<0.001	***
Var(Intercept)	2252.4	3201.1		2254.6	4713.7				
Cov(Time,Intercept)		-171.1		-206.9	-116.1				
Var(Time)		14.6		6.9	27.2				
Residuals	163.3	112.5		92.2	139.1				
AIC	2762.4	2712.7							
BIC	2773.6	2735.2							
-2 x logLik	2756.4	2700.7							
N.Obs (Level-1)	314	314							
N.ID (Level-2)	63	63							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$

Table 5. Comparison of the null (Model 1 in Table 18) and the reference growth model (Model 3in Table 18) for mathematical literacy

	NULL	**REFERENCE GROWTH**							
MATHEMATICS	*Estimate*	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
Intercept	470.9	467.1	(8.3)	450.5	483.7	56.04	69.41	<0.001	***
Time		1.41	(0.78)	-0.13	2.99	1.81	69.41	<0.1	.
Var(Intercept)	3175	4320.4		3081.4	6299.2				
Cov(Time,Intercept)		-270.5		-318.4	-206.6				
Var(Time)		29.8		18.6	48.0				
Residuals	102.6	47.6		38.2	60.4				
AIC	2375.6	2297.1							
BIC	2386.5	2318.8							
-2 x logLik	2396.6	2285.1							
N.Obs (Level-1)	276	276							
N.ID (Level-2)	63	63							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$

Table 6. Model 16, the significant effect of changes in SES on changes in reading outcomes

READING	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
(Intercept)	468.2	(8.8)	450.4	486.2	52.99	84.12	<0.001	***
SES.mean	33.70	(8.60)	16.59	50.83	3.92	84.12	<0.001	***
SES.change	21.95	(6.10)	9.89	34.42	3.60	84.12	<0.001	***
OECD	26.50	(9.47)	6.55	46.49	2.80	84.12	<0.01	**
Var(Intercept)	1316.6		932.1	1938.9				
Cov(SES.change,Intercept)	-719.8		-1685.4	-164.4				
Var(SES.change)	1403.0		664.4	2616.1				
Residuals	106.8		87.6	131.9				
AIC	2670.4							
BIC	2700.4							
-2 x logLik	2654.4							
N.Obs (Level-1)	312							
N.ID (Level-2)	63							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$

Table 7. Model 19, the relationship between average changes in SES and mathematics compared to the effect of time

MATHEMATICS	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
(Intercept)	483.1	(11.5)	460.3	506.3	42.18	97.35	<0.001	***
SES.mean	41.21	(10.48)	20.28	62.30	3.93	97.35	<0.001	***
SES.change	21.00	(6.01)	8.97	33.10	3.49	97.35	<0.001	***
Time	-1.04	(0.97)	-2.97	0.96	-1.07	97.35	0.29	

OECD	11.50	(11.58)	-12.93	36.05	0.99	97.35	0.32	
Var(Intercept)	2933.5		2041.2	4386.2				
Cov(Time,Intercept)	-188.9		-354.9	-82.5				
Var(Time)	24.3		14.5	40.2				
Residuals	46.9		37.6	59.5				
AIC	2266.2							
BIC	2298.7							
-2 x logLik	2248.2							
N.Obs (Level-1)	275							
N.ID (Level-2)	63							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$

Table 8. Model 48b for reading

READING	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
(Intercept)	775.6	(142.9)	481.6	1066.8	5.43	107.15	<0.001	***
FEM.mean	-619.25	(285.06)	-1198.98	-34.89	-2.17	107.15	<0.05	*
SES.mean	27.92	(8.64)	10.66	45.38	3.23	107.15	<0.01	**
Time	0.94	(0.90)	-0.89	2.80	1.05	107.15	0.29	
SES.change	17.59	(7.73)	3.90	34.50	2.28	107.15	<0.05	*
IMM.change	-49.60	(23.30)	-96.44	-2.73	-2.13	107.15	<0.05	*
OECD	21.49	(9.19)	1.91	41.12	2.34	107.15	<0.05	*
Var(Intercept)	1726.5		1054.3	2562.2				
Cov(Time,Intercept)	-73.3		DNC	DNC				
Cov(SES.change,Intercept)	-476.5		DNC	DNC				
Var(Time)	3.9		0.2	21.5				
Cov(SES.change,Time)	37.7		-1.5	194.8				

