

ANNEX A7

Comparing reading, mathematics and science performance across PISA cycles

The methodology underpinning the analysis of trends in performance in international studies of education is complex. In order to ensure the comparability of PISA results across different assessment years, a number of conditions must be met.

In particular, successive assessments of the same subject must include a sufficient number of common assessment items, and these items must retain their measurement properties over time, so that results can be reported on a common scale. The set of items included must adequately cover the different aspects of the framework for each domain.

Furthermore, the sample of students in assessments carried out in different years must be equally representative of the target population; only results from samples that meet the strict standards set by PISA can be compared over time. Even though they participated in successive PISA assessments, some countries and economies cannot compare all of their PISA results over time.

Even when PISA samples accurately reflect the target population (that of 15-year-olds enrolled in grade 7 or above), changes in enrolment rates and demographics can affect the interpretation of trends. For this reason, Chapter 9 in this volume also discusses contextual changes alongside trends in performance, and presents adjusted trends that account for changes in the student population in addition to the basic, non-adjusted performance trends.

Comparisons over time can also be affected by changes in assessment conditions or in the methods used to estimate students' performance on the PISA scale. In particular, from 2015 onward, PISA introduced computer-based testing as the main mode of assessment. It also adopted a more flexible model for scaling response data, and treated items that were left unanswered at the end of test forms as if they were not part of the test, rather than as incorrectly answered. (Such items were considered incorrect in previous cycles for the purpose of estimating students' position on the PISA scale.) Instead of re-estimating past results based on new methods, PISA incorporates the uncertainty associated with these changes when computing the significance of trend estimates (see the section on "link errors" below, and Chapter 2).

Finally, comparisons of assessment results through years that correspond to different assessment frameworks may also reflect the shifting emphasis of the test. For example, differences between PISA 2015 (and earlier) and PISA 2018 results in reading, or between PISA 2012 and PISA 2018 results in science reflect not only whether students have become better at mastering the common assessment items used for linking the assessments (which reflect the earlier assessment framework), they also reflect students' relative performance (compared to other students, in other countries) on aspects of proficiency that are emphasised in the most recent assessment framework.

LINK ERRORS

Link errors are estimates that quantify the uncertainty involved in comparisons that involve different calibrations of the same scale (e.g. the PISA 2009 and the PISA 2018 calibrations of the reading scale). Standard errors for estimates of changes in performance and trends across PISA cycles take this uncertainty into account.

As in past cycles, only the uncertainty around the location of scores from past PISA cycles on the 2018 reporting scale is reflected in the link error. Because this uncertainty about the position in the distribution (a change in the intercept) is cancelled out when looking at location-invariant estimates (such as estimates of the variance, the inter-quartile range, gender gaps, regression coefficients, correlation coefficients, etc.), standard errors for these estimates do not include the linking error.

Link error for scores between two PISA assessments


Link errors for PISA 2018 were estimated based on the comparison of rescaled country/economy means per domain with the corresponding means derived from public use files and produced under the original scaling of each cycle. This approach for estimating the link errors was used for the first time in PISA 2015 (OECD, 2017, p. 237_[1]). The number of observations used for the computation of each link error equals the number of countries with results in both cycles. Because of the sparse nature of the data underlying the computation of the link error, a robust estimate of the standard deviation was used, based on the S_n statistic (Rousseeuw and Croux, 1993_[2]).

Table I.A7.1 **Link errors for comparisons between PISA 2018 and previous assessments**

Comparison	Reading	Mathematics	Science
PISA 2000 to 2018	4.04		
PISA 2003 to 2018	7.77	2.80	
PISA 2006 to 2018	5.24	3.18	3.47
PISA 2009 to 2018	3.52	3.54	3.59
PISA 2012 to 2018	3.74	3.34	4.01
PISA 2015 to 2018	3.93	2.33	1.51

Note: Comparisons between PISA 2018 scores and previous assessments can only be made to when the subject first became a major domain or later assessment cycles. As a result, comparisons of mathematics and science performance between PISA 2000 and PISA 2018, for example, are not possible.

Source: *PISA 2018 Technical Report* (OECD, forthcoming_[3]).

StatLink  <https://doi.org/10.1787/888934028957>

Link error for other types of comparisons of student performance

In PISA, link errors for comparisons across two assessments are considered to be the same across the scale: the link error is the same for a scale score of 400 as for a scale score of 600. However, not all quantities of interest are reported on the PISA scale; and some comparisons involve more than two assessments. How is the proportion of students scoring above a particular cut-off value affected by the link error? How are regression-based trends affected by link errors?

The link error for regression-based trends in performance and for comparisons based on non-linear transformations of scale scores can be estimated by simulation, based on the link error for comparison of scores between two PISA assessments. In particular, Table I.A7.2 (available on line) presents the estimates of the link error for the comparison of the percentage of students performing below Level 2 and at or above Level 5, while Table I.A7.3 presents the magnitude of the link error associated with the estimation of the average three-year trend (see below for a definition of the average three-year-trend).

