Dire Straits-Education Reforms

Ideology, Vested Interests and Evidence

Montserrat Gomendio
and Jose Ignacio Wert

https://www.openbookpublishers.com

ISBN Paperback: 978-1-80064-930-9

ISBN Hardback: 978-1-80064-931-6

ISBN Digital (PDF): 978-1-80064-932-3

ISBN Digital ebook (EPUB): 978-1-80064-933-0

ISBN Digital ebook (AZW3): 978-1-80064-934-7

ISBN XML: 978-1-80064-935-4

ISBN HTML: 978-1-80064-936-1

DOI: 10.11647/OBP.0332

Cover image: Kimberly Farmer, A collection of books (2017), https://unsplash.com/photos/lUaaKCUANVI. Cover design: Jeevanjot Kaur Nagpal.

4. ILSAs: Do They Count?

4.1 What Do International Metrics Measure?

International large-scale assessments (ILSAs) were started by the International Association for the Evaluation of Educational Achievement (IEA). In 1995 the first TIMSS (Trends in International Mathematics and Science Study) survey was conducted in over forty countries at five grade levels (third, fourth, seventh and eighth grades, as well as the final year of secondary school). Students were assessed in mathematics and science and a parallel analysis of the school curricula was conducted with the aim of finding links between the two. TIMSS was subsequently designed as a "quasi-longitudinal study" assessing a cohort of students in fourth grade and four years later again in eighth grade. The survey has been conducted every four years and in 2015 the sample included fifty-seven countries and seven regional jurisdictions. A different survey, "TIMSS Advanced", targets the final year of secondary school (twelfth grade in many countries) and assesses student achievement in advanced mathematics and physics. The IEA has also developed an international survey to assess reading literacy among fourth grade students every five years (PIRLS: Progress in International Reading Literacy Study). It started in 2001 with thirty-five participating countries and in 2016 it covered fifty countries and eleven regions.

In 2000, the Organisation for Economic Cooperation and Development (OECD) started its own survey, PISA (Programme for International Student Assessment). It assesses three domains: scientific literacy, mathematics and reading literacy, and in each cycle one of these is the main domain. It runs every three years, which means that each subject is treated as the main domain every nine years, and evaluates fifteen-year-olds irrespective of their grade. Thus, it compares students

 https://doi.org/10.11647/OBP.0332.04

who have been at school for different lengths of time depending on the age at which compulsory schooling starts in different countries. In addition, the proportion of fifteen-year-olds in different grades included in the sample varies depending on the rate of grade repetition in each country. The geographical coverage of PISA has grown from forty-three countries/economies in 2000 to seventy-nine in 2018.

Thus, the three major international large-scale assessments (PIRLS, TIMSS and PISA) measure the same domains (reading, mathematics and science), but the methodology, length of the cycle and the target population (as defined by student age or grade) are different. The IEA surveys (PIRLS and TIMSS) sample all students in each classroom focusing on specific grades and have been designed to analyse the extent to which students have acquired curriculum-based content (Martin *et al.*, 2016; Mullis *et al.*, 2016 and 2017). On the other hand, PISA samples fifteen-year-olds in different grades (eighth, ninth, tenth and eleventh grades) and has defined its goal as an assessment of how the knowledge and skills acquired are applied to meet real-life challenges and to solve problems in unfamiliar settings (OECD, 2001 and 2019c). Unlike IEA surveys, PISA does not attempt to relate differences in curricular content between countries to student outcomes. Instead, the knowledge and skills considered relevant for "knowledge-based societies" are decided by groups of experts. This approach recognises explicitly that PISA scores are the result of the combined impact of school, home and the social environment, making the links between PISA results and school policies more tenuous. Despite this, PISA claims to be more policy-oriented than IEA's assessments and in fact PISA publications include many analyses to try to identify which good practices distinguish well-performing countries (OECD 2016b, 2019c, and 2019d). Participating countries first included mostly OECD members, i.e. largely high-income countries, but then expanded beyond the OECD perimeter to include low- and middle-income countries, which (as we shall see) required adjustments to the methodology.

The value of ILSAs lies in providing international benchmarks, which allow comparisons between countries in student performance using the same metrics. The fact that the main ILSAs measure student performance periodically also makes it possible to analyse trends over time. Initially these international surveys faced skepticism because of

the deeply ingrained belief that education systems are too different to allow any meaningful comparison.

Critics also argued that their methodologies were flawed and that differences between countries focused too much on a narrow set of subjects and failed to capture important outcomes of the education systems. As an increasing number of countries has joined these international surveys, trust in them has grown, as has their influence on the narrative around good practices in education policy. This is mainly because they have promoted much-needed analyses on the good practices that lead to improvements in certain countries and the policies that top-performing countries have implemented (Cordero *et al.*, 2013 and 2018; Gustafsson and Rosen, 2014; Hanushek and Woessmann, 2011 and 2014; Hopfenbeck *et al.*, 2018; Johansson, 2016; Klieme, 2013; Lockheed and Wagemaker, 2013; Strietholt *et al.*, 2014). It is important to remember that drawing causal inferences remains controversial mainly due to the cross-sectional nature of the samples.

Over the years the public profile of international surveys, and PISA in particular, has grown. Media and policymakers eagerly await the results of each cycle to find out how their countries perform in relation to others and whether student outcomes have improved or declined since the last cycle. This has increased awareness among policymakers and citizens of the quality of their education systems. It has also contributed to a shift in the debate about education, from an emphasis on inputs (amount of resources invested) to an emphasis on outputs (student performance).

But the heightened media and political impact also has its drawbacks. It inevitably leads to a very narrow focus on the ranking between countries, and to oversimplistic hypotheses concerning the impact of policies implemented by different governments. In the worst-case scenario, it also leads to destructive blame games when countries perform poorly; this is a major concern among low- and middle-income countries which expect to perform badly but wish to be able to measure the progress of their education systems. Thus, international surveys, particularly PISA, have become powerful tools in the political debate. This is a reality that must be acknowledged and raises the bar for ILSAs to be reliable and accountable.

4.2. ILSAs: What Do They Tell Us?

Differences between and within Countries

Differences between Countries

International surveys have revealed large differences in student performance between countries which are equivalent to several years of schooling, showing that differences in the quality of education systems are much larger than expected. The difference between the top-performing country and the lowest-performing country is equivalent to more than seven years of schooling according to PISA (OECD, 2016a); in other words, what an eight-year-old has learned in a country with a good-quality education system is roughly equivalent to what a fifteen-year-old knows in a low-performing system. Thus, differences in the quality of education systems mean that students in different countries end compulsory education with a shocking difference in knowledge and skills. These findings show that years of schooling is not a reliable proxy measure for students' levels of knowledge and skills, because how much students learn in a school year differs greatly from country to country. In other words, education systems differ to a large degree in their effectiveness, or productivity, which are measured as learning progress made by students per time unit.

One might also question the usefulness of viewing the educational attainment of adults as the main measure of a country's human capital and talent, since students at the end of any educational stage (including compulsory education, i.e. lower-secondary) will have very different levels of performance in different countries. The worrying conclusion is that, while the expansion of educational opportunities has led to high returns in terms of skills and knowledge in countries with good-quality education systems, universal access to school and improved enrolment rates at higher levels of education have delivered very poor results in terms of human capital growth among low-performers.

As the number of countries participating in ILSAs has increased over time, the top performers have changed with different cycles, but some trends remain very solid. Perhaps the most telling and consistent of them has been the excellent performance among students from East Asia. The first country from the region to participate in international surveys was

Japan, which achieved the very top positions from the beginning: second in mathematics in 1964 (First International Mathematics Study among thirteen-year-olds) and first in the second mathematics study (1980–1982). It also outperformed most participating countries in science from very early on, achieving first in the First Science Study in 1970 for both ten-year-old and fourteen-year-old students (1970 First Science Study) and maintaining the very top position in the next science survey in 1984. As other countries from East Asia joined, they were recognised as top performers: Hong Kong in 1982 in mathematics and science, Singapore and South Korea in science in 1984.

The outstanding levels of performance among East Asian countries became even more apparent in TIMSS 1995, when the four top performers were all from the region: Singapore, Korea, Japan and Hong Kong, both in eighth-grade (out of thirty-nine participating countries) and fourth-grade mathematics (out of twenty-five countries) (Harmon *et al*. 1997). In TIMSS 1999 the best-performing countries in mathematics were Singapore, Korea, Taiwan, Hong Kong, and Japan (out of thirty-eight participants) (Martin *et al.*, 2000; Mullis *et al.*, 2000). These countries remained top performers in mathematics and science in the next cycles of TIMSS and, as other countries from the region joined, most emerged as top performers. In the last cycles (2015 and 2019) the top-performing countries in mathematics (fourth grade and eighth grade) were Singapore, Korea, Hong Kong, Chinese Taipei and Japan (Mullis *et al.*, 2016; Mullis *et al.*, 2020). The gap in performance between this block and other participating countries was substantial. These countries were also among the top performers in science but did not occupy all of the top positions as a block in this subject (Martin *et al.*, 2016; Mullis *et al.*, 2020).

The results in PISA confirm the same trends, but with a slightly different composition of participating countries from East Asia. In the first PISA cycle (2000) Japan and Korea (both OECD members) achieved the top positions in the ranking and remained top performers in successive cycles (OECD, 2001; OECD, 2019c; OECD, 2019d). Hong Kong-China became the top performer when it joined in the next cycle (2003), and also remained among the top performers in the following year (OECD, 2004). Shanghai and Singapore broke the mould from 2009 onwards (OECD, 2010), with Shanghai outperforming all other countries in all three domains in 2009 and 2012 (OECD, 2014c),

and Singapore becoming the top performer in all three domains in 2015 (OECD, 2016a). In 2018, China chose to be represented by four provinces (Beijing, Shanghai, Jiansu and Zheijiang, or B-S-J-Z), which outperformed all other countries in all domains (OECD, 2019c). Just to give an idea of the extent to which countries in Asia excel, the top performers in mathematics in PISA 2012 were (in descending order): Shanghai-China, Singapore, Hong Kong-China, Chinese Taipei, Korea, Macao-China and Japan. The difference between Shanghai-China and Peru (the highest and lowest performers) was equivalent to over six years of schooling.

The exceptions from East Asia are countries with considerably lower GDP and levels of investment per student, such as the Philippines and Indonesia, which have tended to performer very poorly, and Thailand and Malaysia, which have performed somewhat better both in TIMSS and PISA.

Perhaps the most meaningful change over time is that most of the countries that have joined these surveys in the last cycles are low- and middle-income countries, so the number of low-performing countries has increased substantially over time and the gap between the top- and lowest-performing countries has enlarged. This is particularly pronounced in PISA, which started as a survey designed mainly for OECD countries and then made a proactive effort to expand to include a much broader range of countries. To exemplify the magnitude of this transformation, suffice it to say that in the first PISA cycle, Mexico (an OECD member) and Brazil were the lowest performers and the only two countries from Latin America. By 2018, ten Latin American countries had joined, and they all performed well below OECD levels, but Mexico outperformed twenty-four other participating countries, such as Philippines, Kosovo, Lebanon and Morocco, despite its own performance not actually having improved.

In many ways, Latin America is the opposite of East Asia, because countries in this region show consistently low levels of performance and little improvement over time. In fact, in all international comparisons Latin American students are among the lowest performing of all participating countries. In PISA 2015, all Latin American countries were ranked significantly below the OECD average (OECD, 2016a). Within this group the highest performer was Chile and the lowest was the

Dominican Republic. The difference in performance between Chile and the OECD average was over one year of schooling, while the difference between the lowest performer (Dominican Republic) and the top performer (Singapore) was over five years of schooling. Furthermore, there is no relationship between years of schooling and student performance across Latin American countries, given that students seem to make little progress in learning in each individual year that they spend at school (Hanushek and Woessmann, 2015). Thus, policies which try to compensate for the low quality of education by adding more years of compulsory schooling will not lead to any significant improvements in student performance.

Do Different ILSAs Tell the same story?

Studies comparing how countries perform in both PISA and TIMSS have shown that country averages are strongly correlated in the two years in which both surveys were conducted simultaneously: 2003 (Wu, 2010) and 2015 (Klieme, 2016). Thus, both surveys seem to provide similar information on how students from different countries perform in mathematics and science. Relatively minor differences can partly be attributed to the design of the surveys, and partly to the fact that all students in TIMSS are in the same grade and have experienced the same years of schooling, while PISA targets fifteen-year-olds irrespective of grade (countries show major differences in the rate of grade repetition and in the age at which compulsory schooling starts) (Wu, 2010).

The general picture that arises from both surveys shows three clusters of countries: East Asian countries are top-performers, European and North American countries together with New Zealand and Australia are mid-performers, and countries in Africa, Latin America and the Near Middle East are poor performers.

Looking in more detail at countries´ performances, it seems that East Asian countries perform comparatively better in TIMSS than PISA, while some Nordic and English-speaking countries seem to perform better in PISA than TIMSS (Klieme, 2016; Wu, 2010). Many hypotheses have been put forward to explain these differences, but the most widely accepted suggests that students in East Asian countries may focus more on learning the curricular content, which is more accurately captured by

TIMSS, while in Nordic and English-speaking countries more emphasis is placed on problem solving, which may be better assessed by PISA. A relationship has also been reported between fourth graders' performance in PIRLS 2011 and reading performance in PISA 2018, which are assumed to correspond roughly to the same birth cohort of students (OECD, 2019c). It is worth highlighting that East Asian countries as a group do not outperform countries in Europe and English-speaking countries in reading to the extent they do in mathematics and science.

Differences within Countries

As we have seen, differences between countries in student performance are large, but differences between students within the same country are even larger. In many decentralised countries, major differences between the quality of different regions' education systems are a chief source of inequality. Differences in levels of student performance lead to large differences in the skill levels of the adult population which, in turn, are related to employment levels, economic growth and prosperity (Cheshire *et al.*, 2014; OECD, 2016d; OECD, 2019b).

Many low-performing regions fall into the so-called 'low skills trap' since their labour markets are based on low-skilled jobs and there are few incentives for education systems to become more demanding and efficient; in these contexts students feel that the returns of education are low and tend to leave school as soon as they reach the age at which attendance is no longer compulsory, in many cases before having attained the compulsory education diploma (OECD, 2015c; OECD, 2017b).

These differences have been analysed in more detail in Spain and Italy, since both countries show large regional differences, and most regions have an extended sample in PISA that allows meaningful comparisons. To simplify what is a very complex scenario, in both countries, poorer regions in the south tend to perform badly, while richer regions in the north achieve better student outcomes. However, in Spain some relatively poor regions in the north (Galicia) and the centre (Castilla y León) are the best performers in PISA. Also, as we shall see, there is no relationship at the regional level between the level of investment in education and student outcomes.