Var(SES.change)	495.1		44.7	1981.5				
Residuals	103.2		84.6	127.2				
AIC	2669.2							
BIC	2721.6							
-2 x logLik	2641.2							
N.Obs (Level-1)	312							
N.ID (Level-2)	63							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$; DNC = Did not converge

Table 9. Model 48b for mathematics

MATHEMATICS	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
(Intercept)	916.5	(169.0)	572.0	1257.7	5.42	108.4	<0.001	***
FEM.mean	-859.49	(338.05)	-1541.05	-171.32	-2.54	108.4	<0.05	*
SES.mean	39.74	(10.61)	18.00	61.91	3.75	108.4	<0.001	***
Time	-1.97	(0.95)	-4.06	0.05	-2.07	108.4	<0.05	*
SES.change	31.74	(7.93)	15.69	49.58	4.00	108.4	<0.001	***
IMM.change	-52.83	(29.75)	-112.62	6.76	-1.78	108.4	<0.1	.
OECD	8.61	(10.92)	-14.86	32.74	0.79	108.4	0.43	
Var(Intercept)	2174.0		1357.0	3505.1				
Cov(Time,Intercept)	-51.3		-55.6	167.5				
Cov(SES.change,Intercept)	-903.9		DNC	DNC				
Var(Time)	16.6		4.5	41.8				
Cov(SES.change,Time)	-28.1		-17.5	274.1				
Var(SES.change)	1046.4		123.9	3162.3				
Residuals	41.4		32.5	53.4				
AIC								
BIC								

-2 x logLik								
N.Obs (Level-1)	275							
N.ID (Level-2)	63							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$; DNC = Did not converge

Table 10. A multilevel growth model of PISA mathematics accounting for changes in the mathematics curriculum

MATHEMATICS	*Estimate*	*SE*	*CI low*	*CI high*	*t*	*DF*	*p-value*	*SIG*
(Intercept)	463.6	(8.4)	446.8	480.3	55.30	66.26	<0.001	***
Time	1.50	(0.84)	-0.15	3.22	1.80	66.26	0.08	.
Curriculum.yr	1.31	(1.24)	-1.21	3.83	1.05	66.26	0.3	
Var(Intercept)	4078.9		2871.0	6038.5				
Cov(Time,Intercept)	-289.7		-214.4	-359.7				
Cov(Cr.yr,Intercept)	-98.6		DNC	DNC				
Var(Time)	33.4		20.7	53.8				
Cov(Cr.yr,Time)	10.0		DNC	DNC				
Var(Cr.yr)	3.1		DNC	DNC				
Residuals	44.9		35.7	57.5				
AIC	2149.9							
BIC	2185.5							
-2 x logLik	2129.9							
N.Obs (Level-1)	259							
N.ID (Level-2)	59							

*** $p < 0.001$; ** $0.001 \leq p < 0.01$; * $0.01 \leq p < 0.05$; . $0.05 \leq p < 0.1$. DNC = Did not converge

Table 11. Overall effect sizes (in $|t|$ values)

	Time only (Model 3)	Time + demographics (Model 48b)	Time + curriculum
Mathematics	1.81	2.07	1.80

Table 12. Pearson's correlations between country mean scores in reading and mathematics from PISA 2000 to PISA 2015.

