2

How PISA results are reported: What is a PISA score?

This chapter presents information about the methods behind the analysis of PISA data and how to interpret the score values; it does not contain results of the PISA 2018 tests. The chapter summarises the test-development and scaling procedures used to ensure that results are comparable across countries and with the results of previous PISA assessments, and explains how the score values can be interpreted.

HOW DOES PISA DEFINE A REPORTING SCALE?

This section summarises the test-development and scaling procedures used to ensure that PISA score points – the unit in which results of the PISA 2018 test are reported – are comparable across countries and with the results of previous PISA assessments. These procedures are described in greater detail in Annex A1 and in the *PISA 2018 Technical Report* (OECD, forthcoming[1]). The test-development procedures described in this section apply, in particular, to the computer-based test, which was used in the vast majority of countries/economies (70 out of 79). The differences between the paper-based test and the computer-based test are described in Annex A5.

How test questions were developed and selected

The first step in defining a reporting scale in PISA is developing a framework for each subject assessed. This framework provides a definition of what it means to be proficient in the subject;[1] delimits and organises the subject according to different dimensions (e.g. the cognitive component skills that underpin proficiency, the types of situations in which proficiency manifests itself, etc.); and identifies factors that have been found, in previous studies, to relate to proficiency in the subject. The framework also suggests the kind of test items (tasks or problems) that can be used within the constraints of the PISA design (e.g. length of the assessment, target population) to measure what students can do in the subject at different levels of proficiency (OECD, 2019[2]).

This test framework is developed by a group of international experts for each subject and is agreed upon by the participating countries. For the assessment of reading, mathematics and science, the framework is revisited every third assessment. For PISA 2018, the reading framework was redeveloped, while the mathematics and science frameworks remained identical to those used in 2015.[2] This new framework for the assessment of reading is summarised in Chapter 1 of this volume.

Once the participating countries and economies agree on the framework, the actual tasks (or items) used to assess proficiency in the subject are proposed by a consortium of testing organisations. This consortium, under contract by the OECD on behalf of participating governments, develops new items and selects items from existing tests, particularly previous PISA tests of the same subject. The expert group that developed the framework reviews the testing instruments – i.e. single items or tasks, as well as the complete electronic test forms and paper booklets – to confirm that they meet the requirements and specifications of the framework. All participating countries and economies review all of the draft items to confirm that the content, cognitive demands and contexts of the items are appropriate for a test for 15-year-olds.

It is inevitable that not all tasks in the PISA assessment are equally appropriate in different cultural contexts, and equally relevant in different curricular and instructional contexts. To address this dilemma, PISA asked experts from every participating country/economy to identify those draft tasks that they considered most appropriate for an international test. These ratings were considered when selecting items for the assessment.

Items that passed these qualitative reviews by national and international experts were translated, and these translations were carefully verified by the PISA consortium.[3] The items were then presented to a sample of 15-year-old students in all participating countries as part of a field trial to ensure that they met stringent quantitative standards of technical quality and international comparability. In particular, the field trial served to verify the psychometric equivalence of the items and test across countries, which was further examined before scaling the results of the main study (see Annex A6).

All countries that participated in the PISA 2018 assessment had to review the test material for curricular relevance, appropriateness and potential interest for 15-year-olds; and all countries were required to conduct a field trial. After the qualitative review and then again after the field trial, material was considered for rejection, revision or retention in the pool of potential items. The international expert group for each subject then formulated recommendations as to which items should be included in the main assessments. The final set of selected items was also subject to review by all countries and economies (see Annex A6). During those reviews, countries/economies provided recommendations regarding the items' suitability for assessing the competencies enumerated in the framework; the items' acceptability and appropriateness in their own national context; and the overall quality of the assessment items, all to ensure that they were of the highest standard possible. This selection was balanced across the various dimensions specified in the framework and spanned various levels of difficulty, so that the entire pool of items could measure performance across all component skills and across a broad range of contexts and student abilities. For further details, see the *PISA 2018 Technical Report* (OECD, forthcoming[1]).

How the electronic test forms were designed

All students completed two hours of testing in two or three subjects.[4] In order to ensure that the assessment covered a wide range of content, with the understanding that each student could complete only a limited set of tasks, the full set of tasks was distributed across several different electronic test forms with overlapping content. Each student thus completed only a fraction of all items, depending on which test form he or she was assigned. This design ensures that PISA can provide valid and reliable estimates of performance at aggregate levels when considering many students together (e.g. all students in a country, or with a particular background characteristic in common).

