8

Where did performance change between 2015 and 2018?

This chapter discusses short-term changes in student performance – both in mean performance and in the performance distribution – in the PISA assessment between 2015 and 2018.

This volume has so far focused on performance in reading, mathematics and science as measured by the 2018 round of the PISA assessment. However, PISA allows for more than just a snapshot of an education system's performance at a given moment: as a long-term study, dating back to 2000, PISA gives countries and economies an opportunity to see how their performance has evolved over the course of almost two decades.

Chapter 9 discusses long-term trends in student performance. This chapter examines changes in performance between the previous PISA assessment, which took place in 2015, and the latest 2018 assessment. Any changes in performance over such a short period of time can likely be related to, if not attributed to, changes in education policy, in the learning environment (both in and outside of school), and in the composition of student populations that affected children who were 15 years old between 2015 and 2018 (i.e. those born between 1999 and 2002).

What the data tell us

- On average across OECD countries, mean performance in reading, mathematics and science remained stable between 2015 and 2018.
- There were large differences between individual countries and economies in how their performance changed between 2015 and 2018. For example, mean performance in reading improved in 4 countries and economies (Macao [China], the Republic of North Macedonia, Singapore and Turkey), declined in 13 countries/economies, and remained stable in the remaining 46 countries/economies.
- Between 2015 and 2018, the performance distribution in both reading and mathematics widened, on average across OECD countries; but the performance distribution in science neither widened nor narrowed significantly during that period.

In order to attribute changes in performance across PISA cycles to changes in student learning or to differences in the composition of student populations, the PISA test and how it was conducted would have had to remain equivalent from cycle to cycle. Differences in how the test was conducted – such as the length of the test, whether students take the test on paper or on computer, or whether they sit the test in the morning or the afternoon – could affect student motivation and performance; therefore, these differences must be monitored and minimised.

Overall, PISA 2018 and PISA 2015 were conducted in much the same way:

- As in 2015, the vast majority of students who sat the PISA 2018 assessment answered questions in just two subjects, devoting one hour to each: the major domain (reading in 2018, science in 2015) and one other domain (OECD, forthcoming[1]). In previous rounds of PISA, the number of subjects varied more across students: while large numbers of students were tested in two subjects, a significant proportion was tested in three subjects within the same two-hour testing period.
- The assessment was primarily conducted on computer in both 2015 and 2018, whereas it was conducted on paper in 2012 and earlier. While measures were taken in 2015 to align the computer-based tests with the original paper-based scales, these measures were implemented mainly at the international level. Country-specific differences in familiarity with computers, or in student motivation when taking the test on computer or on paper, could still interfere with performance trends (OECD, 2016[2]). For most countries, this mode-related source of uncertainty no longer existed when comparing 2015 and 2018. Furthermore, test administration was more regimented when computer-based assessments were used as there was less room to deviate from standard procedures (e.g. when to take breaks and how long such breaks last).

Annex A8 further explores differences in students' effort and motivation across countries and over time.

At the same time, it is important to assess the impact of using different test items in different years,[1] resulting in performance scales that are not identical. This potential source of error when examining changes in PISA results is summarised by the **link error**, which provides an estimate of the shift in the same subject scale used in two different years. The reporting scales for a subject between two different years are uniformly misaligned by a certain amount. The magnitude of this amount and its direction (i.e. whether a score in one year is equivalent to a higher or lower score in the other year) is unknown, but it is on the order of the size of the link error. For example, the link error between 2015 and 2018 in the reading assessment is roughly four score points, and hence a change of up to eight score points in a country's mean reading performance between 2015 and 2018 would not be significant as it could easily be attributed to the link error.

The link error between 2015 and 2018, however, is noticeably smaller than the link error between other PISA cycles (e.g. between 2012 and 2018, or between 2012 and 2015). In addition to the two reasons listed above concerning how PISA was conducted in 2015 and 2018, there are two further reasons for this smaller link error:

- There were more items in common between the 2015 and 2018 assessments than there were between previous sets of assessments.
  - The 2015 and 2018 mathematics assessments were virtually identical, as they were both minor domains based on the 2012 PISA mathematics framework.
  - The items in the 2018 science assessment were a subset of the items in the 2015 science assessment; the majority of these items were created in 2015 to reflect the updated PISA 2015 science framework and hence differ from items used in assessments prior to 2015.
  - Although new items were developed for the PISA 2018 reading assessment to reflect its new framework (see Chapter 1), a large number were retained from the items developed for PISA 2009 and used between PISA 2009 and 2015.
- In contrast to the procedures used in previous cycles, the characteristics of trend items (those items that were also used in previous cycles of PISA; in this case, in 2015) were assumed to be identical in 2015 and 2018. In practice, item characteristics in 2018 were assumed to be identical to those in 2015, unless there was sufficient evidence of non-equivalence. This resulted in more consistent measurement scales across cycles, and reduced the link error.[2] Items that were unique to one year, however, did not aid in linking scales across years.