		Reading 2000–2015						**Mathematics 2003–2015**				
	Year	*00*	*03*	*06*	*09*	*12*	*15*	*03*	*06*	*09*	*12*	*15*
Reading 00–15	*00*	1.00										
	03	0.94	1.00									
	06	0.91	0.97	1.00								
	09	0.95	0.96	0.97	1.00							
	12	0.93	0.92	0.94	0.97	1.00						
	15	0.92	0.93	0.91	0.94	0.95	1.00					
Maths 03–15	*03*	0.92	0.95	0.94	0.93	0.92	0.91	1.00				
	06	0.92	0.94	0.96	0.95	0.94	0.91	0.99	1.00			
	09	0.93	0.94	0.94	0.96	0.96	0.93	0.98	0.98	1.00		
	12	0.89	0.90	0.91	0.93	0.96	0.92	0.95	0.96	0.98	1.00	
	15	0.89	0.89	0.88	0.90	0.94	0.95	0.95	0.95	0.96	0.98	1.00

Source: own calculation. The correlations were computed with pairwise deletion of missing data, which means that the subsets being correlated between each pair of assessment cycles varied by size and countries, from a minimum of 32 (between 2000 and 2003 scores) to a maximum of 63 (in 2012).

Table 13. A list of country score changes above ± 40 points

CHANGE	COUNTRY	OECD?	DOMAIN	PERIOD
-44.53	Argentina	No	Reading	2000–2006
40.09	Japan	Yes	Reading	2006–2012
40.99	Serbia	No	Reading	2006–2009
41.67	Romania	No	Reading	2006–2012
42.49	Liechtenstein	No	Reading	2000–2003
42.61	Peru	No	Reading	2000–2009
45.10	Serbia	No	Reading	2006–2012
45.11	Albania	No	Reading	2000–2012
47.13	Israel	Yes	Reading	2006–2012
50.16	Qatar	No	Maths	2006–2009
57.07	Peru	No	Reading	2000–2012
58.49	Qatar	No	Maths	2006–2012
59.502	Qatar	No	Reading	2006–2009
75.29	Qatar	No	Reading	2006–2012

Source: Reproduced from Aloisi (2016, pp. 139–140)

Appendix

Table 14. PISA country sample used in this study, with reading and mathematics scores

			Reading						Mathematics				
ID	CNT	NAME	2000	2003	2006	2009	2012	2015	2003	2006	2009	2012	2015
1	AUS	Australia	528	525	513	515	512	503	524	520	514	504	494
2	AUT	Austria	492	491	490	*NA*	490	485	506	505	*NA*	506	497
3	BEL	Belgium	507	507	501	506	509	499	529	520	515	515	507
4	CAN	Canada	534	528	527	524	523	527	532	527	527	518	516
5	CHL	Chile	410	*NA*	442	449	441	459	*NA*	411	421	423	423
6	CZE	Czech Republic	492	489	483	478	493	487	516	510	493	499	492
7	DNK	Denmark	497	492	494	495	496	500	514	513	503	500	511