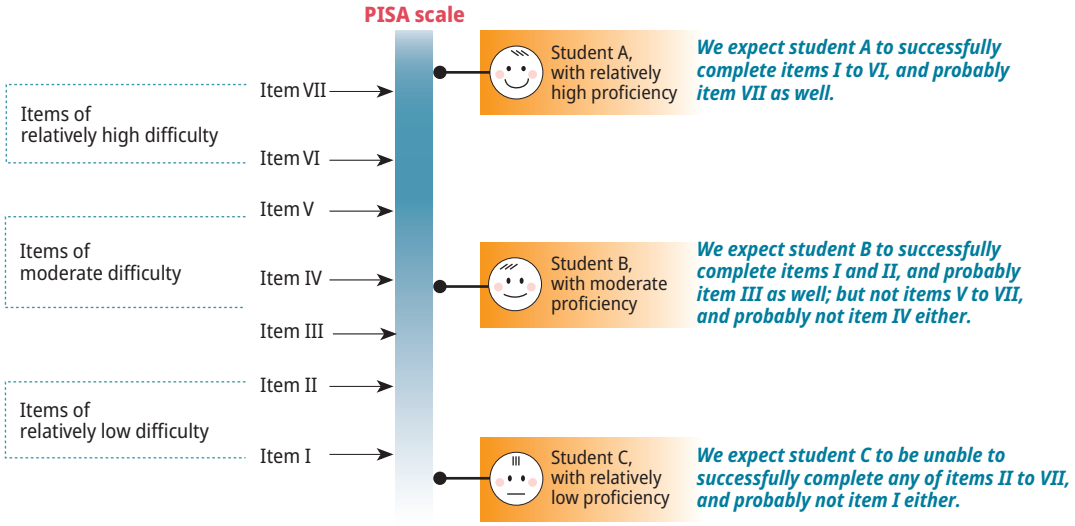
All forms contained an hour-long sequence of reading questions in the first or second part of the two-hour test, with the other hour used to assess one or sometimes two remaining subjects, which were randomly assigned. The exact sequence of test questions in reading was determined by a combination of random assignment and assignment based on performance in the initial stages of the reading assessment (see the section "How does the PISA adaptive test of reading work?" in Chapter 1). In all other subjects, the assignment of questions to students was determined by a single random draw, amongst a predetermined set of item sequences, so that each question was presented to students with equal probability and at different points during the test.

From test questions to PISA scores

PISA reports both the difficulty of questions and the proficiency of test-takers on a single continuous scale (Figure I.2.1), based on item-response theory models (see Annex A1). By showing the difficulty of each question on this scale, it is possible to locate the level of proficiency in the subject that the question demands. By showing the proficiency of test-takers on the same scale, it is possible to describe each test-taker's level of skill or literacy by the type of tasks that he or she can perform correctly most of the time.

Estimates of student proficiency are based on the kinds of tasks students are expected to perform successfully. This means that students are likely to be able to successfully answer questions located at or below the level of difficulty associated with their own position on the scale.[5]

Figure I.2.1 **Relationship between questions and student performance on a scale**



INTERPRETING DIFFERENCES IN PISA SCORES

PISA scores do not have a substantive meaning as they are not physical units, such as metres or grams. Instead, they are set in relation to the variation in results observed across all test participants. There is theoretically no minimum or maximum score in PISA; rather, the results are scaled to fit approximately normal distributions, with means around 500 score points and standard deviations around 100 score points. In statistical terms, a one-point difference on the PISA scale therefore corresponds to an effect size (Cohen's *d*) of 0.01; and a 10-point difference to an effect size of 0.10.

Determining proficiency levels for reporting and interpreting large differences in scores

To help users interpret what students' scores mean in substantive terms, PISA scales are divided into proficiency levels. For example, for PISA 2018, the range of difficulty of reading tasks is represented by eight levels of reading literacy: the simplest tasks in the assessment correspond to Level 1c; Levels 1b, 1a, 2, 3, 4, 5 and 6 correspond to increasingly more difficult tasks.

For each proficiency level identified in this way, descriptions were generated to define the kinds of knowledge and skills needed to complete those tasks successfully. Individuals who are proficient within the range of Level 1c are likely to be able to complete Level 1c tasks, but are unlikely to be able to complete tasks at higher levels. Level 6 includes tasks that pose the greatest challenge in terms of the skills needed to complete them successfully. Students with scores in this range are likely to be able to complete tasks located at this level and all the other tasks in the domain in question (see the following chapters for a detailed description of the proficiency levels in reading, mathematics and science).