In summary, changes between 2015 and 2018 were more precisely estimated than changes involving earlier years. This chapter explores these short-term changes in performance.

CHANGES BETWEEN 2015 AND 2018 IN MEAN PERFORMANCE

Figure I.8.1 shows the changes between 2015 and 2018 in mean performance in reading. On average across OECD countries, mean performance in reading did not change significantly during the period. The decline in performance was most pronounced in Georgia and Indonesia, where it exceeded 20 score points; it exceeded 10 score points in Colombia, the Dominican Republic, Japan, Luxembourg, the Netherlands, Norway, the Russian Federation (hereafter "Russia") and Thailand.

By contrast, several countries/economies saw significant improvements in reading performance. The largest were seen in the Republic of North Macedonia (hereafter "North Macedonia") (41 score points) and Turkey (37 score points), while improvements of between 10 and 20 score points were observed in Macao (China) and Singapore (Figure I.8.1).

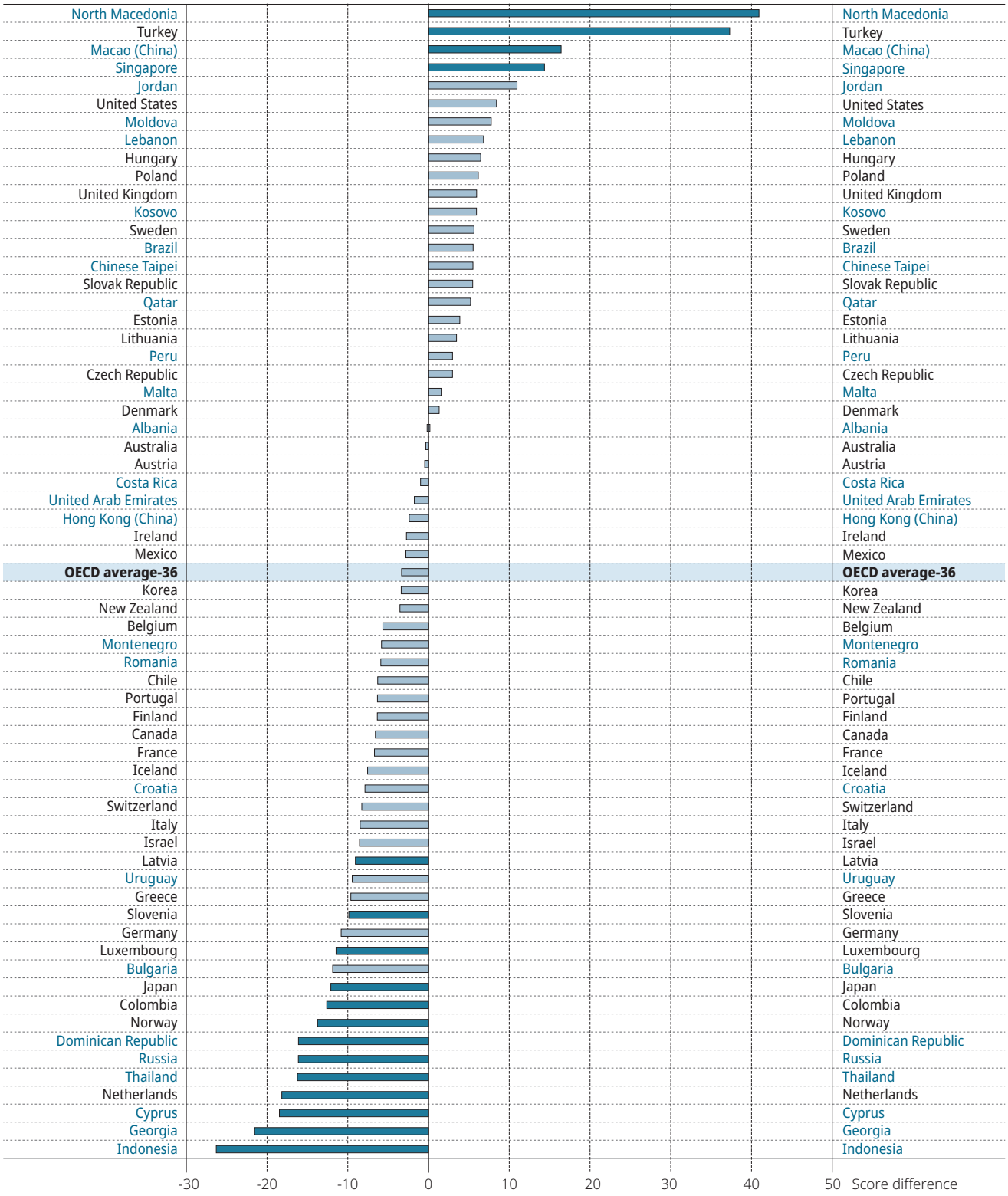
On average across OECD countries, no significant change in either mathematics or science performance was observed between 2015 and 2018. Mathematics performance declined in only three countries/economies (Malta, Romania and Chinese Taipei) during the period, while it improved by over 10 score points in 11 countries/economies (Albania, Jordan, Latvia, Macao [China], Montenegro, North Macedonia, Peru, Poland, Qatar, the Slovak Republic and Turkey). Improvement was notable in Turkey (33 score points), Albania (24 score points) and North Macedonia (23 score points) (Table I.B1.10).