8	EST	Estonia	NA	NA	501	501	516	519	NA	515	512	521	520
9	FIN	Finland	546	543	547	536	524	526	544	548	541	519	511
10	FRA	France	505	496	488	496	505	499	511	496	497	495	493
11	DEU	Germany	484	491	495	497	508	509	503	504	513	514	506
12	GRC	Greece	474	472	460	483	477	467	445	459	466	453	454
13	HUN	Hungary	480	482	482	494	488	470	490	491	490	477	477
14	ISL	Iceland	507	492	484	500	483	482	515	506	507	493	488
15	IRL	Ireland	527	515	517	496	523	521	503	501	487	501	504
16	ISR	Israel	452	NA	439	474	486	479	NA	442	447	466	470
17	ITA	Italy	487	476	469	486	490	485	466	462	483	485	490
18	JPN	Japan	522	498	498	520	538	516	534	523	529	536	532
19	KOR	Korea	525	534	556	539	536	517	542	547	546	554	524
20	LUX	Luxembourg	NA	479	479	472	488	481	493	490	489	490	486
21	MEX	Mexico	422	400	410	425	424	423	385	406	419	413	408
22	NLD	Netherlands	NA	513	507	508	511	503	538	531	526	523	512
23	NZL	New Zealand	529	522	521	521	512	509	523	522	519	500	495
24	NOR	Norway	505	500	484	503	504	513	495	490	498	489	502
25	POL	Poland	479	497	508	500	518	506	490	495	495	518	504
26	PRT	Portugal	470	478	472	489	488	498	466	466	487	487	492
27	SVK	Slovak Republic	NA	469	466	477	463	453	498	492	497	482	475
28	SVN	Slovenia	NA	NA	494	483	481	505	NA	504	501	501	510
29	ESP	Spain	493	481	461	481	488	496	485	480	483	484	486
30	SWE	Sweden	516	514	507	497	483	500	509	502	494	478	494
31	CHE	Switzerland	494	499	499	501	509	492	527	530	534	531	521
32	TUR	Turkey	NA	441	447	464	475	428	423	424	445	448	420
33	GBR	United Kingdom	NA	NA	495	494	499	498	NA	495	492	494	492
34	USA	United States	504	495	NA	500	498	497	483	474	487	481	470
35	*ALB*	*Albania*	349	NA	NA	385	394	405	NA	NA	377	394	413
36	*ARG*	*Argentina*	418	NA	374	398	396	NA	NA	381	388	388	NA
37	*BRA*	*Brazil*	396	403	393	412	410	407	356	370	386	391	377
38	*BGR*	*Bulgaria*	430	NA	402	429	436	432	NA	413	428	439	441
39	*TAP*	*Chinese Taipei*	NA	NA	496	495	523	497	NA	549	543	560	542
40	*COL*	*Colombia*	NA	NA	385	413	403	425	NA	370	381	376	390
41	*CRI*	*Costa Rica*	NA	NA	NA	443	441	427	NA	NA	409	407	400
42	*HRV*	*Croatia*	NA	NA	477	476	485	487	NA	467	460	471	464
43	*HKG*	*Hong Kong-China*	525	510	536	533	545	527	550	547	555	561	548
44	*IDN*	*Indonesia*	371	382	393	402	396	397	360	391	371	375	386
45	*JOR*	*Jordan*	NA	NA	401	405	399	408	NA	384	387	386	380
46	*KAZ*	*Kazakhstan*	NA	NA	NA	390	393	NA	NA	NA	405	432	NA
47	*LVA*	*Latvia*	458	491	479	484	489	488	483	486	482	491	482
48	*LIE*	*Liechtenstein*	483	525	510	499	516	NA	536	525	536	535	NA
49	*LTU*	*Lithuania*	NA	NA	470	468	477	472	NA	486	477	479	478
50	*MAC*	*Macao-China*	NA	498	492	487	509	509	527	525	525	538	544
51	*MYS*	*Malaysia*	NA	NA	NA	414	398	NA	NA	NA	404	421	NA
52	*MNE*	*Montenegro*	NA	NA	392	408	422	427	NA	399	403	410	418
53	*PER*	*Peru*	327	NA	NA	370	384	398	NA	NA	365	368	387
54	*QAT*	*Qatar*	NA	NA	312	372	388	402	NA	318	368	376	402
55	*ROU*	*Romania*	428	NA	396	424	438	434	NA	415	427	445	444
56	*RUS*	*Russian Federation*	462	442	440	459	475	495	468	476	468	482	494
57	*SRB*	*Serbia*	NA	NA	401	442	446	NA	NA	435	442	449	NA
58	*QCN*	*Shanghai-China*	NA	NA	NA	556	570	NA	NA	NA	600	613	NA
59	*SGP*	*Singapore*	NA	NA	NA	526	542	535	NA	NA	562	573	564
60	*THA*	*Thailand*	431	420	417	421	441	409	417	417	419	427	415
61	*TUN*	*Tunisia*	NA	375	380	404	404	361	359	365	371	388	367

62	*ARE*	*United Arab Emirates*	*NA*	*NA*	*NA*	423	432	434	*NA*	*NA*	411	423	427
63	*URY*	*Uruguay*	*NA*	434	413	426	411	437	422	427	427	409	418

Source: PISA database. Non-OECD (partner) countries are in *italic*.