Data at the regional level show that the PISA average for Spain hides major differences between regions (Gomendio, 2021; OECD, 2015c; Wert, 2019). Thus, in PISA 2015 the difference between the top-performing region in science (Castilla y León) and the lowest performing region (Andalucía) is the equivalent of more than one year of schooling. Eleven out of the seventeen Spanish regions perform above the OECD average, and six perform below it. In Italy, regional differences are even larger and are equivalent to more than two years of schooling. Thus, mediocre average results at the national level conceal diversity within countries where some regions are actually top performers, while others are low performers according to PISA rankings.

In Spain, rates of grade repetition are very high (2015: 36.1% in Spain vs 13% OECD average) and show large regional differences, with the number of students repeating at least one grade ranging from 25% to 45% (Gomendio, 2021; Wert, 2019). It is important to remember that PISA samples fifteen-year-olds in different grades and that students who repeat a grade perform much worse. To understand the extent to which grade repetition influences PISA scores in Spain, it should be noted that in PISA 2015 the sample included 67.9% of fifteen-year-olds in tenth grade, while 23.4% were one year behind and 8.6% were two years behind (OECD, 2016a). Regional differences in the rate of grade repetition explain to a large extent the variation in student outcomes in PISA. Grade repetition, in turn, is a good proxy for rates of early school leaving. Students who repeat grades are much likelier to drop out of school later on, become NEETs and suffer high rates of unemployment. Thus, there is a clear relationship between regional levels of student performance, differences in rates of grade repetition and early school leaving which, in turn, have a major impact on rates of youth unemployment.

In contrast, in Italy grade repetition rates are lower but truancy rates are higher, and both explain to a large extent differences between regions in PISA scores (Hippe *et al.*, 2018). These large regional differences are by no means unique to Southern Europe. In Canada the difference between the top-performing region (Alberta) and the lowest-performing region (Saskatchewan) is also larger than one year of schooling, while in the United States the level of performance of Massachusetts is so much

higher than Puerto Rico that it is equivalent to more than three years of schooling (OECD, 2016a).

In order to understand the level of variation in student performance within countries, it is important to look at the proportion of students that reach different levels of proficiency. International surveys establish thresholds below which students are assumed to have failed to achieve the most basic skills, and above which students can be considered excellent. A comparative analysis with data from seventy-seven countries which participated in different international surveys shows that among top-performing countries the share of students who do not acquire basic skills in mathematics and science is less than 5%, while among low-performing countries the share of functionally illiterate students ranges from 40% to 80% (Hanushek and Woessmann, 2015). On the other hand, the share of excellent student ranges from 10% to 20% among top-performing countries to almost non-existent among low-performing countries. These findings highlight the fact that low-performing countries are not only unable to allow excellent students to reach their full potential, but they also fail to equip a large share of students with the most basic skills.

4.3. ILSAs: Trends Over Time

For most countries, comparing how students perform over time is crucial, since this is the main source of information through which to infer whether the implementation of certain policies has had the expected positive outcomes. Changes over time are also the main focus of political battles, as different political parties engage in debates about which government was responsible for improvements or declines. However, in this crucial aspect ILSAs differ to a large extent, leaving policymakers to decide which survey is more reliable or better-suited to measure the impact of specific policies.

In a nutshell, while PISA finds no significant change over time (2000–2018), both TIMMS and PIRLS detect improvements in most participating countries. To understand the extent to which these surveys diverge in the changes that they detect or fail to detect, we will just provide an overall summary avoiding excessive technicalities.

When the last two cycles of PISA are compared, mean performance for the sixty-three countries that participated in both 2015 and 2018 remained stable in reading, mathematics and science (OECD, 2019c). When changes are analysed separately for participating countries, we find that only four countries improved in reading between 2015 and 2018, while thirteen declined and forty-six remained stable. Furthermore, in twenty-four countries out of the sixty-three, no changes were observed in any of the three domains. When longer periods are considered, there are sixty-five countries that participated both in PISA 2018 and at least one other PISA cycle before PISA 2015. Out of these sixty-five countries, seven countries/economies improved in all three domains, seven declined in all domains, and thirteen showed no changes in any of the three domains. When only OECD countries are considered, PISA detects no changes between 2000 and 2018.

In contrast, data from PIRLS and TIMSS show clear improvements overall. From 2015 until 2019, out of the forty-five countries participating in both cycles in TIMSS (fourth grade mathematics) fourteen improved, only eight declined and twenty-three did not change substantially; among eighth grade students (mathematics), out of thirty-three countries, thirteen improved and only four declined (Mullis *et al.*, 2020). When a longer timeframe is considered (2007–2019), out of twenty-one countries participating in fourth grade mathematics, fourteen improved and none declined, and out of twenty-three countries participating in eighth grade mathematics, sixteen improved and only two declined. Similar patterns of change over time emerged for student performance in science.

In summary, when the last two cycles are considered only 6% of participating countries improve according to PISA, while 40% of participating countries improve in TIMSS. When changes since the first cycle are considered, only 10% of PISA participating countries improve, while 50% of TIMSS participating countries improve (since 1995). Obviously, the participating countries in both surveys are not identical, although there is a substantial degree of overlap. Despite this, the contrast between the flatness in PISA trends and the positive TIMSS and PIRLS trends points to differences between surveys, rather than between participating countries.

This conclusion is supported by evidence from specific countries which shows that the trends over time identified by TIMSS and PISA are strikingly different. Australia and South Korea show a substantial decline over successive cycles in mathematics according to PISA (2003–2018), while they show clear improvements over the same period according to TIMSS (2003–2019). In the case of Chile, Japan, Lithuania and the US, PISA findings show no improvement between 2003 and 2018, while TIMSS reveals major improvements over a similar timeframe (2003–2019). Thus, a pattern emerges in which TIMSS uncovers improvements where PISA fails to detect changes, or TIMSS shows no changes where PISA finds declines. In both cases, TIMSS reveals a more positive evolution over time for many countries than PISA.

Further analyses reveal that until 2011/2012 the number of countries which improved or declined was more or less the same when comparing TIMSS and PISA. However, since 2015 more countries showed declines in PISA and, in some cases, the same countries showed improvements in TIMSS (Klieme, 2016). The conclusion from this study is that the lack of sensitivity to changes shown by PISA is the consequence of a new mode of assessment adopted in PISA 2015. Detailed studies conducted within countries have shown that, in countries such as Germany, the methodological changes implemented in PISA 2015, which include moving from paper to computer-based assessments, as well as changes in the way students' scores were calculated, had a negative impact (Robitzsch *et al.*, 2020). Further changes introduced in 2018 seem to have been even more disruptive, leading to the withdrawal of PISA results for countries such as Spain and Vietnam.

In 2018 PISA introduced substantial changes aimed at improving the sensitivity of the survey at low levels of student performance, in order to deal with the problems generated by the fact that many low- and middle-income countries had joined PISA. Since these countries tended to have poor levels of performance, PISA was of little use beyond stating the obvious. Furthermore, governments faced strong criticism from political opponents when the poor results were made public, so the political costs of engaging with PISA were high, and the benefits limited. This became a constraint to the ambitious targets that PISA had set in terms of increasing the number of participating countries. In an effort to maximise the amount of information that could be provided

to the increasing share of low-performing countries, PISA introduced a number of changes to improve its sensitivity at low levels of student performance. However, these changes may have been made at the expense of the consistency required to detect changes over time and, in at least a few countries, at the expense of the reliability of the PISA 2018 results.

In 2018, PISA merged some items from the PISA for Development framework (OECD, 2017c), which was developed to measure low levels of performance among fifteen-year-olds (in and out of school) in low- and middle-income countries. These included a new section on "reading-fluency", which in theory was designed to assess in more detail the reading skills of students in the lower proficiency levels (OECD, 2019c, p. 270).

In practice, this section seemed designed to assess whether students had the cognitive skills to distinguish if short sentences make sense or not, rather than reading fluency. Examples include short sentences such as "airplanes are made of dogs", to which students had to reply 'yes' or 'no'. A significant proportion of students in some countries, such as Spain, gave patterned responses (all 'yes' or all 'no'), but then continued onto more difficult items and responded according to their true level of proficiency (OECD 2019c, Annex A9). Although the section on Spain claims that this problem is unique to this country (OECD, 2019c, p. 208), in a different section the OECD reports that this pattern of behaviour ("straightlining") was also present in over 2% of the high-performing students in at least seven other countries (including top performers such as South Korea) and even higher in countries such as Kazakhstan (6%) and the Dominican Republic (5%) (OECD, 2019c, p. 202). No data are provided on the prevalence of straightlining behaviour among all students. The OECD recognises that it is possible that some students "did not read the instructions carefully" or that "the unusual response format of the reading fluency tasks triggered disengaged response behavior".

Any problems with this initial section may have had major implications for the whole assessment because in 2018 PISA introduced another major change: it was designed for the first time as an "adaptive test". This means that students were assigned to comparatively easy or comparatively difficult stages later on, depending on how they

performed at initial stages. This contrasts with PISA 2015 and previous cycles, when the test format did not change over the course of the assessment based on how students had performed at previous stages. It is also worth mentioning that this adaptive testing cannot be used in the paper-based assessments. Thus, any anomalies in the first section improperly labelled as "reading fluency" may have led not only to low scores, but more importantly to mistakes in how students were assigned to easy or difficult tests for the rest of the assessment.

In the case of Spain, the OECD withdrew the results for the main domain (reading) from the official launch of its report in December 2019, after complaints from several regional governments about major problems with the data. This led to unprecedented concerns about the unreliability and unaccountability of PISA (*El Mundo*: "La Comunidad de Madrid pide a la OCDE que retire todo el informe PISA por errores de un calibre considerable: Toda la prueba está contaminada", 29 Nov 2019; *El Mundo*: "Las sombras de PISA: ¿hay que creerse el informe tras los errores detectados?", 2 December 2019; *El País*: "Madrid pide que no se publique ningún dato de PISA porque todo está contaminado", 30 November 2019; *La Razón*: "Madrid llama chapucera a la OCDE por el informe PISA", 2 December 2019). Surprisingly, the OECD then published the same results in July 2020, although it made it clear that the results were not comparable to previous cycles (OECD, 2019c, Annex A9). The OECD claimed that Spanish students were "negatively disposed towards the PISA test and did not try their best to demonstrate their proficiency", thus failing to assume any responsibility. Furthermore, according to the OECD the same results that were initially withdrawn were published months later at the request of the Ministry of Education, generating widespread concern about the OECD giving in to political pressures from a government that wished to use unreliable data to justify the need for an education reform that had been announced well before PISA 2018 data were even available (*El País*: "Falta de interés y cansancio en mitad de los exámenes finales: así explica la OCDE las "anomalías" del informe PISA en España", 23 July 2020; *El Mundo*: "La OCDE atribuye los errores del PISA a la "disposición negativa" de los alumnos españoles por coincidir varios exámenes", 23 July, 2020).

In 2018, the PISA results for a different group of countries were also deemed unreliable or did not meet 'PISA technical standards'.

These included Vietnam, which was praised as a top performer in PISA 2012, and a number of relevant countries such as Hong Kong (China), Netherlands, Portugal, the United Kingdom and the United States. While the results for Vietnam have not been published by the OECD, the government has informed national media of the PISA scores that were provided by the OECD (*Viet Nam News*: "VN gets high scores but not named in PISA 2018 rankings", 6 Dec 2019). The results for the other countries were published by the OECD because they were accepted as "largely comparable" (OECD, 2019c), but raised major concerns (for a detailed analysis of the UK see Jerrim, 2021).

The OECD concludes that the lack of progress detected by PISA is the result of countries not implementing the right policies, by which it means the policies that the OECD recommends (OECD, 2019c). In the words of OECD Education Director, Andreas Schleicher, PISA has:

> become the world´s premier yardstick for evaluating the quality, equity and efficiency of school systems, and an influential force in education reform. It has helped policy makers lower the cost of political action by backing difficult decisions with evidence—but it has also raised the political cost of inaction by exposing areas where policy and practice are unsatisfactory. (OECD, 2019c)

In our opinion, it is unfair to make governments responsible for the apparent lack of progress made by countries, given that the OECD has not provided satisfactory explanations about the impact of methodological changes and the reliability of the scores. For many governments, PISA is a high-stakes exam of the kind that PISA itself no longer supports when making recommendations to countries about student assessments. By participating in PISA, governments expose themselves to huge media impact and to the blame games that are so often part of the political debate. This means that the results will have major implications about how particular education policies or reforms are perceived by societies. In exchange, PISA must remain accountable when the results generate reasonable doubts. As an evidence-based organization, the OECD should also examine the possibility that PISA has lost the sensitivity required to detect the changes unveiled by other ILSAs.

Let us assume, just for the sake of the argument, that the OECD is right in that countries have not improved the quality of their education systems because they have failed to implement those policies which,

according to PISA, lead to better student outcomes. This would imply that the OECD has not achieved the self-proclaimed status of a global player in education, since countries have not listened to or acted upon the lessons that PISA has to offer (indeed, this is PISA´s own conclusion). But there is an even more worrying hypothesis not contemplated by PISA, that those countries which have followed the OECD's recommendations have not noticed improvements in student performance.

In the following section, we will examine in more detail the argument that PISA has substantially changed the balance of costs and benefits derived from implementing education reforms by "backing difficult decisions with evidence". Given the doubts about the reliability of PISA 2018, we will focus mostly on earlier cycles.

4.4. Evidence from ILSAs on Effective Policies

In this section, we will review the evidence available on education policies which has led to improvements in student outcomes, focusing mostly on the data generated by ILSAs. While the OECD has made great efforts to make PISA an "influential force in education reform", the IEA does not focus on drawing conclusions about which policies lead to better outcomes, beyond more specific analyses of the curriculum. Thus, over successive cycles, the number of analyses aimed at identifying which policies are linked to better student outcomes has grown in PISA publications. These include links between student outcomes, which are measured directly, and factors about the school and home environment which are addressed in questionnaires answered by students and principals. Since the OECD advises governments directly and PISA has a substantial media impact, these conclusions have reached many policymakers and have influenced public opinion.

However, PISA statistical analyses are almost exclusively correlations which cannot establish causal effects. To overcome this limitation, a number of researchers have used more sophisticated statistical techniques to more reliably identify causal factors, and many of them have included in their analyses not just PISA data, but also data from other ILSAs. In this section we will review the policy recommendations elaborated by the OECD based on PISA data and analyse how robust the evidence on which they are based is. We will also look at the conclusions

of non-OECD researchers who have independently analysed data from ILSAs. We do not intend to review all of the available literature, so we will only refer to studies which are not based on ILSA data when they are required to support or refute specific conclusions. Our main purpose is to assess the robustness of those policy recommendations as evidence-based pieces of advice.