Country-level improvements in science performance were far less common. Improvements of 10 points or more between 2015 and 2018 were observed in only four countries/economies: Turkey (43 score points), North Macedonia (29 score points), Jordan (21 score points) and Macao (China) (15 score points). Science performance declined by at least 10 score points in seven countries/economies: Georgia (28 score points), Bulgaria (22 score points), Chinese Taipei (17 score points), Kosovo (14 score points), Italy (13 score points), Albania (10 score points) and Switzerland (10 score points) (Table I.B1.12).

Most countries and economies did not observe significant changes in performance between 2015 and 2018, when considering each subject independently. This is to be expected. A lack of improvement over three years is not necessarily a cause for concern: education is cumulative and any changes in policy are both incremental and can take years, if not an entire cohort of school-aged children, to have an effect. Moreover, the precision with which differences can be measured means that differences that may be significant in the long term are not deemed significant in the short term. Indeed, in 24 countries and economies out of the 63 that took part in both PISA 2015 and 2018 (Austria, Belgium, Brazil, the Czech Republic, Chile, Costa Rica, Croatia, Estonia, France, Germany, Greece, Hong Kong [China], Hungary, Ireland, Israel, Korea, Lebanon, Lithuania, Mexico, Moldova, New Zealand, Sweden, the United Arab Emirates and the United States), no significant change in performance was observed, between 2015 and 2018, in any of the three core subjects that PISA assessed (Table I.8.1).

During the period, performance improved across all three subjects in Macao (China), North Macedonia and Turkey; and performance improved across two subjects and stayed stable in the third in Jordan and Poland (Table I.8.1).

Figure I.8.1 **Change between 2015 and 2018 in mean reading performance**



Notes: Statistically significant differences between PISA 2015 and PISA 2018 are shown in a darker tone (see Annex A3).

The change in reading performance between 2015 and 2018 for Spain is not reported; see Annex A9. OECD average-36 refers to the arithmetic average across all OECD countries, excluding Spain.

Countries and economies are ranked in descending order of the change in reading performance between PISA 2015 and 2018.

Source: OECD, PISA 2018 Database, Table I.B1.10.

StatLink https://doi.org/10.1787/888934028691

Encouragingly, in no country or economy did performance decline across all three subjects. However, in seven countries/economies – Georgia, Japan, Luxembourg, Malta, Norway, Slovenia and Chinese Taipei – performance declined in two subjects and remained stable in the third (Table I.8.1).

Table I.8.1 **Change between 2015 and 2018 in mean performance in reading, mathematics and science**