Table 15. Expert questionnaire on education reforms

Dear colleague,

Peter Tymms and I are investigating the impact of curriculum policy changes on student results in international assessments such as PISA and TIMSS.

We would be extremely grateful if you could answer two very brief questions to the best of your knowledge. Your responses will be treated confidentially and anonymously.

1. Were there any major changes in the READING curriculum that might have affected (positively or negatively) the PISA outcomes between 2000 and 2015? If yes, in what years were these curricular changes implemented?

2. Were there any major changes in the MATHEMATICS curriculum that might have affected (positively or negatively) the PISA or the TIMSS (Grade 8) outcomes between 2000 and 2015? If yes, in what years were these curricular changes implemented?

Thank you. Your contribution is critical to this study. Please feel free to expand your answers as necessary, forward this email to other colleagues, or get in touch with us for any query you might have.

Table 16. Modal grade of the PISA population 2000–2015

ID	Name	Modal	ID	Name	Modal	ID	Name	Modal
1	Australia	10	22	Netherlands	10	43	Hong Kong-China	10
2	Austria	10	23	New Zealand	11*	44	Indonesia	10*
3	Belgium	10	24	Norway	10	45	Jordan	10
4	Canada	10	25	Poland	9	46	Kazakhstan	9
5	Chile	10	26	Portugal	10	47	Latvia	9*
6	Czech Republic	10*	27	Slovak Republic	10	48	Liechtenstein	9
7	Denmark	9	28	Slovenia	10	49	Lithuania	9
8	Estonia	9	29	Spain	10	50	Macao-China	10
9	Finland	9	30	Sweden	9	51	Malaysia	10
10	France	10	31	Switzerland	9	52	Montenegro	9
11	Germany	9	32	Turkey	10	53	Peru	10
12	Greece	10	33	United Kingdom	11	54	Qatar	10
13	Hungary	9	34	United States	10	55	Romania	9
14	Iceland	10	35	Albania	10	56	Russian Federation	9*
15	Ireland	9	36	Argentina	10	57	Serbia	9
16	Israel	10	37	Brazil	9*	58	Shanghai-China	10
17	Italy	10	38	Bulgaria	9	59	Singapore	10

18	Japan	10	39	Chinese Taipei	10	60	Thailand	10*
19	Korea	10	40	Colombia	10	61	Tunisia	10
20	Luxembourg	9	41	Costa Rica	9	62	United Arab Emirates	10
21	Mexico	10	42	Croatia	9	63	Uruguay	10

*The modal grade was fixed at this value for trend analyses, but varied with time.

Table 17. Values taken by the curriculum variable for mathematics in each country and PISA cycle

	Reform year	**PISA 2003**	**PISA 2006**	**PISA 2009**	**PISA 2012**	**PISA 2015**
Australia	00-09	0.2	0.2	0.2	0.2	0
Austria	00, 09	0.8	0	0.2	0.8	0
Belgium	01(Fr), 07(Fr)	0.2	0.6	0.2	0.6	0
Canada	97-01, 11	0.2	1	0	0	1
Chile	02, 10	0.4	1	0	0.6	1
Czech Republic	96, 07	0	0	0.6	1	0
Denmark	00, 07	0.8	0	0.6	1	0
Estonia	97–98, 13	1	0	0	0	0.6
Finland	06, 10	0	0.2	0.8	0.6	1
France	08	0	0	0.4	1	0
Germany	05	0	0.4	1	0	0
Greece	03	0.2	0.8	0	0	0