Each proficiency level corresponds to a range of about 80 score points. Hence, score-point differences of 80 points can be interpreted as the difference in described skills and knowledge between successive proficiency levels.

Interpreting small differences in scores

Smaller differences in PISA scores cannot be expressed in terms of the difference in skills and knowledge between proficiency levels. However, they can still be compared with each other to conclude, for example, that the gender gap in one country is smaller than the average gender gap across OECD countries, or that the score-point difference between students with and without a tertiary-educated parent is larger than the score-point difference between students with and without an immigrant background.[6] For all differences, but particularly for small differences, it is also important to verify their "statistical significance" (see below).

In order to attach a substantive or practical meaning to differences of less than 80 points, it is tempting to compare them to some benchmark differences of recognised practical significance, expressed in the same units, such as the average achievement gain that children make from one year to the next (Bloom et al., 2008[3]). However, there is considerable uncertainty about how PISA score-point differences translate into a metric such as "years of schooling", and the empirical evidence is limited to a few countries and subjects.

There are, indeed, many difficulties involved in estimating the "typical" progress of a 15-year-old student from one year to the next or from one grade to the next in an international study such as PISA. Just as the quality of education differs across countries, so does the rate at which students progress through their schooling. A single number is unlikely to constitute a common benchmark for all countries. Furthermore, in any particular country, the observed difference between grades may be influenced by the particular grades considered. For example, the difference may depend on whether the student has transitioned from lower secondary to upper secondary school or has remained at the same level of education.

Because the PISA sample is defined by a particular age group, rather than a particular grade, in many countries, students who sit the PISA assessment are distributed across two or more grade levels. Based on this variation, past reports have estimated the average score-point difference across adjacent grades for countries in which a sizeable number of 15-year-olds are enrolled in at least two different grades. These estimates take into account some socio-economic and demographic differences that are also observed across grades. On average across countries, the difference between adjacent grades is about 40 score points. For more information see Table A1.2 in OECD (2013[4]; 2010[5]; 2007[6]).

But comparisons of performance amongst students of the same age across different grades cannot describe how much students gain, in PISA points, over a school year. Indeed, the students who are enrolled below the modal (or most common) grade for 15-year-olds differ in many ways from the students who are the same age but are enrolled in the modal grade for 15-year olds, as do those who are enrolled above the modal grade. Even analyses that account for differences in socio-economic and cultural status, gender and immigrant background can only imperfectly account for differences in motivation, aspirations, engagement and many other intangible factors that influence what students know, the grade in which they are enrolled, and how well they do on the PISA test.

Two types of studies can provide a better measure of the grade equivalence of PISA scores: longitudinal follow-up studies, where the same students who sat the PISA test are re-assessed later in their education, and cross-sectional designs that compare representative samples of students across adjacent age groups and grades.

In Germany, a longitudinal follow-up of the PISA 2003 cohort assessed the same 9th-grade students who participated in PISA one year later, when they were in the 10th grade. The comparisons showed that over this one-year period (which corresponds both to a different age and a different grade) students gained about 25 score points in the PISA mathematics test, on average, and progressed by a similar amount (21 points) in a test of science (Prenzel et al., 2006[7]).

In Canada, the Youth in Transition Study (YITS) followed the first PISA cohort, which sat the PISA 2000 test in reading, over their further study and work career. The most recent data were collected in 2009, when these young adults were 24, and included a re-assessment of their reading score. The mean score in reading amongst 24-year-olds in 2009 was 598 points, compared to a mean score of 541 points for the same young adults when they were 15 years old and in school (OECD, 2012[8]). This shows that students continue to progress in the competencies assessed in PISA beyond age 15. At the same time, it is not possible to know how this progress developed over the years (e.g. whether progress was continuous or whether more progress was made while students were still in secondary school than after they left secondary school). It must also be borne in mind that the PISA test does not measure the more specialised kinds of knowledge and skills that young adults acquire between the ages of 15 and 24.