	Reading	Mathematics	Science
Mean performance improved between 2015 and 2018	Macao (China), North Macedonia, Singapore, Turkey	Albania, Iceland, Jordan, Latvia, Macao (China), Montenegro, North Macedonia, Peru, Poland, Qatar, the Slovak Republic, Turkey, the United Kingdom	Jordan, Macao (China), North Macedonia, Poland, Turkey
Mean performance did not change significantly between 2015 and 2018	**OECD average-36**, Albania, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, Costa Rica, Croatia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hong Kong (China), Hungary, Iceland, Ireland, Israel, Italy, Jordan, Korea, Kosovo, Lebanon, Lithuania, Malta, Mexico, Moldova, Montenegro, New Zealand, Peru, Poland, Portugal, Qatar, Romania, the Slovak Republic, Sweden, Switzerland, Chinese Taipei, the United Arab Emirates, the United Kingdom, the United States, Uruguay	**OECD average-37**, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, Colombia, Costa Rica, Croatia, the Czech Republic, Denmark, the Dominican Republic, Estonia, Finland, France, Georgia, Germany, Greece, Hong Kong (China), Hungary, Indonesia, Ireland, Israel, Italy, Japan, Korea, Kosovo, Lebanon, Lithuania, Luxembourg, Mexico, Moldova, the Netherlands, New Zealand, Norway, Portugal, Russia, Singapore, Slovenia, Spain, Sweden, Switzerland, Thailand, the United Arab Emirates, the United States, Uruguay	**OECD average-37**, Austria, Belgium, Brazil, Chile, Colombia, Costa Rica, Croatia, the Czech Republic, the Dominican Republic, Estonia, France, Germany, Greece, Hong Kong (China), Hungary, Iceland, Indonesia, Ireland, Israel, Korea, Latvia, Lebanon, Lithuania, Mexico, Montenegro, Moldova, the Netherlands, New Zealand, Peru, Qatar, Romania, Russia, Singapore, the Slovak Republic, Sweden, Thailand, the United Arab Emirates, the United Kingdom, the United States
Mean performance declined between 2015 and 2018	Colombia, the Dominican Republic, Georgia, Indonesia, Japan, Latvia, Luxembourg, the Netherlands, Norway, Russia, Slovenia, Thailand	Malta, Romania, Chinese Taipei	Albania, Australia, Bulgaria, Canada, Denmark, Finland, Georgia, Italy, Japan, Kosovo, Luxembourg, Malta, Norway, Portugal, Slovenia, Spain, Switzerland, Chinese Taipei, Uruguay

Notes: The change in reading performance between 2015 and 2018 for Spain is not reported; see Annex A9. OECD average-36 refers to the arithmetic average across all OECD countries, excluding Spain.
Source: OECD, PISA 2018 Database, Tables I.B1.10, I.B1.11 and I.B1.12.

CHANGES BETWEEN 2015 AND 2018 IN THE PERFORMANCE DISTRIBUTION

The stability in mean performance across OECD countries and in most PISA-participating education systems masks changes in the distribution of student performance. One way this can be seen is by examining the percentiles of student performance. The 10th percentile is the point on the scale below which 10% of students score. In other words, if all students were ranked from lowest- to highest-scoring, the 10th percentile would be the highest-scoring of the lowest-performing 10% of students. Likewise, the 90th percentile is the point on the scale below which 90% of students score (or, conversely, above which only 10% of students score). The median, or 50th percentile, divides the performance distribution into two equal halves, one above and one below that position on the scale.

The subject whose scales should be most comparable between 2015 and 2018 is mathematics, as the assessment items were virtually identical.[3] No significant change was observed in any of the percentiles of the performance distribution between the 10th and 90th, on average across OECD countries, indicating that neither the strongest- nor the weakest-performing students saw an improvement or a decline in performance between 2015 and 2018. However, the inter-decile range (the gap between the 10th and 90th percentiles, and a measure of the dispersion of student performance) increased by 4 score points between 2015 and 2018, on average across OECD countries (Tables I.B1.14 and I.B1.29). This is possible because, although changes in percentiles across time are affected by the offset between scales in different years (i.e. the link error), which can render their measurements less precise, the inter-decile range is not affected by this offset and is thus measured with greater precision.

There was no significant narrowing or widening in the dispersion of the performance distribution in science between 2015 and 2018, on average across OECD countries. There was also no significant change in performance amongst either the strongest- or the weakest-performing students (Tables I.B1.15 and I.B1.30).

Results from PISA 2015 and 2018 indicated that, on average across OECD countries, the score-point difference in reading performance between weaker and stronger students increased by 11 points during that period. The lower level of precision in measuring changes in performance over time, however, made it impossible to state with confidence that stronger students saw an improvement in their performance, or that weaker students saw a decline in theirs (Tables I.B1.13 and I.B1.28).[4, 5]

The discussion above only applies to the average trend across OECD countries; the distribution in performance in individual countries and economies has evolved differently. For example, the inter-decile range in mathematics performance widened significantly in 8 countries and economies (as did the OECD average), while it narrowed significantly in 2 countries/economies and did not change significantly in the remaining 53 countries/economies for which comparable data for 2015 and 2018 were available (Table I.8.2).