Hungary	00, 07	0.8	0	0.6	1	0
Iceland	07	0	0	0.6	1	0
Ireland	00, 12	0.8	0	0	0.2	0.8
Israel	04, 10	0	0.6	1	0.4	1
Italy	02, 08, 13	0.4	1	0.4	1	0.6
Japan	02, 10	0.4	1	0	0.4	1
Korea	02	0.4	1	0	0	0
Luxembourg	07	0	0	0.6	1	0
Netherlands	98, 06, 10	1	0.2	0.8	0.6	1
New Zealand	10	0	0	0	0.6	1
Norway	97, 06	0	0.2	0.8	0	0
Poland	02, 09	0.4	1	0.2	0.8	0
Portugal	07	0	0	0	1	0
Slovak Republic	97, 08	0	0	0.4	1	0
Slovenia	03, 11	0.2	0.8	0	0.4	1
Spain	06	0	0.2	0.8	0	0
Sweden	00, 11	0.8	0	0	0.4	1
Turkey	98, 08	1	0	0.4	1	0
United Kingdom	00, 07-08, 10, 12	0.2	0	0.4	1	0.2
United States	01, 11	0.6	1	0	0.4	1
Argentina	96, 06	0	0.2	0.8	0	0
Bulgaria	98	1	0	0	0	0

Chinese Taipei	97, 03, 08	0.2	0.8	0.4	1	0
Colombia	03	0.2	0.8	0	0	0
Croatia	11	0	0	0	0.4	1
Hong Kong-China	02	0	1	0	0	0
Indonesia	04, 06	0	0.6	1	0	0
Jordan	05	0	0.4	1	0	0
Kazakhstan	03, 12	0.2	0.8	0	0.2	0.8
Latvia	05, 13	0	0	1	0	0.6
Liechtenstein	99, 05, 10	0	0.4	1	0.6	1
Lithuania	97, 03, 10	0.2	0.8	0	0.6	1
Malaysia	03	0.2	0.8	0	0	0
Montenegro	06	0	0.2	0.8	0	0
Peru	08	0	0	0.4	1	0
Qatar	05	0	0.4	1	0	0
Romania	99, 04, 09	1	0.6	1	0.8	0
Russian Federation	00, 04, 11	0.8	0.6	1	0.4	1
Serbia	06	0	0.2	0.8	0	0
Singapore	02, 08	0.4	1	0.4	1	0
Thailand	01, 08	0.6	1	0	0.6	1
Tunisia	00, 08	0.8	0	0.4	1	0
United Arab Emirates	00, 07, 10	0.8	0	0.4	1	1

Table 18. Model-building approach

Model n.	Time	OECD	FEM mean	FEM change	SES mean	SES change	IMM mean	IMM change	FL mean	FL change	GRADE mean	GRADE change
1												
2	F											
3	R											
4		F										
5	F	F										
6	R	F										
7			F									
8			F									
9		F	R									
10			F	F								
11			F	R								
12			F									
13		F	F									
14					F	F						
15					F	R						
16		F			F	R						
17	F				F	F						
18	R				F	F						
19	R	F			F	F						
20	F				F	R						
21	R				F	R						
22							F					
23								F				
24								R				
25	R							R				
26									F			
27										F		
28										R		
29											F	
30												F
31												R
32												R
33							F		F			
34								F		F		
35	R							F		F		
36			F		F							
37	R		F		F							
38			F		F							R
39			F		F			R				
40	R		F		F			F				
41	R		F		F			R				
42	R	F	F		F			R				
43			F		F	R						
44	R		F		F	R						
45	R	F	F		F	R						
46			F		F	R		R				
47	R		F		F	F		R				