In France, in 2012, 14-year-old students in 9th grade were assessed as part of a national extension to the PISA sample at the same time as 15-year-old students. The comparison of 14-year-old students in 9th grade (the modal grade for 14-year-old students in France) with students who were enrolled in the general academic track in 10th grade (15-year-old students) shows a 44 score-point

difference in mathematics (Keskpaik and Salles, 2013[9]). This represents an upper bound on the average progression between the 9th and 10th grades in France, because some of the 14-year-olds who were included in the comparison went on to repeat 9th grade or moved to a vocational track in 10th grade, and these were likely to be amongst the lower-performing students in that group.

Because of the limited evidence about differences in PISA scores across school grades, for the same (or otherwise similar) students, and of the variability in these differences that is expected across subjects and countries, this report refrains from expressing PISA score differences in terms of an exact "years-of-schooling" equivalent. It uses the evidence from the cited studies only to establish an order of magnitude amongst differences that are statistically significant.[7]

WHEN IS A DIFFERENCE STATISTICALLY SIGNIFICANT? THREE SOURCES OF UNCERTAINTY IN COMPARISONS OF PISA SCORES

The results of the PISA assessments are estimates because they are obtained from samples of students, rather than from a census of all students, and because they are obtained using a limited set of assessment tasks, not the universe of all possible assessment tasks. A difference is called statistically significant if it is unlikely that such a difference could be observed in the estimates based on samples when, in fact, no true difference exists in the populations from which the samples are drawn.[8]

When students are sampled and assessment tasks are selected with scientific rigour, it is possible to determine the magnitude of the uncertainty associated with the estimate and to represent it as a "confidence interval", i.e. a range so defined that there is only a small probability (typically, less than 5%) for the true value to lie above its upper bound or below its lower bound. The confidence interval needs to be taken into account when making comparisons between estimates, or between an estimate and a particular benchmark value, so that differences that may arise simply due to the sampling of students and items are not interpreted as real differences in the populations. The designs of the PISA test and sample are determined with the aim of reducing, as much as possible, the statistical error associated with country-level statistics and therefore to narrow the confidence interval. Two sources of uncertainty are taken into account:

- *Sampling error:* The aim of a system-level assessment such as PISA is to generalise the results based on samples to the larger target population. The sampling methods used in PISA ensure not only that the samples are representative, and provide a valid estimate of the mean score and distribution of the population, but also that the error due to sampling is minimised, within the given budget and design constraints. The sampling error decreases the greater the number of schools and (to a lesser extent) of students included in the assessment. (In PISA, schools are the primary sampling unit, and students are sampled only from within the schools selected in the first stage of sampling.) The sampling error associated with a country's mean performance estimate is, for most countries, around two to three PISA score points. For the OECD average (which is based on 37 independent national samples) the sampling error is reduced to about 0.4 of a PISA score point.
- *Measurement error* (also called imputation error): No test is perfect or can fully measure proficiency in broad subjects such as reading, mathematics or science. The use of a limited number of items to assess proficiency in these subjects introduces some measurement uncertainty: would the use of a different set of items have resulted in different performance? This uncertainty is quantified in PISA. Amongst other things, it decreases with the number of items in a subject that underlie an estimate of proficiency. It is therefore somewhat larger for the minor subjects in an assessment than for major ones, and it is larger for individual students (who see only a fraction of all test items) than for country means (which are based on all test items). It also decreases with the amount of background information available. For estimates of country means, the imputation error is smaller than the sampling error (around 0.5 of a PISA score point in reading, and 0.8 of a point in mathematics and science).

When comparing results across different PISA assessments, an additional source of uncertainty must be taken into account. Indeed, even if different PISA assessments use the same unit for measuring performance (the metric for reading literacy, for example, was defined in PISA 2000, when reading was, for the first time, the major focus of the PISA test), the test instruments and items change in each assessment, as do the calibration samples and sometimes the statistical models used for scaling results. To make the results directly comparable over time, scales need to be equated. This means that results are transformed so that they can be expressed on the same metric. The *link error* quantifies the uncertainty around the equating of scales.