Moreover, there were various reasons for why the inter-decile range changed (or did not change) in these countries/economies. For example, the following could explain why the inter-decile range widened between 2015 and 2018:

- Weaker students became weaker and stronger students became stronger.
- Weaker students became weaker but no significant change was observed amongst stronger students.
- Stronger students became stronger but no significant change was observed amongst weaker students.
- All students across the distribution became weaker, but weaker students showed a greater decline in performance than stronger students did.
- All students across the distribution became stronger, but stronger students showed greater improvement in performance than weaker students did.
- There were no significant changes observed at individual percentages (i.e. no significant changes observed amongst either stronger or weaker students) but the overall distribution grew wider.[6]

Table I.8.2 lists countries and economies by whether their performance distributions in reading, mathematics and science narrowed, widened or did not change significantly in dispersion (as measured by the inter-decile range). It also shows whether the change, or lack thereof, was primarily due to changes amongst weaker students, stronger students or both (or in the case of a lack of change, neither). For example, stronger students became stronger but there was no significant change in the performance of weaker students in the United Arab Emirates in mathematics (Table I.B1.13).

The only country where the performance distribution widened between 2015 and 2018 in all three subjects was the United Arab Emirates; it widened in two subjects and remained stable in the third in Canada, Germany, Hong Kong (China) and Romania.[7] There was no country in which the performance distribution narrowed in all three subjects, although it narrowed in two subjects and remained stable in the third in Bulgaria, France, Georgia, Malta and Montenegro (Table I.8.2).

Table I.8.2 [1/2] **Change between 2015 and 2018 in the performance distribution in reading, mathematics and science**

	Reading	Mathematics	Science
Widening of the distribution			
Weaker students became weaker; stronger students became stronger	Hong Kong (China)		
Weaker students became weaker; no significant change amongst stronger students	Canada, Finland, Germany, Iceland, Israel, Latvia, Norway	Germany, Luxembourg, Romania	Romania, the United Arab Emirates
Stronger students became stronger; no significant change amongst weaker students	Australia, Estonia, Macao (China), Poland, Singapore, Sweden, Chinese Taipei, United Arab Emirates, United States	United Arab Emirates	
Almost all students became weaker, but weaker students declined more so than stronger students did	Netherlands, Russia		
Almost all students became stronger, but stronger students improved more than weaker students did	Turkey		North Macedonia
No significant change at individual points along the distribution, although overall widening of the dispersion	**OECD average-36**, Denmark, Ireland, Mexico, Switzerland	**OECD average-37**, Canada, Costa Rica, Norway, Thailand	Hong Kong (China), Qatar

Table I.8.2 [2/2] **Change between 2015 and 2018 in the performance distribution in reading, mathematics and science**

	Reading	Mathematics	Science
No change in the dispersion of the distribution			
No significant change along most individual points of the distribution	Austria, Belgium, Brazil, Chile, Colombia, Costa Rica, Croatia, Czech Republic, Greece, Hungary, Italy, Korea, Lebanon, Lithuania, Malta, Moldova, New Zealand, Peru, Portugal, Qatar, Romania, Slovak Republic, Slovenia, United Kingdom, Uruguay	Australia, Austria, Belgium, Brazil, Bulgaria, Chile, Colombia, Croatia, Denmark, Dominican Republic, Estonia, Finland, France, Georgia, Greece, Hong Kong (China), Hungary, Iceland, Indonesia, Ireland, Israel, Italy, Japan, Korea, Kosovo, Lebanon, Lithuania, Mexico, Moldova, Netherlands, New Zealand, Portugal, Russia, Singapore, Slovenia, Spain, Sweden, Switzerland, Uruguay	**OECD average-37**, Austria, Belgium, Brazil, Chile, Colombia, Costa Rica, Croatia, Czech Republic, Dominican Republic, Estonia, Finland, Germany, Hungary, Iceland, Indonesia, Ireland, Israel, Japan, Korea, Latvia, Lebanon, Lithuania, Mexico, Moldova, Netherlands, New Zealand, Norway, Peru, Russia, Slovak Republic, Sweden, Switzerland, Thailand, United Kingdom, United States
Stronger students became weaker; no significant change amongst weaker students			Luxembourg, Portugal
Stronger students became stronger; no significant change amongst weaker students		Czech Republic, United Kingdom, United States	
Most students became weaker	Dominican Republic, Indonesia, Japan, Luxembourg, Thailand	Chinese Taipei	Albania, Australia, Canada, Denmark, Italy, Spain, Chinese Taipei, Uruguay
Most students became stronger	North Macedonia	Albania, Jordan, Latvia, Macao (China), North Macedonia, Peru, Poland, Qatar, Slovak Republic, Turkey	Jordan, Macao (China), Montenegro, Poland, Turkey
Narrowing of the distribution			
Weaker students became stronger; stronger students became weaker	Albania		
Stronger students became weaker; no significant change amongst weaker students	Bulgaria, France, Montenegro	Malta	France, Greece, Malta, Singapore, Slovenia
Weaker students became stronger; no significant change amongst stronger students	Jordan, Kosovo		
Almost all students became weaker, but stronger students declined by more than weaker students did	Georgia		Bulgaria, Georgia, Kosovo
Almost all students became stronger, but weaker students improved by more than stronger students did		Montenegro	