It is important to note that as the number of PISA participating countries has increased over time, some conclusions have changed. It is also worth pointing out that as the diversity of participating countries has increased, the pertinence of extrapolating good practices directly between countries has been questioned.

Investment in Education

Since 2006 PISA has considered the relationship between student outcomes and countries' GDP or investment in education (measured as investment per student from the ages of six to fifteen). Although the percentage of GDP allocated to education is a widely used measure, it is heavily influenced by demographic factors. Thus, a similar percentage of GDP invested in education will result in high investments per student in countries with ageing populations, and low investments per student in countries with larger cohorts of young people. To avoid this confounding factor, we focus on levels of investment per student.

Investment per student can be analysed in purely monetary terms (absolute investment), in monetary terms corrected by purchasing power parity (investment relative to prices in any given country), or in relation to either per capita GDP or per capita public expenditure (investment relative to income or public expenditure). A combination of all three metrics provides quite a complex outlook, not just on how much a country invests in education, but also on how education is prioritised (or not) in public policy.

In 2006 and 2009, no relationship was found between investment in education and student performance (OECD, 2007 and 2010), but as more countries joined, from 2012 to 2015 a clear pattern began to emerge. PISA data show that below a certain threshold (which is established as 50,000 USD, after accounting for purchasing power parities [PPP]), there is a strong positive relationship between investment per student

and performance in PISA, which all but disappears above this threshold (OECD, 2014a and 2016a).

Most of the countries below this threshold are low- and middle-income countries that have not reached universal access to education and/or countries where students only spend a few years in school. Therefore, these countries are still at the stage where further investment is needed to build schools, provide them with the necessary resources and hire more teachers. In PISA 2012 and 2015, this included all participating Latin American countries and others such as Thailand, where only between 50% and 70% of fifteen-year-olds are enrolled at school. Still, the fact that all countries below this threshold have low levels of student performance means that, for the share of fifteen-year-olds that remain in school, the quality is very low. This implies that there is a minimum level of investment below which the limited resources are not enough to develop a quality education system. But there seems to be one exception.

The only outlier is Vietnam, a country which, despite having one of the lowest levels of investment per student, achieved PISA scores similar to those of countries above the threshold, such as Germany and Canada, and higher than the United States, Portugal or Sweden. However, while among top-performing countries most, if not all, students are in school at the age of fifteen, in Vietnam less than 50% of fifteen-year-olds are enrolled at school. It is in fact the only top-performing country that has such a small proportion of fifteen-year-olds in school, followed by China (B-S-J-G) where 64% of fifteen-year-olds are in school. It seems reasonable to assume that fifteen-year-olds who have already left school (or never attended in the first place) have a low level of performance, so including these out-of-school students in the sample would dramatically lower the performance of these two countries which PISA regards as top-performers. Furthermore, it raises the question of whether these countries achieve such high performance levels precisely because disadvantaged students or students from rural areas are not integrated into the schooling system (OECD, 2016a).

The most revealing finding is the fact that, above a relatively low level of investment, there is no relationship whatsoever between investment per student and student performance. These countries represent a wide range of levels of investment, from just over 50,000 to almost 200,000

USD (PPP) invested per student between the ages of six and fifteen. Thus, countries that invest up to four times more than others do not achieve better student outcomes. The group of countries which is above this threshold is large and diverse: all of Europe (including the UK), the United States, Canada, Australia, New Zealand and most countries in East Asia. Some countries which are just above the threshold in terms of investment are top performers (such as Estonia), while others invest much more and obtain poor results (such as Luxembourg). And while certain intermediate factors may explain such differences (Estonia is a very homogeneous and egalitarian society; Luxembourg is very unequal and the share of immigrant students, at 55%, is by far the largest among OECD countries) they clearly demonstrate that investment *per se* is no guarantee of success.

Further support for this conclusion comes from studies carried out at a more granular level, which have compared the level of investment per student for different regions within the same country. One of the advantages of these studies is that regions in any one country are more similar to each other in terms of their education system, its institutional structure and the mechanisms that define how it is funded, than different countries participating in large international surveys. Thus, studies which compare regions avoid many of the confounding factors that studies which compare countries encounter. In Spain, where regions decide how much to invest in education from a lump sum transferred by the central government (which covers education, health and social affairs), there are remarkable differences in the level of investment per student between regions: some regions invest twice as much as others. Despite these large differences, there is no relationship between levels of investment in education and student performance (Gomendio, 2021; Wert, 2019).

Another way to analyse the impact of investment on student outcomes is to look at increases or decreases over time in levels of investment, and whether or not they are aligned with student outcomes. In the 2012 cycle, PISA found no relationship between changes in investment between 2003 and 2012 and changes in PISA scores. Although the vast majority of countries significantly increased education investment over this period, many of them experienced a decline in student performance. An independent analysis of changes in expenditure per student from 2000

until 2010 and changes in PISA reading scores from 2000 until 2012 also suggests no relationship between the two (Hanushek and Woessmann, 2015).

Most reviews of the vast amount of work which has analysed in different ways the impact of educational expenditure on student outcomes concludes that this lack of relationship is a very robust finding (Hanushek, 2003; Hanushek and Woessmann, 2011; Woessmann, 2007a). Detailed studies of changes in investment in specific countries over longer periods of time using more sources of data have also failed to show that those changes lead to changes in student outcomes. In the US, there have been dramatic increases in spending per student from 1960 to 2016 (expenditure has more than quadrupled over that period), but student performance has remained rather stable and similar to the OECD average (Hanushek, 2021). In Spain, the education budget doubled between 2000 and 2009, but mediocre student outcomes did not significantly change during this period (Gomendio, 2021; Wert, 2019). Conversely, after the global financial crisis, regions in Spain started to reduce the budgets assigned to education and did not increase them again until 2016; contrary to all expectations, student performance had improved in mathematics and science in TIMSS 2015, further improvements in reading were detected in PIRLS 2016, and PISA 2015 also detected improvements of a lesser magnitude (Gomendio, 2021).

The evidence showing that investment *per se* is unrelated to student outcomes is the most solid evidence available about what does not work in education. These findings contradict the most widely accepted premise in any debate on education: the higher the input (investment) the better the outcome (student performance). They also contradict the reverse premise: that budget cuts in education will inevitably lead to a decline in student outcomes. To explain why the total amount of resources is not a determinant of student outcomes, it has been argued that what is most important is how resources are invested. But what does this mean?

To analyse this claim in more detail, it is important to understand how investment in education is allocated. More than 90% of total expenditure on education is devoted to current expenditure (average across OECD countries) given that education is labour-intensive. In primary and secondary education, around 61% of current expenditure is allocated to

funding teachers, about 16% is allocated to compensating other staff and 23% to other expenditure, such as meals and transportation for students (OECD, 2016b). Thus, the majority of resources assigned to education depend on two factors: the number of teachers (which is, in turn, the product of the number of students and the ratio of students per teacher) and teacher salaries. In the next sections, we will analyse evidence of the impact of teacher salaries and class size on student performance.

Teacher Quality and Salaries

It is widely accepted that the success of any education system relies to a large extent on the quality of its teachers. However, the concept has proven to be elusive (Gomendio, 2017). The best evidence comes from longitudinal studies, which have tracked student performance over time. These "value-added analyses" have shown that there are large differences between teachers in terms of classroom outcomes: differences in the progress made by students with weak teachers when compared to those with great teachers may represent as much as one grade (Hanushek, 1992; Hanushek and Rivkin, 2010 and 2012; Rivkin *et al.*, 2005; Rockoff, 2004). In turn, these differences in learning progress thanks to exposure to effective teachers have a large impact on access to higher levels of education (university) and higher income (Chetty *et al.*, 2014). Thus, we know that teachers make a difference, but what makes teachers different?

There are so far no concrete conclusions in the quest to identify which traits make teachers effective. Both PISA´s own analyses and others have found no relationship between traits which are easy to quantify, such as teacher education, certification or professional development, and student outcomes (Chingos and Peterson, 2011; Glewwe *et al.*, 2014; Hanushek and Rivkin, 2006; Hanushek and Woessmann, 2015; Harris and Sass, 2011). This is probably due to the fact that in most countries, teachers hold university degrees and have some form of professional development, but these similarities mask large differences in the training requirements, as well as the quality and content of degrees and in-service training.

According to school principals who participated in PISA 2015, the average student in OECD countries attends a school where 84% of

teachers have been fully certified, with some countries reaching over 90% (such as Ireland, Japan and Australia), and a few falling below 60% (such as Mexico and Chile) (Gomendio, 2017). However, the fact that teachers hold university degrees in most countries should not lead to the conclusion that they all have similar levels of knowledge and skills. There are major national differences in terms of how demanding education systems' entry requirements are, and the levels of knowledge and skills that trainee teachers acquire by the end of their degree. While in Latin America, students applying to education degrees generally have lower grades in university entrance exams than students applying to other degrees (Bruns and Luque, 2015), in other countries education degrees are much more selective. It is also important to consider that there are large differences in quality: the skills acquired by university graduates in low-quality systems are lower than those of secondary students in high-quality systems (OECD, 2016d).

Similarly, while some countries have very effective models of professional development, others do not. Most countries follow a rather traditional model offering courses and workshops which do not have any impact on their teaching practices or knowledge levels (Gomendio, 2017; Opfer, 2016). But some top performers have developed very effective models of teacher training and professional development: in Singapore, teachers are entitled to 100 paid hours of professional development each year, and the National Institute of Education, as well as the Academy of Teachers, provides high-quality training for the upskilling of teachers (Gomendio, 2017).

Given the importance of having effective teachers to achieve high levels of student performance, it is surprising how little direct information there is comparing teachers´ knowledge and skills in different countries. The one international survey that assessed teachers´ knowledge focused on mathematics (TEDS-M, 2008) and showed large differences between countries both in primary and secondary education, with teachers in Singapore and Switzerland reaching the highest scores, and teachers in Chile and Philippines receiving very low scores (Tatto, 2014; Tatto *et al.*, 2012).

It is remarkable that, despite the efforts made by ILSAs to understand effective teaching practices, none has been able to find direct links to student outcomes. The OECD Teaching and Learning International

Survey (TALIS) asks teachers about their working conditions and subjective perceptions of their "effectiveness", but there is no major programme in place to link these findings with student outcomes (as measured by PISA). The so-called TALIS-PISA link has a very limited number of participating countries and does not provide clear-cut results (OECD, 2021a).

The only relevant international comparative analysis in this context has used data from the survey of adult skills (PIAAC), which does not include teachers as a target subpopulation, but does include a small proportion in the general sample. This study has found a strong correlation between the skill levels of teachers (PIAAC data) and student performance (PISA data) across countries (Hanushek *et al.*, 2019). These findings clearly show that teachers' skill levels differ to a large extent, and that these differences do matter, since only highly skilled teachers are able to achieve good student outcomes. They also show that degrees and certificates are not good indicators of the real skill levels of teachers, because of large differences in the quality of those degrees between countries. There are different ways in which teachers in different countries may achieve different skill levels. Since the survey of adult skills (PIAAC) shows that there are large differences between countries in terms of skill levels of adult populations, teachers´ skills could be merely a reflection of these population differences. In other words, teachers may be more skilled in some countries just because they are part of an adult population with higher skill levels.

Alternatively, teachers in different countries may represent different levels of skills within their country's range: in some countries the education system may allow university graduates with relatively low skill levels to become teachers, while more demanding education systems may ensure that (among those with a university degree) only those who have achieved high skill levels can become teachers. This study shows that differences between countries in terms of teachers' skill levels are mainly the result of policy choices on where teachers fall on the spectrum of a country´s university graduates. Out of thirty-one countries included in this analysis, teachers in Finland have the highest skill levels because they score highly amongst Finnish graduates, who already perform higher than many other countries. In contrast, Denmark has a similar skills distribution, but teachers have lower skill levels than

other university graduates. If we consider countries where the skill levels of the population are lower, such as Chile, teachers have relatively high skill levels compared to other university graduates, while in Italy, teachers come from the lower range of the skills distribution spectrum.

These findings highlight the importance of establishing mechanisms and incentives to ensure that good candidates are attracted into the teaching profession and that their education and training is demanding. In the policy debate, the conclusion of an influential McKinsey report has now become a *cliché*: "the quality of an educational system cannot exceed the quality of its teachers" (Barber and Mourshed, 2007). This report also concludes that "the top-performing systems recruit their teachers from the top third of each cohort graduate from their school system" (Barber and Mourshed, 2007). However, the findings of a wide-ranging comparative study show that some countries, such as Singapore and South Korea, perform better than expected from the skills of the teaching force, while others such as Sweden or Greece perform worse than expected (Hanushek *et al.*, 2019).

Thus, while teachers have a major impact on student outcomes, other aspects of the education system also play an important role, such as high curricular standards and effective student assessments. The findings also show that in no country do teachers fall at the very top of the national distribution of graduates. However, this finding should be treated with caution, since the PIAAC survey assesses the adult population from the age of sixteen to sixty-four. Countries which have started to implement policies to attract highly skilled graduates into the profession during the last decades may only see the impact of this selective approach among young teachers. If these countries have a large proportion of old teachers included in the overall sample, the effects of new policies may be diluted. Thus, these results do not dispute the fact that in countries like Singapore and Finland, in the last decades only 20% of secondary school students who apply to teacher education programmes are accepted, and all applicants fall within the top range of student performance (Barber and Mourshed, 2007).

Given that only highly skilled teachers can achieve good student outcomes, it is often assumed that high salaries are required to attract good candidates into the teaching profession and to retain the most effective teachers. This has led to a substantial increase in teachers´

salaries among OECD countries between 2000 and 2010. For countries for which data are available, teacher salaries continued to increase from 2010 until 2014, despite the financial crisis (OECD, 2016b). Although PISA´s correlations have found no significant relationship between teachers´ salaries and student outcomes, in some cycles it has claimed that they are linked (e.g. OECD, 2013a, Fig. IV.1.10, p. 43), while in others it has recognised that they are not (OECD, 2016b and 2019c), which makes policy recommendations rather confusing. According to PISA 2015, countries such as Finland, Japan or Canada achieve good student outcomes with average teachers´ salaries (relative to per capita GDP), while countries such as the United Arab Emirates, Qatar or Mexico have poor student performance despite higher relative teacher salaries (PISA 2016d, Vol. II, Fig II.6.7). Other studies looking at the relationship between student outcomes and teachers´ salaries have found no clear link (Hanushek and Woessmann, 2015).

It has been suggested that incentives such as performance-related teacher pay may be more important than absolute values. Cross-country studies have indeed found this to be the case, since students have better outcomes in countries where teachers receive performance-related pay, and introducing performance-related pay has improved outcomes in a number of countries (Atkinson *et al.*, 2009; Hanushek and Woessmann, 2015; Podgursky and Springer, 2007; Woessmann, 2011). Performance-related pay could represent an incentive for existing teachers to work harder (referred to as 'effort' margin), or it could make the teaching profession more attractive to candidates who are likely to benefit from such working conditions while making the system more effective at retaining effective teachers (the so-called 'selection' margin).