The link error represents uncertainty around the scale values ("is a score of 432 in PISA 2018 the same 432 as in PISA 2015?") and is therefore independent of the size of the student sample. As a result, it is the same for estimates based on individual countries, on subpopulations or on the OECD average.[9] For comparisons between reading results in PISA 2018 and reading results in past PISA assessments, the link error corresponds to at least 3.5 score points, making it by far the most significant source of uncertainty in trend comparisons. The link error is considerably smaller only for comparisons between PISA 2018 and PISA 2015 mathematics and science results (about 2.3 score points in mathematics and 1.5 point in science). The reduction in the uncertainty around trend comparisons is the result of improvements to the test design (in particular, a greater number of trend

items common to the two assessments) and to the scaling procedure (with the introduction of concurrent calibration) introduced in PISA 2015, and of the absence of framework revisions (the frameworks for assessing mathematics and science remained unchanged from 2015). This reduced uncertainty can explain why a particular score difference may not be considered statistically significant when it is observed between PISA 2018 and PISA 2012, while a score difference of the same magnitude is considered statistically significant when it is observed between PISA 2018 and PISA 2015 (link errors for all possible score comparisons are provided in Annex A7).

Notes

1. Proficiency in reading, mathematics and science is not conceived as an attribute that a student has or does not have; rather, as an attribute that can be acquired to a greater or lesser extent.
2. The PISA 2018 paper-based instruments were based on the PISA 2009 reading framework and the PISA 2006 science framework. Only the mathematics framework was common to both the paper- and computer-based tests in 2018.
3. "Translation" also refers here to the adaptation process; see Chapter 5 in the *PISA 2018 Technical Report* (OECD, forthcoming[1]).
4. In some countries, students with special needs received a one-hour test. This so-called "UH form" consisted of questions from the three domains of reading, mathematics and science.
5. "Unlikely", in this context, refers to a probability below 62% (see Annex A1). The farther above the student's position on the scale a question lies, the lower the probability that the student will answer successfully.
6. Comparisons of score-point differences around similar scale locations should be preferred to comparisons of gaps at different scale locations. Indeed, comparisons of gaps at different scale locations rely on equal-interval properties of the reporting scale (i.e. the idea that the difference between 300 and 350 is, in some sense, the same difference as between 700 and 750) that may not be warranted (Braun and von Davier, 2017[10]; Jacob and Rothstein, 2016[11]).
7. Woessman (2016, p. 6[12]) writes: "As a rule of thumb, learning gains on most national and international tests during one year are equal to between one-quarter and one-third of a standard deviation, which is 25-30 points on the PISA scale". This is, admittedly, a broad generalisation; without taking it too literally, this "rule of thumb" can be used to gain a sense of magnitude for score-point differences.
8. Some small countries/economies actually do conduct a census of schools and, in some cases, of students. Even in these countries/economies, PISA respondents may not coincide with the full, desired target population due to non-response and non-participation.
9. In PISA the link error is assumed to be constant across the scale. For PISA 2018 (as was the case for PISA 2015), link errors are estimated based on the variation in country means across distinct scale calibrations (see Annex A7).

References

Bloom, H. et al. (2008), "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions", *Journal of Research on Educational Effectiveness*, Vol. 1/4, pp. 289-328, http://dx.doi.org/10.1080/19345740802400072. [3]

Braun, H. and **M. von Davier** (2017), "The use of test scores from large-scale assessment surveys: psychometric and statistical considerations", *Large-scale Assessments in Education*, Vol. 5/1, http://dx.doi.org/10.1186/s40536-017-0050-x. [10]

Jacob, B. and **J. Rothstein** (2016), "The Measurement of Student Ability in Modern Assessment Systems", *Journal of Economic Perspectives*, Vol. 30/3, pp. 85-108, http://dx.doi.org/10.1257/jep.30.3.85. [11]

Keskpaik, S. and **F. Salles** (2013), "Les élèves de 15 ans en France selon PISA 2012 en culture mathématique: baisse des performances et augmentation des inégalités depuis 2003", *Note d'information*, Vol. 13/31. [9]

OECD (2019), *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, https://dx.doi.org/10.1787/b25efab8-en. [2]

OECD (2013), *PISA 2012 Results: What Makes Schools Successful (Volume IV): Resources, Policies and Practices*, PISA, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264201156-en. [4]

OECD (2012), *Learning beyond Fifteen: Ten Years after PISA*, PISA, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264172104-en. [8]

OECD (2010), *PISA 2009 Results: What Makes a School Successful?: Resources, Policies and Practices (Volume IV)*, PISA, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264091559-en. [5]

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264040014-en. [6]

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

Prenzel, M. et al. (eds.) (2006), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*, Waxmann Verlag GmbH. [7]

Woessmann, L. (2016), "The Importance of School Systems: Evidence from International Differences in Student Achievement", *Journal of Economic Perspectives*, Vol. 30/3, pp. 3-32, http://dx.doi.org/10.1257/jep.30.3.3. [12]