Notes: The change in reading performance between 2015 and 2018 for Spain is not reported; see annex A9. OECD average-36 refers to the arithmetic average across all OECD countries, excluding Spain.

Changes in the dispersion of the distribution – widening, narrowing or no change – are measured by the inter-decile range, or the difference in score points between the 90th percentile and the 10th percentile of the student-performance distribution.

Changes in the location of individual percentiles between 2015 and 2018 are estimated with less precision than changes in the mean. For some countries/economies, a significant change in mean performance was observed during the period even though changes in points along the distribution could not be deemed significant.

It is also possible that there was no significant change in the dispersion of the distribution, but that one of the extremities (i.e. the 10th or the 90th percentile) changed significantly, while the other did not. It should be kept in mind that the difference between significant and non-significant changes is, itself, often non-significant.

When there was either a widening or a narrowing of the distribution, there was a change amongst weaker students if student performance at either the 10th or 25th percentile improved or declined and that at the other percentile moved in the same direction or did not change significantly. Likewise, there was a change amongst stronger students if student performance at either the 75th or 90th percentile improved or declined and if that at the other percentile moved in the same direction or did not change significantly. In order to classify a country/economy as one where almost all students became weaker or stronger, at least four of the percentiles examined (the 10th, 25th, 50th, 75th and 90th percentiles) must have declined or improved.

When there was no change in the dispersion of the distribution, at least three individual points along the distribution that were examined (the 10th, 25th, 50th, 75th and 90th percentiles) must have declined or improved in order to say that most students became weaker or stronger in that country/economy. When there was no change in the dispersion of the distribution, student performance at both the 10th and 25th percentiles had to move in the same direction in order to say that weaker students became stronger or weaker; likewise performance at both the 75th and 90th percentiles had to move in the same direction in order to say that stronger students became stronger or weaker.

Source: OECD, PISA 2018 Database, Tables I.B1.13, I.B1.14 and I.B1.15.

Box I.8.1. Reading trends and changes in the reading framework

This chapter discusses changes in reading performance between 2015 and 2018 as if they reflected an evolution in students' abilities during the period. This is likely to be the case for changes in mathematics and science performance, as the 2015 and 2018 assessments in these two subjects were identical or a representative subset of one another. But the framework for the reading assessment changed between 2015 and 2018; hence the evolution in reading performance might be attributable to those changes – particularly in students' relative strengths and weaknesses in certain aspects of reading proficiency that were more or less emphasised in 2018 compared to 2015.[8]