The evidence seems to suggest that the latter is more important in leading to an improvement in student outcomes. Several studies have shown that, when teachers have high initial salaries but flat trajectories (i.e. small increases thereafter), teaching turns into a low-risk/low-returns profession that is unattractive for highly skilled and ambitious individuals (Bruns and Luque; 2015; Corcoran, Evans and Schwab, 2004; Eide, Goldhaber, and Brewer, 2004; Fredriksson and Ockert, 2007; Hernani-Limarino, 2005; Hoxby and Leigh, 2004). Thus, incentive-based policies that enhance teacher accountability can improve student outcomes at a fraction of the cost of reforms that uniformly increase

teacher salaries across the board (Bruns *et al.*, 2011, Bruns and Luque, 2015).

Class Size

Class size explains to a large extent why, beyond a certain threshold, the amount of resources invested in education is unrelated to student outcomes. The measure of reducing class size involves the highest cost, because it requires hiring more teachers, and more often than not the benefits (if any) are small.

It is a widespread assumption that large classes may constrain the degree of attention that teachers may devote to each of their students, that this may lead to less support for struggling students, and overall to poor concentration among students, or even a lack of discipline. As a result, large class sizes are assumed to lead to poor student outcomes. This is such a strong belief that governments have made huge financial investments to decrease class size over time. Between 2005 and 2014, the average class size among OECD countries decreased and, despite the 2008 financial crisis, class size continued to decrease between 2010 and 2014 (OECD, 2016b). As a consequence, the average class size in public schools among OECD countries was twenty-one in primary schools and twenty-three in lower-secondary schools in 2014 (OECD, 2016b).

There have been massive investments to decrease class size over time despite the lack of evidence linking it to student outcomes. No single PISA cycle has shown a significant correlation between the two variables when countries are compared, but the policy recommendations have changed over time. Comparisons within countries in PISA 2015 showed that, in most countries, students in schools with larger classes tend to perform better (OECD, 2016b and 2016c). However, these analyses should be treated with caution because it is unclear how or whether they accounted for the fact that larger class sizes are found in schools in rich neighbourhoods, in urban areas and in public schools.

In the same PISA cycle, a comparison between countries showed no relationship between class size and student performance, since some top performers in East Asia have classes of over thirty-five students (B-S-J-G China, Japan, Chinese Taipei, Macao-China), but some low performers also have similar class sizes (Dominican Republic, Brazil, Mexico or

Turkey). Conversely, countries with class sizes of less than twenty-five students include both top performers such as Estonia or Finland and poor performers such as Greece and Moldova (PISA, 2016d).

Thus, in its 2015 cycle PISA points out that large class sizes may result in positive trade-offs, such as freeing up time for teachers to prepare their lessons, or to engage in peer learning and professional development. They may have other benefits such as exposing many students to high-quality teachers. The OECD concludes that, since large classes lead to excellent performance in schools in East Asia, and across OECD countries students in large classes perform better, "governments should seriously consider the opportunity costs of reducing class size" (OECD, 2016d).

In contrast, the conclusions from PISA 2018 clearly recommend fewer students per class in order to improve outcomes, despite the data showing only weak and statistically insignificant correlations between both variables and the consistent finding that the top-performing systems in PISA have very large class sizes (OECD, 2019e). The mixed messages in the last PISA cycle are probably the result of the addition of many low-performing countries with large class sizes, such as the Philippines, Panama and Saudi Arabia (with around forty to forty-five students per class). This is a clear example of the limitations of correlational approaches, and of the contradictory policy recommendations that follow as the sample of countries changes over time due to the increased participation of low- or middle-income countries which tend to perform poorly.

More robust analyses using data from a variety of ILSAs to compare different countries have found that class size does not impact student performance (Cordero *et al.*, 2018; Hanushek and Woessmann, 2015; Woessmann, 2007; Woessmann and West, 2006), as have inter-regional studies within certain countries (Gomendio, 2021; Wert, 2019). As Nobel Prize winner Michael Kremer bluntly put it, adding "more-of-the-same inputs" (whether teachers, textbooks or other resources) has no impact on student performance (Kremer *et al.*, 2013). This conclusion comes from an experimental (RCT: randomised control trial) study in Kenya in which new teachers were hired on temporary contracts to reduce class size; despite a reduction in class size from eighty-two to forty-four students, those students who were randomly assigned to remain with

the same teacher did not show any improvements, while learning did improve among those students placed with new teachers, probably due to the latter's incentive to perform well and prove themselves because of their short-term contracts (Duflo *et al.*, 2015).

This experimental study clearly shows that in education systems where teachers have low skill levels, smaller classrooms will not solve the core problem. Conversely, in high-quality education systems, highly skilled teachers still achieve good student outcomes in large classrooms. This seems to be the case in East Asian countries, but whether this is only related to the fact that teachers are very effective, or to a more complex set of issues such as the high degree of discipline in the classroom, remains to be seen. These findings have led to intense academic disputes, but have had no impact on policy, since most countries have continued to reduce class size by hiring more teachers, despite the low benefits and high costs involved. We will discuss why in the next chapter.

Student Assessments

In most OECD education systems, there are national external standardised assessments for students at the end of lower- or upper-secondary level, or both. Central government is responsible for standardising both lower- and upper-secondary evaluations in most countries, although in decentralised systems this responsibility has been transferred to states/regions (e.g. Belgium, Germany and the US). The results of national assessments are used to obtain degrees and to determine students' entry to a higher grade or education level. In many education systems, the results of upper-secondary examinations are also used to grant access to tertiary institutions or degrees.

Over time PISA has changed its own conclusions on the impact of assessments. In 2006 and 2009, PISA correlations showed that external standardised evaluations had a large and positive impact on student performance, but in subsequent cycles PISA warned against the dangers of "high stakes" exams, i.e. student assessments with academic consequences (OECD, 2007, 2010b and 2013b). Apparently, the reason for this change in policy recommendations is that analyses carried out in later cycles focused on the uses of standardised tests and concluded that they had a negative impact on student performance if they were used

to adapt teaching to students´ needs, to identify aspects of instruction or the curriculum that could be improved, to make decisions about retaining or promoting students, or to make judgements about teachers´ effectiveness (OECD, 2016d, Fig. II.4.24). This leaves the question of what assessments should be used for.

In contrast, other analyses have consistently found that countries which have curriculum-based external exit exams tend to outperform countries without them (Bishop, 1997 and 2006; Hanushek and Woessmann, 2011 and 2015; Woessmann, 2018). In decentralised countries such as Canada and Germany, students perform better in regions with external exit exams, and strong accountability systems in states in the US improve student performance (Bishop, 1999; Graham and Husted, 1993; Hanushek and Woessmann, 2015; Jacob, 2005; Lüdemann, 2011; Piopiunik *et al.*, 2012). These improvements occur because standardised external evaluations are powerful signals for both students and teachers of the level of knowledge and skills expected at the end of each educational stage, allowing them to align their level of effort with these goals, and promoting practices which support students who are struggling to reach these targets. When these evaluations have direct consequences for students they also serve as powerful incentives for them to make the necessary effort to learn, hold teachers and principals accountable for the results, provide the evidence required to evaluate school and classroom practices, and allow policymakers to identify which schools or areas of the education system are performing well and which are falling behind and require improvements (Bishop, 2006; Fuchs and Woessmann, 2007). Student assessments also provide the necessary evidence about learning gains to evaluate whether policy decisions are having the expected positive outcomes across the system.

Critics argue that standardised tests may reinforce the advantages of schools with students from high socio-economic backgrounds, that they may demotivate low-performing students, or that teachers may narrow their teaching to the goals set by them (so-called 'teaching to the test'); these potential negative effects are presumed to be magnified when 'high stakes' (i.e. academic consequences) are linked to exit examinations (Clarke *et al.*, 2000 and 2003; Dee and Jacob, 2006; Dufaux, 2012; Hooge, Burns and Wilkoszewski, 2012; Jacob, 2005; Koretz, 2005; Koretz *et al.*, 1991; Ladd and Walsh, 2002; OECD, 2013b; Papay *et al.*, 2008). While

these criticisms highlight the need to design the tests adequately and to apply their results constructively to improve the quality of the education system, they do not provide any evidence that student performance is better when there are no evaluations. Furthermore, 'high stakes' make students and teachers care about exit exams and incentivise the whole education system to achieve the set standards. Diluting the consequences of exams by implementing evaluations with low standards or no academic impact would defeat the purpose of improving student performance. The important question here is: where is the evidence that a lack of evaluation leads to better outcomes or reduces the risk of students from low socio-economic backgrounds dropping out because they fear that they may not reach such ambitious targets?

There are few countries with no evaluations, but they provide a very firm answer. Spain and Greece are exceptions within the EU in that they have not implemented standardised evaluations. In these two cases, the reasons for not doing so are similar: fear of negative consequences outweighs the possibility of positive consequences. Furthermore, in these countries there have been dictatorships in the not-too-distant history which still cast a long shadow over the perceptions of many educational issues. In Spain, as explained in the chapter on ideology, the concept of external student evaluations is immediately associated with those in place during the Franco regime, which were specifically designed as bottlenecks to limit the number of students going to university. Thus, any form of evaluation is assumed to have the goal of segregating students and is perceived as a barrier designed to prevent students from low socio-economic backgrounds from going to university (Gomendio, 2021; Wert, 2019). In Greece, the fact that poor results in evaluations were in the distant past used to dismiss teachers taints any debate on the positive impact of evaluations with the fear of punitive outcomes for teachers (OECD, 2017d).

So, what are the real consequences of not having evaluations? They go far beyond the mediocre performance of students in these countries, because in fact they magnify the effects that they are intended to avoid. Traditionally, Spain and (to a lesser extent) Greece have suffered high rates of early school leaving in relation to other European countries, thus generating the worst form of inequity in any education system. These exceptions to the general rule show that, in the absence of clear and

uniform standards for all students, anyone struggling may go unnoticed, and the system lacks incentives to support them because the goals are non-existent. As a consequence, the gap between those students with difficult starting points and others widens as they grow older, until they lose all hope that the education system has anything to offer them and ultimately drop out.

Thus, a lack of standardised evaluations leads to the worst form of segregation for disadvantaged students, and to poor results overall. Some countries in Latin America also lack evaluations, and the arguments for not implementing them are very similar: i.e., as a measure that supposedly protects students from discrimination and teachers from unfair consequences. In South America, this also leads to poor performance among students, large inequities due to the strong impact of family socio-economic background and low quality of teaching (Bruns and Luque, 2015).

Since national evaluations define the same standards for all students, they also minimise the risk of geographical inequalities. Spain is an interesting counterexample, as an unfortunate exception within the EU in that it lacks national and regional standardised external evaluations. As a result, Spain has huge regional disparities in student outcomes (Gomendio, 2021; Wert, 2019). Thus, national standardised evaluations are also the main tool that central governments have to ensure equity, i.e. that students in different regions achieve similar standards (Gomendio, 2017).

The debate surrounding the correct uses of student assessments seems to have had a clear impact since, according to PISA 2018 (OECD, 2019d), there has been a decline in the frequency with which student assessments are used to compare school performance and to make decisions about promoting or retaining students. There has also been a very marked decline in their use in judging teachers´ effectiveness.

School Autonomy

Countries with good-quality education systems train highly skilled teachers and professional principals (who tend to enjoy high levels of school autonomy) to ensure that they have the flexibility to make the

most appropriate decisions for their student population in terms of curricula, pedagogical methods and allocation of resources.

Based on the fact that high-quality education systems tend to grant schools a high degree of autonomy, PISA makes a general recommendation that countries give more autonomy to schools in order to improve student performance (OECD, 2013a, 2016d and 2019d). Over the cycles, conclusions have focused on those aspects on which decision-making responsibilities should be transferred to schools, such as budgetary resources, curricular content or assessments. More importantly, since 2009 PISA has established a relationship between autonomy and accountability, making it clear that both should go hand-in-hand. This has been the trend in most OECD countries which have increased school autonomy, while at the same time implementing greater accountability in terms of outcomes (student performance). In this way, principals and teachers have increasingly been able to make decisions in their schools that they feel are best-suited to the specific needs of their students. At the same time, regional or central governments have developed more elaborate ways in which to evaluate whether their policy decisions do lead to better student performance and to implement corrective mechanisms if they do not. Other studies also support the conclusion that greater school autonomy will only lead to improved student performance when strong accountability measures are implemented, because this prevents opportunistic decision-making behaviour by agents who may pursue their own interests rather than seeking to improve student performance (Hanushek and Woessmann, 2015; Woessmann, 2007).

This is one of the policy recommendations that has been widely applauded and has become part of the policy package that is recommended to many low-performing countries (OECD, 2018a). This is unfortunate, since there is a large amount of evidence showing that school autonomy will only bring benefits when principals and teachers are prepared to use those responsibilities in an effective way (Hanushek, Link and Woessmann, 2013; Hanushek and Woessmann, 2015). This requires highly skilled teachers and principals who have been trained to take on leadership responsibilities.

In other words, the fact that schools have a large degree of autonomy in countries such as Finland, the Netherlands or Hong Kong does not

mean that granting more autonomy to schools in Greece, Turkey or Mexico would improve their results. In fact, it would mean the opposite. Other studies have shown that school autonomy improves student performance in countries with high-quality education systems but has a negative impact upon student performance in developing countries (Hanushek, Link and Woessmann, 2013). This is a classic example of the mistake of extrapolating those practices which work in mature, high-quality education systems and importing them to low-performing systems before they are ready to take the required steps. Improving education systems requires a carefully orchestrated sequence of steps. School autonomy is one of the last steps in that sequence, because it is first necessary to deal with the quality of teachers and to build up principals' capacity to be true leaders.

This kind of policy advice on school autonomy highlights the errors that are often made when features that are common among top-performing systems are transformed into recommendations to low-performing systems without careful attention to the context. School autonomy in itself does not improve student outcomes; necessary preconditions are that teachers have already achieved a high level of skills and principals have been trained as leaders. A closely linked example is the frequent recommendation that teachers should be allowed to innovate. There is evidence for this argument in Finland, where teacher innovation is regarded as one of the key features of the system's success (Gomendio, 2017). But in countries where teachers have not achieved a similar level of skills, they need guidance much more than the freedom to innovate (Barber and Mourshed, 2007). This is why the choice of quality textbooks in low- and middle-income countries is key. Textbooks conversely play a much less significant role in countries where teachers are ready to be creative, innovate and use a wide array of educational resources.