48	R		F		F	R		F				
48b	R	F	F		F	R		F				
49	R		F		F	R		F				
50			F		F	R		R				R
51	R	F	F					R				
52	R	F	F		F (not split)			F				
53	R	F	F		F (not split)			F				
54	R		F		F (not split)			R				
55	R	F	F		F (not split)			R				
56	R	F	F		F (not split)			F				
57	R	F	F		F (not split)			F				

F = Modelled as a fixed effect (equal for all countries); R = Modelled as a random effect (varying by country); not split = The variable was not decomposed as per equations (9) and (10) in this case. Model 1 is the Null model and Model 3 is the Reference model

Figure 5. Trends in PISA reading 2000–2015

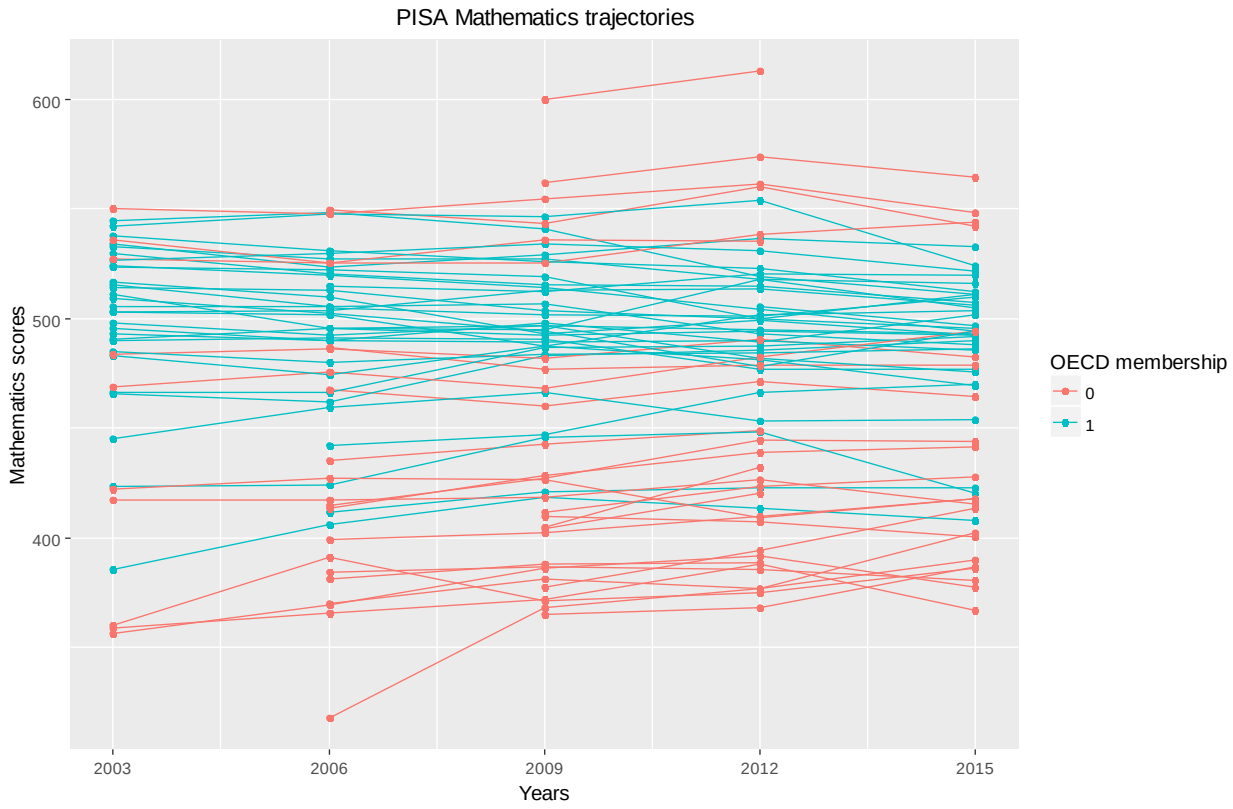

Figure 6. Trends in PISA mathematics 2003–2015