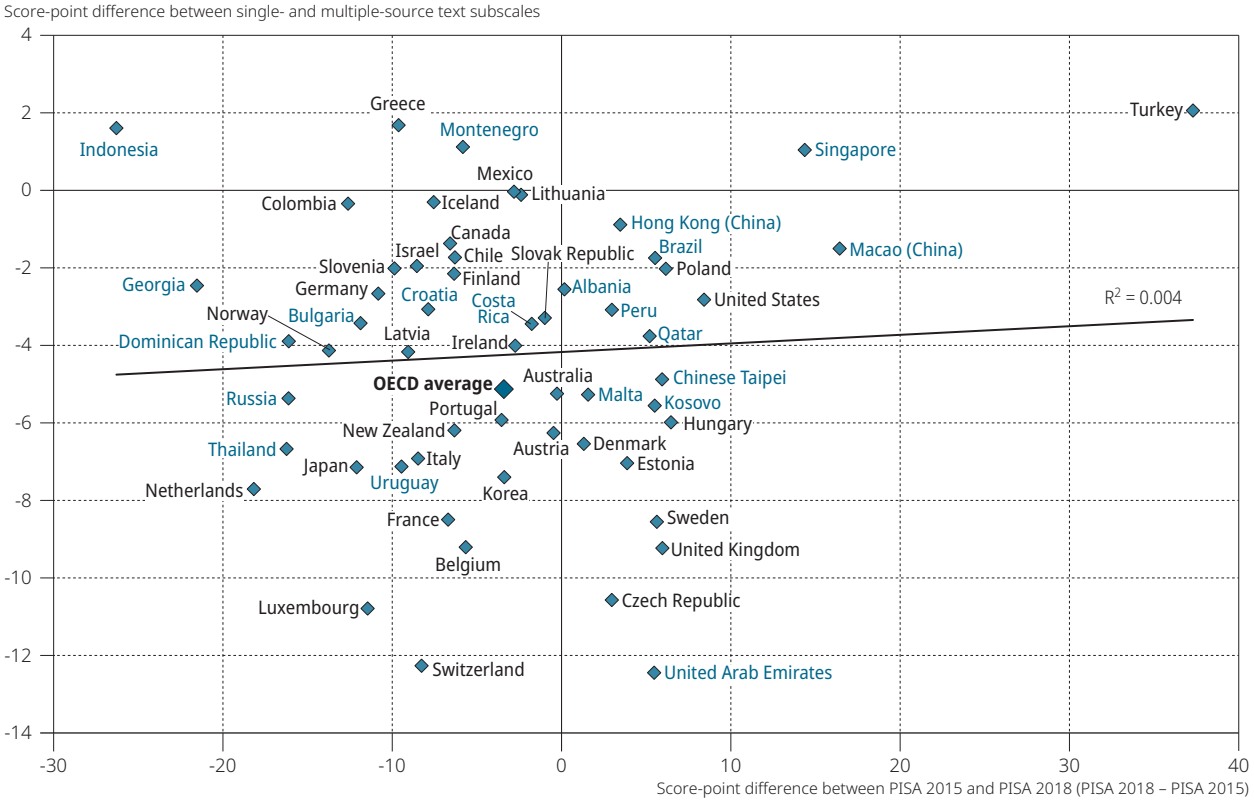
There were two main changes to the framework between 2015 and 2018: the greater focus on multiple-source texts and the inclusion of reading-fluency items. As discussed in Chapter 1, the greater focus on multiple-source texts was made possible by delivering the assessment on computer. Countries and economies whose students were relatively weaker in reading multiple-source texts, for example, might be expected to have more negative trends between 2015 and 2018 than countries whose students were relatively stronger in reading such texts.

Fortunately, it was possible to examine whether this first change to the framework affected student performance. PISA 2018 included a subscale for items that required only a single source and a subscale for those that required reading multiple sources. If changes in performance between 2015 and 2018 were in large part due to the changes in the framework, they would be observed in a correlation between the change in performance and the difference in subscale scores, as both would reflect differences between using single and multiple sources.[9]

Figure I.8.2 shows a scatterplot of the differences in the single- and multiple-source subscales in PISA 2018 versus the change in reading performance between PISA 2015 and PISA 2018. There is no noticeable correlation between the two variables. As a result, it is possible to conclude that the greater emphasis on multiple-source texts in PISA 2018 had a limited impact on changes in reading performance.

Figure I.8.2 Change in reading performance and score differences on reading subscales

Change between 2015 and 2018; performance difference in single- and multiple-source reading subscales



Source: OECD, PISA 2018 Database, Tables I.B1.10, I.B1.19 and I.B1.20.

StatLink https://doi.org/10.1787/888934028710

As mentioned above, the other main change in the framework between 2015 and 2018 was the inclusion of reading-fluency items. These items were presented at the beginning of the assessment. They measured whether students could quickly determine whether certain sentences, such as "The red car had a flat tire" or "Airplanes are made of dogs", made sense. These items were used to determine a student's overall reading score but were not part of any subscale. Hence, the part of a student's score that cannot be explained by his or her subscale scores can be taken as a proxy for his or her accuracy in answering reading-fluency items.[10]

There was no correlation between the change in countries/economies' average reading performance between 2015 and 2018 and the estimated accuracy in answering reading-fluency items. Indeed, R^2 values never exceeded 0.04, regardless of how the estimated accuracy was computed or which subscales (reading process or text source) were used, and the (non-significant) direction of the correlation was highly sensitive to the removal of outliers. As with the greater emphasis on multiple-source texts, the inclusion of reading-fluency items does not appear to explain a large degree of the change in reading performance between 2015 and 2018.