It is surprising that policy recommendations regarding the autonomy of schools and principals tend to avoid advising that principals should be able to choose the teachers that join their schools, and dismiss those that are underperforming. In fact, many education systems have developed mechanisms that allow more senior teachers to choose the school where they work, but do not grant principals the power to select teachers or dismiss low-performing teachers. A few pilot experiments

have shown the positive impact of empowering principals to make decisions on the teachers in their schools. In Chicago, principals were allowed to dismiss teachers who they regarded as unsatisfactory while they were on probation; research showed that dismissed teachers did have higher rates of absenteeism, low performance rates and a negative impact on students, thus corroborating principals' capacity to correctly identify low-performing teachers (Jacob, 2012; Jacob and Lefgren, 2008; Jacob and Levitt, 2003). But the findings went further, since absenteeism also decreased among tenured teachers in these schools. It has also been shown that dismissing the lowest 10% of teachers has a substantial impact on student performance (Hanushek, 2009). In any other sector, the importance that leaders attach to their autonomy to build their teams is well-established and is not up for debate. We will discuss possible reasons for the generalised refusal to grant this power to principals in the next chapter.

School Choice: Public vs Private Schools

Most education systems allow the co-existence of different types of schools, which fall into three categories: (*i*) public schools are funded and managed by government; (*ii*) the so-called government-dependent private schools by PISA (also known as charter schools in some countries) are funded by government and managed by NGOs or religious organisations; and (*iii*) private schools are for-profit and are privately owned and run.

When only two broad categories are considered, government-dependent private schools can be considered either public or private. Since the mere existence of these schools is a controversial issue in many countries, it is revealing that supporters tend to label them as 'public' and detractors as 'private'. As a matter of fact, in many analyses PISA considers both charter (government-funded) and private (for-profit) schools as a single category, and compares this too broad and basically heterogeneous category against public schools. The consequences of this are two-fold. First, the analyses are not granular enough to consider in full the three categories and to draw clear conclusions about charter schools, which are the focus of much controversy in some countries. But, second, PISA takes a side about the nature of these schools, supporting

the critical stance that some radical advocates of public education express when arguing that public funds should only be allocated to schools run by governments.

In our view, charter schools are funded by governments, who are responsible for ensuring that all citizens can exercise their right to education by providing adequate school places. Thus, they are part of the network of public schools that must ensure that all children have access to quality education and are bound to follow most of the basic rules defined by governments, just like schools managed by governments. Our position in relation to the definition does not in any way convey a bias or prejudice about government-dependent private schools, but we wish to flag it, because it affects how analyses are carried out and conclusions are drawn.

There is great variation among countries in the proportion of students that attend different types of schools and in the extent to which parents can choose the school that they think is most appropriate for their children. Among OECD countries, about 82% of fifteen-year-old students attend public schools, around 14% attend government-dependent private schools, and just over 4% attend private schools (OECD, 2016b, Table II.4.7.). However, these averages hide a large degree of variation between OECD countries. In around half of the countries more than 90% of students attend public schools, and in most countries the proportion of students attending private schools is rather low (ranging between 0% and 10% at most), although there are a few exceptions, such as Japan (28%).

The proportion of students attending government-dependent private schools is over 50% in countries such as Belgium, Chile, Ireland, the Netherlands or the UK. In a few education systems, the vast majority of students attend government-dependent private schools, as is the case in Hong Kong (93.3%) and Macao (China) (83.2%). Among government-dependent private schools, there is also variation in the types of organisations that run schools: across OECD countries 39% of students enrolled in these schools attend schools run by a religious organisation, 53% are in schools run by another non-profit organisation, and 8% attend schools run by a for-profit organisation (OECD, 2016b, Table II.4.7).

In countries such as Belgium and the Netherlands, government-dependent private schools have a long history because they have traditionally been regarded as an effective way for the public education system to offer parents a broad range of choices including, but not limited to, different religious faiths (Fontaine and Urzúa, 2018; Nusche *et al.*, 2015; OECD, 2017e). In this way, historical confrontations about which religion, if any, should be taught at school were solved by allowing different types of schools to co-exist and enabling parents to exert meaningful choices. It has been argued that the principles of freedom of religion paved the way for school choice to become one of the pillars of these education systems (Patrinos, 2011). In other countries, the Church had a historical role in creating the first schools, which were eventually integrated into the public system as government-dependent private schools. This is the case in Ireland and Hong Kong, which have a large proportion of government-dependent private schools, most of which are run by the Catholic Church (Renehan and Williams, 2015; Tan, 1997).

More recently (from the 1990s onwards), a group of countries has introduced reforms aimed at enhancing school choice in order to make the education system more sensitive to the increasingly varied needs of societies which have become more diverse and plural, as well as to enhance quality and stimulate innovation. These countries include New Zealand, Spain, the United Kingdom and the United States. In the United Kingdom, the changes have gone beyond developing a model of government-funded private schools, since Tony Blair's Labour government's introduction of so-called 'academies', which entailed a major change in governance: the responsibility and the funding shifted from local authorities to central government, and new accountability mechanisms were put in place (Adonis, 2012; Wiborg, 2017a). These major changes were originally introduced to tackle the large number of low-performing public schools in the UK. The model was further expanded by consecutive governments and it has grown so rapidly that at present nearly 70% of publicly funded secondary schools are academies.

We will not address in this section the ongoing debate about whether parents should have the right to ensure that they can send their children to schools which are aligned with their views on pedagogical approaches, discipline, values or religious faith, or whether this

choice should be limited because it may increase social and cultural segregation (Elacqua, 2012; Levin, Cornelisz and Hanisch-Cerda, 2013; OECD, 2017e; Renzulli and Evans, 2005; Saporito, 2003). Instead, we will focus on whether school choice does improve student performance by stimulating competition and, in doing so, enhancing efficiency and innovation (Chapman and Salokangas, 2012; Jiménez and Paqueo, 1996). Since reforms which have expanded school choice have also implemented new accountability mechanisms which focus on student performance, rather than the traditional focus on inputs and processes, it is important to consider both simultaneously. We will also look at the evidence for the claim that school choice increases inequality because middle-class families tend to exert their choice and send their children to government-funded, privately managed schools, which detract resources from public schools where students from disadvantaged backgrounds tend to remain.

When addressing this issue, PISA tends to group government-funded private schools with private schools and compare this broad group against public schools. Thus, the analyses are not granular enough to compare the three categories and to draw clear conclusions about government-dependent private schools, which are the focus of much controversy. Data from PISA have consistently shown that student performance is better in private schools (OECD, 2010b, 2013a and 2016b). However, after accounting for socio-economic status, in twenty-two education systems students in public schools score higher than students in private schools, while in nine systems they score lower than students in private schools (OECD, 2016d). This reflects differences in the extent to which students are selected by socio-economic status from country to country.

It is worth noting that the percentage of students enrolled in government-dependent private schools is positively correlated with average scores of student performance at the national level, but there is no association with equity (OECD, 2016d). The positive impact of government-dependent private schools on student achievement is correlated with the greater levels of autonomy granted to these schools (OECD, 2016d), combined with better accountability mechanisms based on outputs, i.e. student performance (Nusche *et al.*, 2015). Further analyses have also established that a major causal factor linking school

choice and the existence of government-dependent private schools to improved student outcomes is enhanced competition between schools to improve student performance so as to become more attractive to parents (Hanushek and Woessmann, 2015, West and Woessmann 2010; Woessmann, 2007b; Woessmann *et al.*, 2009).

A recurrent theme in the policy debate concerning government-dependent private schools is whether they only achieve better student outcomes because they select students according to socio-economic background or level of performance. Several studies show that this is not necessarily the case. In the case of academies in the UK, recent studies have shown that the conversion of underperforming schools to this new model has led to improvements in the performance of students who were already attending the same school before its transition to the new model, and that the degree of improvement is greater among schools that gained larger degrees of autonomy from the conversion (Eyles and Machin, 2019). This work clearly shows that a dramatic change in the governance and accountability mechanisms of these schools improved the performance of those students who attended them prior to their conversion to academies, thus eliminating the possibility that improved student performance was the result of academies selecting high-performing students. Other studies carried out in charter schools on school admission policies based on 'lotteries' have shown that the performance of pupils who were 'lotteried' into charter schools improved, while the performance of those who were not accepted did not (Abdulkadiroglu *et al.*, 2011; Angrist *et al.*, 2010, 2013 and 2016; Dobbie and Fryer, 2011, 2013 and 2014; Hoxby, Murarka and Kang, 2009).

The data also show that government-dependent private schools tend to be much more cost-effective than public schools, since the former tend to provide education at a lower cost per student than the latter (Hanushek and Woessmann, 2015; Howell and Peterson, 2002). Although the reasons for this are to some extent country-specific, in most cases it is the result of a combination of factors including teachers in government-dependent private schools investing more time in teaching, these schools having larger class sizes, and their principals having more control over the hiring of teachers.

However, when government-dependent private schools receive too little funding from government, they may not be able to afford to provide free education and may instead charge tuition fees or add-on fees for extra-curricular activities. Since this undermines the principle of free school choice, it is important that enough funding is provided by governments and that these schools do not charge additional fees or follow selective admissions policies. Regulatory mechanisms should be implemented to prevent government-dependent private schools from targeting families who can afford to pay for their children´s education and/or the best-performing students, since both would lead to a wider inequality gap (OECD, 2017e).

A rather recent phenomenon which seems to be growing fast in some developing countries is the emergence of low-cost private schools, which tend to produce better student outcomes at a much lower cost than public schools (Amjad and MacLeod, 2014; Barber, 2013; Van de Berg *et al.*, 2017).

Student Socio-economic Background

International comparisons have consistently revealed that no education system has been able to prevent the impact of socio-economic background on student performance. The reasons for the strong influence that family background exerts are complex and include many factors, such as the degree of stimulation that children receive from their parents and home environment before they enter school, how much children can learn from their parents own skill levels, the expectations that parents from different backgrounds may have for how their children should perform at school, the quality of the support that parents can provide for their children's learning needs, and the value that they place on education.

According to PISA, the impact of socio-economic background is universal since students from privileged backgrounds perform better than underprivileged students in all countries. However, it is also true that good-quality education systems raise the performance of all students, while even privileged students fail to achieve high levels of performance among low-quality education systems. As a result, PISA comparative data consistently show that students from low socio-economic backgrounds in good-quality education systems outperform

privileged students in low-quality education systems (OECD, 2013a and 2016d). These data suggest that high-quality education systems have more power to improve the performance of poor students than low-quality systems' ability to improve rich students' performance. But there is more to the story than this.

The fact that education quality improves the outcomes of all students challenges the widespread assumption that quality is achieved mostly by improving the performance of privileged students, at the expense of students from low socio-economic backgrounds. On the other hand, privileged students in poor-quality education systems cannot completely escape the overall poor levels of performance. To put it bluntly, money cannot overcome the limitations of underperforming education systems, probably because there are features of the system (such as low-quality teachers and curricula, a lack of accountability mechanisms or assessments with low standards) which are pervasive. It also calls into question the argument that equity can only be achieved by lowering standards to ensure that underprivileged students do not fall behind, fail evaluations, drop out, or fail to obtain degrees. In fact, these findings show that the opposite is true, since students from low socio-economic backgrounds perform very well in good-quality education systems, probably because there are compensatory measures that ensure that they get the support that they need in order to achieve high standards.

But the fact remains that, within each country, socio-economic background always has a major impact on student performance. According to PISA 2015, it is the single greatest influencing factor in student performance, when compared to many others (OECD, 2016d). This has been interpreted by some as a depressing sign that none of the education policies that have been implemented so far have allowed disadvantaged students to overcome their 'fate' as low performers. But there is a brighter side to this rather gloomy interpretation. The extent of the impact of family background varies greatly from country to country. There is an intense debate about whether this is because some societies are more equitable than others (and this is also reflected in the outcomes of education systems), or whether some education systems are more effective at diminishing the influence of social and economic inequities. Since we know that some countries, such as Nordic countries, are

more equitable than others, such as Latin American countries, it seems reasonable to ask to what extent can education systems be expected to compensate for large social and economic disparities.

Other studies have adopted a more sophisticated approach which has shed light on this difficult issue. What is missing in PISA's comparative data is that, when comparing the performance of students from different percentiles of the socio-economic range, the relative wealth of the country is not considered. It seems obvious that the poorest students in Finland are not as poor as those in Colombia. Furthermore, PISA also fails to take into account the fact that some countries still have not reached universal access to education, and in others a substantial proportion of students drop out of school before they reach fifteen years of age. One clear example of this oversight is Vietnam, a country which was hailed as an outstanding example of equity in PISA 2015, when the fact is that over 50% of the fifteen-year-old population does not attend school and is therefore not even assessed by this survey (OECD 2016 b, Fig. I.6.2). In fact, PISA claims that "the world is no longer divided between rich and well-educated nations and poor and badly educated ones: the 10% most disadvantaged students in Vietnam compare favourably to the average student in the OECD area" (OECD, 2016b, p. 4). It is misleading to make these statements when Vietnam has the lowest proportion of fifteen-year-olds enrolled in school of all PISA participating countries, and it seems reasonable to assume that the other 50% not enrolled in schools come from very poor family backgrounds. This is not an isolated example. In most Latin American countries, a substantial proportion of students are no longer in school by the age of fifteen, so any analyses concerning the impact of socio-economic background seriously underestimate its real impact (OECD, 2016b).

Using a different approach that takes into account both national income and household income, it becomes clear that both matter. The performance of primary students (using data from TIMSS, PIRLS, and two other regional multi-country assessments, Latin American LLECE and African PASEC) is strongly correlated with household income in real, purchasing-power-parity dollars across countries, but students with the same level of household resources have different educational outcomes depending on the wealth of their country of residence (Patel and Sandefur, 2020). The reason for this is that student outcomes are

also strongly linked to GDP per capita, except in oil-rich countries where the wealth of the country does not translate into improved student performance (an educational version of the 'curse of the commodities'). As a result, poor students in rich countries perform better than rich students in poor countries. So countries do indeed seem to be divided between rich and well-educated countries and poor and badly educated countries, contrary to the OECD's claim.

This study also provides new insights on the impact of economic inequality within countries. Among countries with high levels of inequity (as measured by the Gini coefficient), the impact of household income on student performance is much greater than among more egalitarian societies. Thus, household income has a greater impact on student performance in countries like Colombia, Brazil or Guatemala than it does in economically equitable societies such as Finland or Norway. It is worth pointing out that economic inequality often goes hand-in-hand with the range of differences in skills of the adult population (OECD, 2016e). While most parents in Finland have high skill levels, only a small proportion of parents in Latin American countries achieve similar skill levels. Thus, the relationship between the degree of inequality and the extent of the impact of parental income is most likely not just about how much parents can invest in education, but also about how much children can learn from their parents and their home environment. This finding has important implications. It suggests that education systems cannot overcome the impact of social and economic inequalities when these are profound.

It also cautions against the risk of establishing causal links between specific education policies and equitable outcomes in egalitarian societies, as well as the risk of assuming that transferring those policies to countries with high levels of inequity will successfully reduce inequality in student outcomes. It seems more likely that social and economic equity permeates education systems which, therefore, do not require major interventions against inequity. The prime example is the widespread assumption that because Finland has equitable education outcomes its policies should be extrapolated to countries where inequality is rampant. It seems more likely that Finland can afford those policies because social and economic inequality is not a major issue. Countries with high levels of economic inequality may require different policies from those with

low levels of economic inequality, because they are each addressing completely different challenges. In a context of strong economic and social disparities, students from low socio-economic backgrounds may need additional support, more personalised attention and more flexible pathways, which are not required in more egalitarian societies.

The fact remains that most education systems aim to achieve quality and equity without trade-offs between the two. Equity has two main dimensions: *fairness* implies that personal circumstances (such as gender, socio-economic or migrant status) should not have any major impact on student outcomes; *inclusion* implies ensuring a basic minimum standard of education for all (Field *et al.*, 2007). This two-dimensional definition is important because it asserts the need to prevent students from falling below a certain threshold, and it also avoids claiming that equity requires similar outcomes. Instead, it emphasises the need to minimise the impact on student outcomes of factors which are known to hinder learning.

We will now review the evidence concerning which factors influence equity in outcomes, and which policies seem more effective at enhancing equity.

Dealing with Student Diversity: Is Diversification a Form of Segregation?

A major challenge for education systems is how to deal effectively with the degree of student heterogeneity found within a single grade or classroom, or how to ensure that struggling students are not left behind while allowing those top performers to advance more rapidly. A number of policies have been developed with this aim.

First, many countries have implemented some form of 'ability grouping', which sorts students according to their level of academic performance in different groups or classes at primary and/or lower-secondary level. The term 'ability grouping' includes a wide array of practices. In its most extreme form, students may be sorted into different classes for all subjects (a practice referred to as 'tracking' in English-speaking countries). Softer versions of ability grouping involve students being divided into different groups within the same class for certain subjects.

Second, all countries have differentiated trajectories which students can choose in upper-secondary education, but some countries start much earlier. Traditionally, the major divide has been between academic programmes and vocational education and training (or apprenticeships), but students are also allowed to choose different paths within the academic track. A few countries have also developed several trajectories which represent different combinations of academic and VET-oriented content. The separation of students into academic and VET trajectories is labelled in PISA publications as 'tracking', generating some confusion with extreme forms of ability grouping.

Third, when these practices are not implemented or do not prove efficient enough in reducing student heterogeneity in performance, some students may lag so far behind that they make little progress in one grade; in these extreme cases most education systems resort to grade repetition, which means that these students remain in the same grade for another year in order to allow them to catch up and increase their chances of continuing to make progress in the education system.

There is an intense controversy both among policymakers and academics about the pros and cons of practices which aim to reduce student heterogeneity in academic performance. The clear advantage is that teachers will find it easier to make progress if they teach a group of students with a similar level of performance who can thus follow a similar pace and have similar needs. When teachers are faced with a heterogeneous group of students their efficiency may be compromised, since they must make choices about whether to focus on the low-performers, the top-performers or a bit of both, thus failing to meet the very diverse needs of their students. But opponents claim that any practices which separate students according to performance will harm low-performing students who will not be allowed to learn from their high-achieving peers, thus exacerbating inequality, and in most cases will lead to discrimination based on socio-economic background or immigrant status. From this viewpoint, these practices are seen as non-inclusive and referred to in a derogatory way as 'segregation'. The recommendations from PISA are consistent with this discourse and therefore discourage countries from any practice which aims to reduce student heterogeneity in performance, because it is assumed that it will lead to segregation and will increase inequity. Thus, PISA does not

recommend ability grouping, early tracking or grade repetition. Let us look at the evidence.

Is Student Heterogeneity within Classrooms Really a Major Problem in Most Education Systems?

A detailed analysis of the education system in Chile may provide a clue (Fontaine and Urzúa, 2018). According to the reports by principals to the PISA questionnaire, in Chile 70% of students are in classrooms where the main barrier to learning is the heterogeneity in student performance. Further in-depth interviews with teachers in Chile reveal that they feel that this degree of diversity in levels of skills and knowledge within schools and classrooms is the main challenge that they face in their struggle to make progress in learning. It is well-known that social inequity is a major issue in Chile, as well as in many Latin American countries. Therefore, the broader issue is whether the education system can compensate for the inequity that is so prevalent in some societies, and how. The authors conclude that, in the context of such a large degree of social and economic disparities, treating all students equally will generate unequal results which do not reflect merit (Fontaine and Urzúa, 2018). Thus, it is possible that education systems can only deal with such levels of heterogeneity by implementing mechanisms which organise students into groups that reduce performance disparities, so that teachers can be more effective in ensuring learning. In other words, a large degree of heterogeneity in student performance may require the implementation of measures to allow teachers to manage it. But is this challenge unique to countries with large social and economic inequality?

If we dig deeper in the principal reports for PISA 2012, which is a key element of the study in Chile, the results are truly shocking. In 86% of countries (fifty-five out of sixty-four countries) more principals identify "teachers having to teach students of heterogeneous ability levels within the same class" as a bigger obstacle to learning than any of the other ten potential barriers, which include "teachers not being well prepared for classes", "teachers having to teach students of diverse ethnic backgrounds (i.e. language, culture) within the same class" and "teachers' low expectations of students" (OECD, 2013a). Thus, student

heterogeneity in terms of ability is regarded as the main barrier to learning in most countries.

Furthermore, in those few countries where principals do not believe that student heterogeneity is the main issue, it is still among the top three barriers. For example, in Australia and Italy, around 35% of principals believe that student heterogeneity is an obstacle to learning which is similar to or slightly higher than the proportion who believe that "staff resisting change" is an obstacle too. Similarly, in the United Kingdom, 14% of principals reported that student heterogeneity is an obstacle to learning, while only a slightly higher proportion believe that "teachers not meeting individual students' needs", "teacher absenteeism" and "staff resisting change" are barriers to learning. In countries where teacher absenteeism is prevalent (e.g. Uruguay and Tunisia), it is cited by a similar proportion of principals as a barrier to learning than student heterogeneity, which gives an idea of the extent to which principals regard the latter as a major problem. In contrast, in many countries where most principals identify student heterogeneity as an obstacle to learning, the distance from other potential barriers tends to be much larger. This is the case in Colombia, Chile, Portugal and Spain, where 70–80% of principals identify student heterogeneity as a barrier to learning, while other potential obstacles are only considered relevant by 25–40% of principals.

These data show that in most countries, most principals believe that student heterogeneity in terms of ability is a major obstacle to learning. This is even the case among top performers such as Finland, Singapore or Hong Kong. The issue of how to allow a heterogeneous classroom to make progress without leaving struggling students behind, or preventing those who excel from continuing to advance, is a universal and major challenge. While providing individualised teaching to each student seems the optimal strategy, this is rarely possible and requires a highly skilled teaching force, plus a combination of technology-enabled resources and technology-savvy teachers, to create a bespoke 'personal learning environment' for every learner. This task is easier said than done. Hence, in most cases teachers may be more effective when student heterogeneity is reduced by separating students into ability groups, different trajectories, or in extreme cases resorting to grade repetition.

Grade Repetition

The approach that PISA adopted when defining its target population is different from other ILSAs in that it assesses fifteen-year-olds irrespective of the actual grade in which they are studying. Thus, PISA evaluates the performance of fifteen-year-old students in the grade corresponding to their age (modal grade), as well as those in lower grades because at some point they have remained in the same grade level for an additional year (either once or several times) due to low academic performance. Only fifteen-year-old students who remain in primary education are excluded from PISA's consideration.

Among OECD countries an average of 11% of students participating in PISA reported that they had repeated a grade at least once, but the variation between countries is very large. In some countries, the rate of grade repetition is below 5% (mainly Nordic countries and countries in East Asia, such as Denmark, Sweden, Finland, Iceland, Estonia, Singapore, Japan and Korea), while in others over 20% of students report having repeated a grade at least once (Spain, Portugal and all Latin American countries participating in PISA 2018). In Colombia over 40% of students have repeated a grade at least once, and in Morocco this figure stands at over 50% (OECD, 2019c).

The OECD refers to grade repetition as "vertical stratification". Students in disadvantaged schools are four times more likely to repeat a grade at least once (20%) than students in advantaged schools (5%), and there is an ongoing debate about whether this is due to the impact of socio-economic background on student performance, or to discrimination against these students because of the low expectations that teachers may have (OECD, 2020c). Unsurprisingly, students who have repeated a grade at least once show substantially lower levels of performance in PISA, corroborating international metrics on the poor academic performance of students who resit a grade. Unsurprisingly too, countries where grade repetition is more prevalent score lower in PISA, since a larger proportion of the fifteen-year-olds in the PISA sample have fallen behind, are in lower grades and show lower levels of performance. Surprisingly, the recommendation that PISA makes based on these findings is that, because students who have repeated a grade

perform at a lower level, countries should avoid grade repetition. This misses the point entirely.

First, the almost exclusive use of correlations in PISA is itself problematic because it leads to a well-known statistical problem: 'reverse causality'. When two variables are positively or negatively associated it is not possible to conclude which is the causal factor, or if both are caused by a third factor which has not been included in the analysis. If a correlation is wrongly used to draw conclusions about causality, a common mistake is to identify one of the two variables as the causal factor, when it is actually the other (i.e. reverse causality). This seems to be one of those cases. Grade repetition and low student performance are associated, not because grade repetition lowers student performance, but rather because low student performance leads to grade repetition.

Second, the goal of grade repetition is to allow students who are lagging so far behind that they cannot follow what is being taught to them to catch up and to have a second chance to learn what they could not manage the first time. But they are expected to catch up with their peers during this second go at the same grade, and not with the former peers who have moved on to the next grade. Thus, the expectation that students who remain in lower grades should perform similarly to those who move on to the modal grade is misplaced, since fifteen-year-old students who have repeated a grade have not been exposed to the same curricular content and teachers as students in the modal grade. Thus, while PISA has much to say about the extent to which education systems equip fifteen-year-olds in each country with the required knowledge and skills (irrespective of their grades), and this seems to us a valuable contribution, it cannot draw the conclusion that grade repetition harms performance by comparing students in different grades with the same metrics. The result is obvious, and the expectation is unfounded.

At the individual level, the relevant question is whether grade repetition does allow students who have fallen behind to catch up, and thus whether it improves their chances of progressing in their education. The counterfactual, i.e. whether students lagging behind would have made greater progress if they had been allowed to move onto the next grade, cannot be tested. At the systemic level it is important to understand why grade repetition is more prevalent in some education systems, what the alternatives are, and what the costs and benefits

are. Since grade repetition is the consequence of differences in the performance levels of students in the same grade which the education system considers insurmountable, it is important to understand whether such large differences are the consequence of very different starting points in compulsory education, due to differences in socio-economic background, immigrant status, or other factors. Alternatively, large differences between students could be due to the education system's failure to compensate early on for different starting points and to provide the support that struggling students need before it is too late.

According to the information provided by the reports from principals and students to PISA 2012 questionnaires, in many of the countries in which grade repetition is rare, a relatively low proportion of principals think that the main impediment to learning is "teachers having to teach students of heterogeneous ability levels within the same class" (OECD, 2013a and 2014c). This is the case among European countries with low rates of grade repetition such as Denmark, Sweden, Iceland or Estonia, where between 39% and 56% of principals identify student heterogeneity as the main obstacle to learning. However, this is not the case in Japan, where grade repetition is forbidden, despite 72% of principals believing that student heterogeneity is the main barrier to learning. In contrast, among countries with high rates of grade repetition, a larger proportion of principals tend to identify student heterogeneity within the classroom as an obstacle to learning (Spain: 66%, Portugal 68%, Chile 71%, Uruguay 75%, Colombia 80%).

These findings suggest that, while grade repetition is a last-resort mechanism, some countries make more frequent use of it, either because the student population is more heterogeneous than in other countries when they start compulsory education, and/or because alternative mechanisms implemented to deal with student heterogeneity (if any) have not been effective by the time students reach the age of fifteen. Still, grade repetition seems to be an inefficient strategy because students who repeat grades are more likely to drop out of school and regions with higher rates of grade repetition also suffer higher rates of early school leaving and youth unemployment (Gomendio, 2021; Wert, 2019). It has also been suggested that students who repeat a grade develop more negative attitudes towards school (Ikeda and García, 2014; Rumberger and Lim, 2008; Thompson and Cunnigham, 2000; West, 2012), although

other studies have found that student retention impacts positively on achievement (Allen *et al.*, 2009). The high costs of grade repetition for the education system compounds its limited efficacy. The total cost of grade repetition can represent 10% or more of some countries' annual national expenditure on primary and secondary education (OECD, 2016b; Wert, 2019).

Grade repetition therefore seems a radical and costly measure, which is not effective because it is quite a rough and crude practice which intends to address low performance after students have fallen dramatically behind by making students go through a whole year of the same curricular content and teaching practices that did not work the first time. However, merely recommending that grade repetition should not occur is not helpful, because it does not address the issue of how to avoid such large differences between students and how to support students who are lagging behind early enough. In other words, education systems need to know what the alternatives are, not just to be told what they should not do.

Spain is a good example of how designing an education system with the theoretical aim of achieving equity has led to one of the least equitable outcomes. It is also very revealing that PISA is blind to the clear signs of inequity in the Spanish system and has reinforced the myth that Spain has sacrificed excellence in the pursuit of equity.

In a nutshell, for decades the Spanish education system has banned all practices that were suspicious of segregating students, such as ability grouping or early tracking. It has also refused to implement external standardised national (or regional) assessments at the end of lower- and upper-secondary education because they are widely regarded as unfairly discriminating against students from poor socio-economic backgrounds. Thus, the education system is not only unable to deal with the diversity of students entering schools, but actually allows differences between students to increase as they age, precisely because it does not allow any differential treatment of students. The lack of assessments in primary education means that students from difficult starting points are not identified early enough and therefore do not get the additional support they need. The lack of standardised evaluations at the end of lower- and upper-secondary education means that there are no clear goals that students need to reach, leaving both students and teachers

without any incentives. As a result, students who are struggling do not have ways to catch up, and those who could become top performers are not given the opportunity to excel.