However, there seem to be some overarching factors affecting student performance. The cross-country correlation between changes in reading and mathematics performance between 2015 and 2018 is 0.62; that between reading and science is 0.67; and that between mathematics and science is 0.75. Factors that affect performance across subjects seem to play a bigger role in explaining changes in reading performance than either the emphasis on multiple-source texts in, or the addition of reading-fluency items to, the PISA 2018 assessment.

Notes

1. Even the same test items may not retain identical measurement properties across PISA cycles. For example, with the passage of time, respondents may become more familiar with what was initially an unusual item format or component of the test, such as an equation editor or a drawing tool; or they may no longer recognise a particular situation (such as writing postcards or using video recorders) as familiar.
2. Item parameters for items common to 2015 and 2018 were initially constrained to the best-fit values used in 2015. The parameters for 2018 were allowed to vary from the parameters used in 2015 if they poorly fit the PISA 2018 data. Student scores from PISA 2015 were not affected by this procedure, i.e. there was no rescaling of PISA 2015 data.
3. As discussed in note 1 above, there may still be differences in the mathematics scales between 2015 and 2018 even if the same test items were used. However, these are more limited in scope, and have less impact on comparisons between years, than changes in the questions used in the assessment (as occurred in reading and science).
4. In this situation, where there was no significant change in the dispersion of the performance distribution, "weaker students" refer to those at the 10th and 25th percentiles, while "stronger students" refer to those at the 75th and 90th percentiles.
5. Adaptive testing (see Chapter 1) was implemented for the reading assessment in 2018, allowing for greater precision in measuring student performance at both the high and low ends of the performance distribution. Measurement of performance at the low end of the distribution was also enhanced through the addition of reading-fluency items at Levels 1b and 1c. Prior to 2018, the measurement of scores at the extremes was affected by greater uncertainty. Adaptive testing, which presents stronger students with more difficult questions and weaker students with easier questions, and reading-fluency items both improved the precision in measuring these students' scores and therefore the ability to detect significant differences amongst high- or low-achieving students. Results in mathematics and science were not affected by either adaptive testing or the introduction of reading-fluency items.

6. This discussion only considers changes that were statistically significant. As mentioned in the main text, changes in performance over time are subject to link error and therefore measured with less precision than changes in the inter-decile range (i.e. a narrowing or widening of the distribution), which are not subject to link error.

7. The performance distribution in reading widened between 2015 and 2018 in more countries/economies (25) than did the performance distributions in either mathematics (8) or science (5). However, the changes observed in reading performance between 2015 and 2018 may also reflect changes in the framework and design of the test, and must therefore be interpreted with caution.

8. This annex is only concerned with countries that delivered the assessment on computer; the paper-based assessment continued to use the same framework as that used between 2009 and 2015.

9. Differences between the two subscales do not have a substantive meaning. It is not possible to say that countries that are stronger at reading multiple-source texts than single-source texts if their multiple-source text subscale score is higher than their single-source text subscale score – much as it is not possible to say that countries are stronger at reading than mathematics if their reading score is higher than their mathematics score. However, as these two subscales were scaled together to give the overall reading scale, their differences can be compared across countries. For example, a country whose multiple-source text subscale score is higher than their single-source text subscale score is *relatively* stronger at reading multiple-source texts than a country where the two subscale scores are identical. For more information, see Chapter 5.

10. There were two ways to estimate the part of a student's reading score that could not be determined by his or her subscale scores. In the first method, the overall reading score was linearly regressed over the subscale scores; the part of the overall score that could not be explained by the subscale scores was captured by the residual of the regression. In the second method, a composite overall score was created through a weighted average of the subscores; the weights came from the approximate composition of the reading assessment on either the text source (65% single-source text and 35% multiple-source text) or the reading process (25% "locating information", 45% "understanding", and 30% "evaluating and reflecting"). Chapter 1 of this volume provides more details on the breakdown of the PISA 2018 reading assessment. In this second method, the part of a student's reading score that could not be determined by the student's subscale scores is thus the difference between the reading score and the composite, weighted-average score.

    The process of creating an overall reading score is not a simple linear combination of various subscores and the reading-fluency portion of the assessment, so neither of these methods truly captures students' performance on reading-fluency questions. However, these two methods gave highly correlated estimates of performance in reading fluency (R^2 between 0.86 and 0.88).

References

OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, https://dx.doi.org/ 10.1787/9789264266490-en. [2]

OECD (forthcoming), *PISA 2018 Technical Report*, OECD Publishing, Paris. [1]