According to PISA, the system is flat, with a small proportion of top-performing students and the same proportion of low performers as the OECD average, which leads to overall mediocre results. This flatness may be wrongly interpreted as reflective of equitable outcomes, since no factor—including socio-economic background—can be identified as having a major impact when levels of performance are uniformly poor. But what PISA fails to detect is that struggling students gradually fall further and further behind until they eventually start repeating grades and ultimately drop out. As a result, the rate of grade repetition in Spain at age fifteen was around 40% from 2000 until 2011, and the rate of early school leaving was 26% in 2011. In conclusion, although it may seem counterintuitive, not implementing practices that allow differential treatment of students according to their academic performance for fear of generating inequality may lead to the worst type of inequality: students being excluded from the education system because they have been lagging behind for years and have lost any motivation or hope that it has something to offer them. These students leave with such low levels of knowledge and skills that they face high levels of unemployment during their lifetimes and are very reluctant to engage in any form of adult learning (Gomendio, 2021; Wert, 2019).

Ability Grouping

Separating students into groups according to their ability for some subjects is the least drastic strategy and, according to PISA, does lead to better student performance without having any negative impact on equity (OECD, 2020c). Among OECD countries, grouping students into different classes is quite common, since 46% of students attend schools whose principal reported this practice, with 38% of students being grouped for some subjects and only 8% for all subjects (OECD, 2016). Ability grouping within classes is even more common: 55% of students attend classes where there is ability grouping, in most cases only for some subjects (50% students) and in a few cases for all subjects (5%). Thus, the benefits of sorting students into more homogeneous groups

with varying levels of difficulty seems to improve student performance, while avoiding the potential costs linked to low-performing students being unable to learn from their higher-achieving peers (Collins and Gan, 2013; Garelick, 2013; Zimmer, 2003).

An experimental study in Kenya sheds light on the controversy around the benefits and costs of separating students into different groups according to their academic performance (Duflo, Dupas and Kremer, 2011). The study was carried out in primary schools that hired an additional teacher and were therefore able to split classes into two (average class size was eighty-three before hiring a new teacher). In half of these schools, students were split according to their academic performance (so-called 'tracking' schools), while in the other half students were randomly assigned to each class ('non-tracking' schools).

The results showed that all students benefited from 'tracking' because teachers were able to make more progress when dealing with a more homogeneous class, while no improvements were observed when class size was reduced but students were randomly assigned to each class. The positive impact on reading and numeracy was clear both for top- and low-performing students. Thus, the benefits for low-performing students clearly offset any potential negative effects of being placed with similarly performing peers. Furthermore, these gains persisted after the programme ended, suggesting that students acquired core skills that facilitated learning later on. Interestingly, the students who benefited the most were low-performing students who were assigned to contract teachers, suggesting that more homogeneous ability groupings and teachers with the right incentives achieve larger gains for low-performing students.

This study was conducted in the context of high levels of student heterogeneity, since students in Kenya differ in age, school readiness and support at home. But the study is unique because its experimental approach allows the establishment of causal relationships that contradict established dogmas: class size reduction *per se* did not have a significant impact on student performance, but assigning students to different classes according to their level of academic performance did. It is possible that in countries were student heterogeneity is smaller, other less drastic strategies—such as online resources or ability grouping

within classes—may be enough to help teachers deal with student diversity.

The important conclusion is that teachers can make greater progress in learning when diversity in student performance is reduced, and this can be accomplished in different ways. As long as these practices reduce heterogeneity by focusing on levels of performance (rather than socio-economic background or immigrant status), they will not increase inequality, because struggling students will benefit the most. The most effective strategy will depend on the level of student heterogeneity in schools which, in turn, depends on external factors such as the degree of social and economic inequality, differences in the levels of educational attainment and skills among parents, proportion of immigrants, and proportion of students enrolled in pre-school education. In conclusion, ability grouping cannot be universally recommended, because it is strongly context-dependent.

Vocational Education and Training (VET) and Apprenticeships

Most education systems have developed "academic or general" and "vocational education and training or apprenticeship" (VET) programmes at school, with the exception of most English-speaking countries which do not offer differentiated VET programmes in school. The main difference is that academic programmes focus on theoretical knowledge, while VET programmes focus on applied skills which are more closely linked to the needs of the labour market. Thus, while academic programmes have traditionally been the main pathway for those who wish to access university, VET programmes have been designed as a more direct route through which to enter the labour market or to continue into tertiary VET.

Education systems in most OECD countries are 'comprehensive', which means that all students follow the same programme until the end of lower-secondary education. Thus, students choose between the academic and VET programmes at the age of sixteen when they move into upper-secondary education. In a few countries, this choice is made much earlier: at ten years old (Austria and Germany), twelve years old (e.g. Belgium, Netherlands, Switzerland, Singapore) or thirteen years

old (e.g. Luxembourg). Most of the countries where the choice is made earlier offer several programmes that cover a range of combinations of theoretical and applied knowledge, while most of the countries where VET is only available in upper-secondary education have two clearly distinguished paths (VET and academic).

Over many cycles PISA has consistently claimed that the performance of fifteen-year-olds in VET programmes is lower than that of students on academic tracks (OECD, 2013a, 2016d and 2020c). This has led to the conclusion that following a VET programme before the end of compulsory education has a negative impact on student performance and increases inequality, because students from low socio-economic backgrounds are more likely to choose or be assigned to VET. Based on these findings, one of PISA´s strongest recommendations is that countries should delay the start of VET programmes as the lesser of two evils (OECD, 2013a, 2016d, 2020c). Since this recommendation and its wide acceptance has had a major impact on the education policy debate, it deserves detailed scrutiny here.

The first issue is that, as mentioned before, in most OECD countries the choice between academic and VET programmes does not take place until students enter upper-secondary education, in most cases at the age of sixteen. Since PISA evaluates fifteen-year-olds, in most countries it cannot assess students in VET programmes. To circumvent this problem, PISA includes as VET students those enrolled in what it calls "pre-vocational" programmes. This is grossly misleading since in most countries these programmes are specifically designed for very low-performing students who are deemed unlikely to obtain a lower-secondary degree. Thus, these programmes are normally designed for a tiny minority of students who need an alternative path to obtain a different educational degree. Despite this questionable tactic, the sample sizes for most countries remain very low: in almost half (46%) of the thirty-five OECD countries considered, the percentage of students in pre-vocational or VET programmes is less than 1%, with many countries having no students at all in this category (OECD, 2016d).

It seems questionable that PISA would draw any solid conclusions from such small sample sizes. But in fact, PISA argues that the negative impact of VET (or pre-vocational programmes) on student performance is greatest among some of those countries with the lowest proportion of

fifteen-year-olds enrolled in such programmes. Ireland (0.8%), Spain (0.9%) and Georgia (1.7%) are among the five countries for which PISA claims that the negative impact on performance is largest (OECD, 2016d, Fig. II.5.10). To generalise from so-called 'pre-vocational' programmes which are designed for a minority of students with very low levels of performance seems first to be another case of reverse causality, and second to be very misleading, since VET programmes have different designs and objectives and target different students.

It is also a matter of concern that countries that are well-known for having developed VET systems at earlier ages and to a much larger extent than most others are treated in PISA 2015 (OECD, 2016b, Table II.5.14) as having very few students enrolled in such programmes. For example, Germany and Switzerland, which are prime examples of European countries with well-developed VET systems from early ages, only have 2.7% and 9.2% of fifteen-year-old students enrolled in VET according to PISA, a much lower proportion of students than has widely been reported for those countries, even by other OECD publications (OECD, 2020d: over 20% of fifteen to twenty-four-year-olds are enrolled in VET in both countries). Other non-OECD countries which have developed a combination of academic and VET programmes from early ages, such as Singapore, have no students enrolled in VET according to PISA. It is unclear whether these problems are to do with the quality of the data or with how programmes have been classified but, in any case, they do not reliably represent those education systems.

The second issue is that what PISA results actually show is that student performance is lower in VET programmes in half of the countries considered (50%), not significantly different from academic programmes in a third of the countries considered, and higher than that of students in academic programmes in 20% of the countries considered (OECD, 2016d). Thus, fifteen-year-olds enrolled in VET or pre-vocational programmes have lower levels of performance in some countries, but by no means in all countries. In Luxembourg, Switzerland, Japan and most Latin American countries, students in VET programmes perform better than students in academic programmes.

Finally, in some of the countries with the highest rates of enrolment of fifteen-year-old students in VET programmes, such as Austria (71%), Italy (50%) and the Czech Republic (33%), the performance of

these students is similar to that of students in academic programmes. Furthemore, if we consider those countries where students can choose between academic and VET programmes at early ages, PISA finds lower performance among VET students in Belgium (twelve years) and the Netherlands (twelve years), no significant difference between academic and VET students in Austria (ten years) and Germany (ten years), and better performance among VET students in Switzerland (twelve years) and Luxembourg (thirteen years) (OECD 2016d).

In conclusion, since PISA assesses the performance of fifteen-year-olds, and in most countries the choice between academic and VET programmes does not take place until upper-secondary education when students are older, this survey cannot properly address the question of whether students enrolled in VET have different levels of performance from those following academic programmes. Even among education systems where differentiation between both types of programmes starts at an early age (i.e. between ten and thirteen years), there is no conclusive evidence that VET students perform worse in PISA. Thus, the widely accepted recommendation that VET should be delayed as much as possible to avoid generating inequalities at early ages seems unfounded.

The point is not whether VET should be delayed in order to postpone any assumed pernicious effects upon student performance as far as possible. The question is which VET models avoid such harmful effects. To understand this, it is necessary to undertake a brief historical overview (Busemeyer and Trampusch, 2011; OECD, 2018b and 2019b).

Traditionally, apprenticeships were designed to train people in a specific set of skills required to enter a trade. In some countries, these apprenticeship systems remain strong and are the responsibility of firms, who set the standards, provide the training, and offer contracts. This is the case for Germany and Switzerland. But in most countries the traditional apprenticeship model declined as education systems expanded and developed vocational education and training programmes which led to educational degrees. Initially, these VET systems were designed as an alternative pathway for students with low academic performance and equipped them with a rather narrow set of technical skills that allowed them to move rapidly into low-skilled manual jobs. In contrast, students with higher academic performance who aspired

to get high-quality, well-paid jobs followed the academic track that allowed access to university. However, this model has become obsolete over time since societies have gradually become more educated and a greater proportion of people have higher levels of skills, and thus aspire to obtain high quality jobs. In parallel, most countries have evolved into knowledge economies where many traditional low-skilled jobs have disappeared due to automation and outsourcing, and a greater share of the labour market consists of middle- and high-skilled jobs (OECD, 2020e).

These changes have led to a major transformation of VET systems in many countries (Busemeyer and Trampusch, 2011; OECD, 2019b, 2020d and 2020e). Modern VET systems are designed for students of all levels of performance, since they prepare them to obtain good quality jobs in high demand. In order to become attractive to a broader range of students, these VET programmes equip participants with strong foundation skills so that they can engage in lifelong learning. This is badly needed in rapidly changing labour markets where people can no longer expect to have a 'job for life' and may even need to move from one sector to another. In fact, modern VET systems offer many advantages in dynamic labour markets, since their strong links with the labour market allow them to more easily track the changes taking place (due to the impact of megatrends) and to respond more efficiently by equipping people with the right skill bundles.

Ideally, education systems should create bridges between academic and VET programmes, so that the latter are not regarded as dead ends, and students in both programmes have the possibility of moving into tertiary education. In addition, VET programmes are more effective when they establish links with the labour market by increasing the amount of time that students spend training at work; this will ensure that they acquire the skills required by the labour market, and will avoid the need for VET schools to constantly update equipment in order to track changes taking place in working environments (OECD, 2018b).

The available data clearly show that VET systems represent smoother transitions to the labour market, since upper-secondary VET graduates enjoy higher employment rates than upper-secondary graduates in academic programmes (OECD, 2020d and 2020e). Furthermore, in more than 30% of OECD countries, upper-secondary VET students have

similar or higher rates of employability than tertiary graduates (OECD, 2020d), highlighting the fact that a university degree is not the only (or necessarily the best) route to a job. The countries in which secondary VET graduates enjoy higher employment rates tend to have a strong component of work-based learning, as is the case in Austria, Germany, Sweden and Switzerland (OECD, 2020d). Some studies suggest that this advantage weakens over people´s lifetimes, possibly because the skillset that VET students acquire becomes obsolete over time, due to technological and structural changes in the labour market (Brunello and Rocco, 2017; Forster *et al.*, 2016; Hanushek *et al.*, 2011 and 2017; Rozer and Bol, 2019). Most VET graduates are employed in middle-skill and low-skill occupations, but 20% of young VET graduates are employed in high-skill occupations (OECD, 2020e). However, this share increases in countries like Germany and Switzerland, where more than one third of VET graduates work in high-skill occupations.

In conclusion, VET systems facilitate school-to-work transitions, resulting in better labour market outcomes for VET graduates compared to general education graduates and, in some countries, even higher than those of tertiary graduates. Countries with strong VET systems which have adapted to the increased demand for high levels of skills from labour markets do ensure that VET graduates work in middle- and high-skills occupations. In parallel, VET systems are effective in reducing dropout rates, since they offer a more applied, work-based learning environment which may be better-suited to students who are not motivated by the academic programmes, or who need to enter the labour market earlier (Henriques *et al.*, 2018; Kulik, 1998). This was clearly the case in Spain, where an education reform which modernised VET and made it more attractive to a wider range of students resulted in a substantial reduction in early school leaving (Gomendio, 2021; Wert, 2019).

4.5. Conclusions

The evidence provided by ILSAs has proven to be very useful for comparing education systems directly and assessing how they evolve over time. These international benchmarks have revealed huge differences in student performance between education systems, raising important questions about which factors improve quality. The vast amount of data

generated have allowed quantitative analyses to identify these factors, and have contributed to a much-needed shift in the educational debate from inputs (i.e. investment) to outputs (i.e. student outcomes).

The international surveys differ in the target population, periodicity and methodology, and focus on evaluating student performance in reading, mathematics and science. While PIRLS and TIMSS assess how much students in specific grades in primary and secondary education learn from the curriculum, PISA claims to measure the extent to which fifteen-year-olds (irrespective of grade) have acquired twenty-first century skills and are able to solve unfamiliar problems in knowledge-based societies. Thus, while PIRLS and TIMSS establish clear links with the curriculum taught in school, PISA openly defines a more ambitious target: to measure what fifteen-year-olds can do with the knowledge acquired, irrespective of whether it has been learned at school, at home, or in their social environment. Despite the more tenuous links between what PISA measures and the learning achieved at school, PISA is more policy-oriented and boasts about its impact on education policies.

When ILSAs are compared in terms of national performance, a very consistent picture with clear geographical differences emerges: top performers are countries in East Asia, low performers are mostly low- and middle-income countries in Latin America, Africa and the Near Middle East, and mid-performers are mostly European and North American countries alongside New Zealand and Australia. International surveys also reveal that differences between regions within countries are sometimes larger than differences between countries. Thus, despite their differences, ILSAs seem measure similar features of student performance.

In contrast, there seem to be significant differences between surveys when trends over time are analysed: while PISA claims that between 2000 (first cycle) and 2018 (last cycle) no significant changes in student performance occurred in most participating countries or in OECD countries, both PIRLS and TIMSS reveal a more positive trend of improved performance in many more countries. The divergence between PISA and the other surveys seems to have become more accentuated in 2015 and 2018 when PISA introduced substantial methodological changes to respond to the needs of an increasing number of low- and middle-income participating countries. Thus, in order to provide more

granular information to low- and middle-income countries about student performance at the lower end of the range, the reliability and comparability of the information provided to high-income countries may have been sacrificed. This highlights the trade-offs when surveys grow rapidly in terms of participating countries and performance levels to the extent that newcomers, who tend to have low levels of performance, need different types of information in order for the survey to be relevant.

In any case, the fact that according to PISA no significant improvements have taken place after almost two decades represents a failure of its self-proclaimed mission: to identify good practices, to advise governments on which policies should be implemented and, in this way, to enhance student performance all over the world. PISA claims that policymakers are at fault because they have failed to implement such good practices, but before shifting the blame to governments, a detailed analysis of PISA recommendations is required to refute alternative hypotheses. What are the main PISA policy recommendations? Are they consistent and solid?

The most robust conclusion from international surveys is that, above a rather low threshold, levels of investment are unrelated to student performance. This holds true for most participating countries, with the exception of the poorer nations. The lack of association also becomes apparent when regions within countries are compared in terms of investment per student. Similarly, changes in levels of investment over time are unrelated to changes in student performance. In other words, increases in investment do not lead to better student outcomes, and decreases in investment do not lead to declines in student outcomes. It is remarkable that the most solid conclusion has had so little impact on the educational debate, which systematically assumes that there is a causal link between levels of investment and quality. It has been argued that what matters is how resources are invested, rather than the absolute amount.

After universal access to education has been achieved and schools and facilities have been built, which is the case in most countries that participate in ILSAs, investment in education mostly translates into investment in teachers (and other staff). As a result, total investment is the result of two main variables: the number of teachers (which is in turn the result of the number of students and class size) and their

salaries. The evidence clearly shows that class size has no impact on student performance but, after making this conclusion clear in all previous cycles, in 2018 PISA does recommend reducing class size for reasons that remain unclear and are not supported by the data provided in this cycle (or any other). Countries in East Asia have very large class sizes because they have made a conscious trade-off: they invest most of their resources in selecting, training and paying a high-quality, albeit reduced, teaching force.

In other countries, this has not been possible because class size is what matters most to unions, since it determines the number of teachers and therefore the size of their membership and ultimately their power. In addition, parents intuitively associate small class sizes with individualised teaching and a higher quality of education. These conflicts of interest have meant a high political cost for increasing or even maintaining class sizes in most countries. This has led to decreases in class size over time, a trend which has major consequences in the medium to long term: resources are needed to pay salaries to a larger number of teachers and therefore selection processes are not as demanding, training is of a lower quality and professional development is poorly elaborated. The dire consequences of this choice are particularly apparent in Latin American countries.

There is also no evidence that teacher salaries are associated with student outcomes, although they need to be above a certain threshold in order to attract good candidates. However, incentives linked to performance do have a positive impact on student learning gains. The additional advantage of such policies is that performance-related pay requires a fraction of the resources that are needed to implement salary increases at the systemic level. However, such incentives are rare.

It is truly remarkable that despite consensus about the relevance of teacher quality to achieve good student outcomes, so little is known about what makes teachers effective. Studies have shown that the impact of a good or bad teacher on student performance is huge, but the precise features that make teachers effective remains unclear to the extent that the OECD refers to this gap in knowledge as "the black box". The main drawback seems to be that few attempts have been made to link student performance to teacher quality, beyond subjective assessments of 'self-efficacy' made by principals and teachers themselves. In addition, those

variables which are easy to quantify such as educational degrees or years of experience do not reveal any clear links, since teachers in most countries have university degrees or certificates but the quality differs dramatically from country to country. One exception is a study which finds a strong relationship between the level of basic skills (numeracy and literacy, as measured by PIAAC) of teachers and student performance, and also shows that high levels of skills among teachers are the result of selection processes which target teaching candidates at the top end of their country's skills distribution.

Perhaps the next most robust finding is that external and standardised student assessments (also known as 'exit exams') are linked to higher levels of student performance. However, the consistency of this result sharply contrasts with how PISA recommendations have evolved over time: in the first cycles PISA reached clear conclusions based on comparative evidence about the positive impact of such student assessments, but eventually started to warn against the negative side-effects of high-stakes exams (such as undue pressure on students and teachers with a negative impact on their wellbeing or potential discrimination against disadvantaged students who may lose motivation when faced with ambitious targets) until it shifted to a narrative that supported so-called 'formative assessments' by teachers.

Clearly such assessments are also useful, but there is no reason why they should not be combined with external assessments, which define the same standards for all schools and teachers, are useful tools for detecting struggling students early enough to provide them with effective support, represent clear incentives for all students and teachers to achieve common targets, and provide information about how different schools or regions are performing using the same metrics in case interventions are required. Many analyses using data from PISA and other ILSAs have shown that exit exams have a clear impact on student performance, so it is unclear why PISA's policy recommendation has changed over time.

The evidence from international surveys also shows that giving more autonomy to schools has a positive impact, but only under certain conditions. The first is that greater autonomy leads to better student outcomes when implemented along with accountability measures. Greater school autonomy means many things: principals may have

more decision-making power in relation to budget allocation, the degree of specialisation in certain knowledge areas, or the amount of time assigned to different subjects, while teachers may be able to decide which materials they will use, their pedagogical practices, internal assessments, and (to a certain extent) curricular content. But it is surprising that in most countries school autonomy is not what matters the most: principals are rarely able to select their teachers, nor do they have the power to dismiss underperforming teachers. When these responsibilities are transferred to principals and teachers, it is important to evaluate whether they make the right decisions to improve student performance. This is why accountability mechanisms, which in most cases are based on the results of standardised external assessments, should go hand-in-hand with more autonomy.

The second condition for school autonomy to work is that both principals and teachers must have high levels of skills and receive the necessary training before assuming new responsibilities. The evidence indicates that while greater school autonomy has a positive impact when teacher quality is at good levels, it has a negative impact in developing countries where low teacher quality implies that the education system is more efficient if there are stricter guidelines about the curriculum, assessments, and classroom materials. PISA often fails to acknowledge the conditionality attached to granting schools greater autonomy in order to ensure their effectiveness, and makes a universal recommendation in favour of high levels of school autonomy.

The extent to which parents should be able to choose the type of school which they think is best for their children is the subject of much controversy. The complexity of the debate is partly due to the fact that for parental choice to be meaningful, there needs to be a diverse array of schools. Such heterogeneity is achieved mainly through government-funded, privately managed (charter) schools. The mere existence of this type of school is a highly charged political issue in many countries, with supporters arguing that they represent the diversity of values prevalent in modern societies, and detractors claiming that they create even more profound divides in societies where cultural integration remains problematic, and that admission policies tend to favour students from privileged backgrounds, leaving disadvantaged students and migrants overrepresented in public schools.

PISA's analyses cannot contribute to this debate because all of their comparisons lump government-funded, privately managed schools and private schools into a single category. More detailed analyses using data from ILSAs clearly show that competition between different types of schools leads to improvements in student performance and provide very solid evidence that government-funded, privately managed schools are more efficient in the sense that they achieve better outcomes with fewer resources. This is partly because they have more autonomy and more accountability, and principals have much more power to choose their team of teachers, an option which is usually lacking in public schools.

In political, academic and media debates, the most contentious issues have to do with the other dimension of education systems: equity. This is due to the strong ideological component of such debates, as well as the difficulties associated with interpreting different ways of measuring it. While it is widely accepted that quality is measured by student performance, equity is multidimensional, and many different measures have been proposed that actually convey very different types of information. All analyses of data from ILSAs reveal that student socio-economic background is the factor that has the greatest impact on student performance. The impact of family socio-economic background is evident in all countries, but to different extents. The broader and most challenging question is to what extent such differences between countries reflect how egalitarian societies are, or whether they are mainly the result of the implementation of policies that minimise the impact of inequity.

What the data tells us is that good-quality education systems raise the performance of all students, but those in the top percentiles of a country's socio-economic distribution perform better than those in lower percentiles. PISA concludes that since differences between countries in student performance are huge, poor students in good-quality education systems perform better than privileged students in countries with low-quality education systems. But this conclusion fails to take into account the fact that students in the lowest percentiles in rich countries are not as poor as those in the equivalent percentiles in poor countries.

More sophisticated analyses using data from ILSAs have provided a more realistic and complex picture: poor students in rich countries (which tend to have higher-quality education systems) actually

perform better than rich students in poor countries. This is probably the consequence of systemic deficiencies, such as low curricular standards, teachers with low levels of skills and poorly designed assessments, which parental resources cannot overcome. These studies also show that in countries with high levels of inequity (as measured by the Gini coefficient) the impact of household income upon student performance is much greater than in more egalitarian societies.

These findings have important implications. They suggest that education systems cannot overcome the impact of social and economic inequalities when these are profound. They also caution against establishing causal links between specific education policies which have been deployed in egalitarian societies with equitable outcomes (since the confounding variable is that high levels of equity are already present in such countries), as well as the risk of assuming that transferring those policies to countries with high levels of inequity will contribute to the reduction of inequality in student outcomes. It seems more likely that social and economic equity permeates education systems which, as a result, do not require major interventions against inequity, while less egalitarian societies face very different challenges that do require specific policies to minimise the impact of inequality.

A major challenge for education systems, which is exacerbated in countries with high levels of inequity, is the question of how to deal effectively with the degree of student heterogeneity found in the same grades and classrooms, and ensure that struggling students are not left behind while those that can become top performers advance at a more rapid pace. In most countries, principals and teachers identify differences between students' levels of performance as the main obstacle to learning, but this challenge is magnified in less egalitarian societies. Thus, a number of policies have been developed to reduce variation in student ability when it compromises learning gains. These include ability grouping, separation of students into academic and vocational tracks, and grade repetition. There is intense controversy both among policymakers and academics about the pros and cons of practices which aim to reduce student heterogeneity in academic performance. The clear advantage is that teachers will find it easier to make progress if they teach a group of students with a similar level of performance who can follow at a similar pace and have similar needs.

When teachers are faced with a heterogeneous group of students, their efficiency may be compromised, since they must make choices about whether to focus on the low-performers, the top-performers or the average students, thus failing to meet the very diverse needs of their students. But opponents claim that any practices which separate students according to performance will harm low-performing students who will not be allowed to learn from their high-achieving peers, thus exacerbating inequality, and in most cases leading to discrimination based on socio-economic background or immigrant status. From this viewpoint, these practices are seen as non-inclusive and are referred to pejoratively as "segregation".

The recommendations from PISA are consistent with this discourse and therefore discourage countries from any practice which aims to reduce heterogeneity in student performance, because it is assumed that this approach will lead to segregation and will increase inequity. Thus, PISA does not recommend ability grouping, early tracking or grade repetition.

These conclusions are not supported by PISA data, so they must be challenged, even if they align with mainstream ideas. In the case of VET, PISA data cannot compare the performance of fifteen-year-olds in academic vs VET programmes because in most countries the latter do not start until the age of sixteen. Thus, the data used to support this conclusion are flimsy at best. In the case of grade repetition, PISA seems to fall into the well-known reverse causality trap: since the performance of students who repeat a grade is lower, grade repetition lowers performance. Obviously, when students repeat a grade it is because their level of performance is much lower, and not the other way round. Finally, conclusions regarding ability grouping suffer from a similar problem: if ability grouping is used more often when student performance levels show huge variation in non-egalitarian societies, the association between the two cannot be used as proof that ability grouping increases inequality.

The available evidence suggests that practices which aim to reduce student heterogeneity and cater for different needs and interests, such as ability grouping and differentiated general and VET programmes, do not decrease student performance. Furthermore, ability grouping seems to benefit low-performing students the most, while VET programmes

can decrease early school leaving and equip students with the skills required to obtain middle- and high-skill jobs without compromising their performance. Obviously, any differential treatment of students carries a hidden risk of discrimination. Poorly designed ability grouping could result in students from low socio-economic backgrounds being unfairly assigned to low-performing groups, therefore limiting their chances of making progress. Similarly, old-fashioned VET systems may target students from underprivileged backgrounds and equip them with such a narrow set of skills that they can only aspire to low-skill jobs. The fear that education systems may fall into these traps does not seem to be supported by the evidence. But it is this fear that leads to recommendations to treat all students equally, which is widely regarded as an inclusive strategy.

Contrary to conventional wisdom, the evidence suggests that not allowing any differentiation may lead to inequitable outcomes, at least in some contexts. It seems reasonable to argue that in countries where there are major differences in the skill levels of the population, differentiation is needed to a greater extent than in more uniform societies. This is the case because in societies where parents' abilities, not only in terms of resources but also in terms of skills, differ to a large extent, children born to parents with low skill levels will have a much more difficult starting point when entering compulsory education. In the worst-case scenario, a lack of ability grouping may leave struggling students behind and, if there are no other alternatives, these students will lag further and further behind until they start repeating grades. A lack of alternative learning paths such as VET programmes that could be more attractive to students seeking more practical training may result in high drop-out rates. The needs of disadvantaged students will not be addressed if they receive the same treatment as other students. This may be a safeguard against potential discrimination but it is by no means a solution to the very real problems. When student heterogeneity becomes an obstacle to learning, offering different pathways allows the education system to have the flexibility to adapt to the diverse needs of the student population.

In conclusion, PISA claims that the evidence it provides about good practices lowers the cost of reforms to policymakers and increases the costs of inaction. The detailed review of the evidence provided by

PISA and other ILSAs unfortunately shows that this is not the case for three main reasons: (a) since most good practices are strongly context-dependent, it is difficult for policymakers to understand precisely what applies to their own country; (b) PISA conclusions are based on its own analyses, which are limited to correlations that cannot establish causal links; and (c) some of the conclusions that PISA draws are not supported by strong and objective data.

Nonetheless, data from ILSAs have proven incredibly useful when more sophisticated statistical techniques have been used, but there are only a few robust conclusions about the factors that do or do not have an impact on student performance: investment in education does not equal quality and the corollary is that class size and teacher salaries do not have any impact; teacher quality matters a lot, but a clear understanding of what it entails is still lacking; student assessments and school choice do have a positive impact; school autonomy has a positive impact only in high-quality education systems and when implemented along with accountability mechanisms; policies that minimise student heterogeneity are required in unequal societies, but not in egalitarian societies, where there are higher levels of student uniformity.