The estimation of the link errors for the percentage of students performing below Level 2 and at or above Level 5 uses the assumption that the magnitude of the uncertainty associated with the linking of scales follows a normal distribution with a mean of 0 and a standard deviation equal to the scale link error shown in Table I.A7.1. From this distribution, 500 errors are drawn and added to the first plausible value of each country's/economy's 2018 students, to represent the 500 possible scenarios in which the only source of differences with respect to 2018 is the uncertainty in the link.

By computing the estimate of interest (such as the percentage of students in a particular proficiency level) for each of the 500 replicates, it is possible to assess how the scale link error influences this estimate. The standard deviation of the 500 replicate estimates is used as the link error for the change in the percentage of students scoring at a particular proficiency level. Because the influence of the scale link error on this estimate depends on the exact shape and density of the performance distribution around the cut-off points, link errors for comparisons of proficiency levels are different for each country, and within countries, for boys and girls.

The estimation of the link errors for regression-based trends similarly uses the assumption that the uncertainty in the link follows a normal distribution with a mean of 0 and a standard deviation equal to the scale link error shown in Table I.A7.1. However, because the interest here lies in trends over more than two assessment years, the covariance between link errors must be considered in addition to the link errors shown in Table I.A7.1.

To simulate data from multiple PISA assessments, 2 000 observations were drawn from a multivariate normal distribution with all means equal to 0 and whose variance/covariance structure is identified by the link error published in Table I.A7.1, and by those between previous PISA reporting scales, published in Table 12.31 of the *PISA 2012 Technical Report* and in Table 12.8 of the *PISA 2015 Technical Report* (OECD, 2014_[4]; OECD, 2017_[1]). These draws represent 2 000 possible scenarios in which the real trend is 0, and the estimated trend entirely reflects the uncertainty in the comparability of scores across scales. Link errors for comparisons of the average three-year trend between PISA 2018 and previous assessments depend on the number of cycles involved in the estimation, but are independent of the shape of the performance distribution within each country.

Comparisons of performance: Difference between two assessments and average three-year trend

To evaluate the evolution of performance, analyses report the change in performance between two cycles and the average three-year trend in performance. When more than five data points are available, curvilinear trend trajectories are also estimated.

Comparisons between two assessments (e.g. a country's/economy's change in performance between PISA 2009 and PISA 2018 or the change in performance of a subgroup) are calculated as:

$$\text{Equation I.A7.1} \quad \Delta_{2018-t} = PISA_{2018} - PISA_t$$

where Δ_{2018-t} is the difference in performance between PISA 2018 and a previous PISA assessment, $PISA_{2018}$ is the mathematics, reading or science score observed in PISA 2018, and $PISA_t$ is the mathematics, reading or science score observed in a previous assessment. (Comparisons are only possible with the year when the subject first became a major domain or later assessments; as a result, comparisons of mathematics performance between PISA 2018 and PISA 2000 are not possible, nor are comparisons of science performance between PISA 2018 and PISA 2000 or PISA 2003.) The standard error of the change in performance ($\sigma(\Delta_{2018-t})$) is:

$$\text{Equation I.A7.2} \quad \sigma(\Delta_{2018-t}) = \sqrt{\sigma_{2018}^2 + \sigma_t^2 + error_{2018,t}^2}$$

where σ_{2018} is the standard error observed for $PISA_{2018}$, σ_t is the standard error observed for $PISA_t$ and $error_{2018,t}^2$ is the link error for comparisons of science, reading or mathematics performance between the PISA 2018 assessment and a previous (t) assessment. The value for $error_{2018,t}^2$ is shown in Table I.A7.1 for most of the comparisons and Table I.A7.2 for comparisons of proficiency levels.

A second set of analyses reported in this volume relates to the average three-year trend in performance. The average three-year trend is the average rate of change observed through a country's/economy's participation in PISA per three-year period – an interval corresponding to the usual interval between two consecutive PISA assessments. Thus, a positive average three-year trend of x points indicates that the country/economy has improved in performance by x points per three-year period since its earliest comparable PISA results. For countries and economies that have participated only in PISA 2015 and PISA 2018, the average three-year trend is equal to the difference between the two assessments.

The average three-year trend in performance is calculated through a regression of the form

$$\text{Equation I.A7.3} \quad PISA_{i,t} = \beta_0 + \beta_1 time_t + \varepsilon_{i,t}$$

where $PISA_{i,t}$ is country i 's location on the science, reading or mathematics scale in year t (mean score or percentile of the score distribution), $time_t$ is a variable measuring time in three-year units, and $\varepsilon_{i,t}$ is an error term indicating the sampling and measurement uncertainty around $PISA_{i,t}$. In the estimation, sampling errors and measurement errors are assumed to be independent across time. Under this specification, the estimate for β_1 indicates the average rate of change per three-year period. Just as a link error is added when drawing comparisons between two PISA assessments, the standard errors for β_1 also include a link error:

$$\text{Equation I.A7.4} \quad \sigma(\beta_1) = \sqrt{\sigma_{s,i}^2(\beta_1) + \sigma_t^2(\beta_1)}$$

where $\sigma_{s,i}(\beta_1)$ is the sampling and imputation error associated with the estimation of β_1 and $\sigma_t^2(\beta_1)$ is the link error associated with the average three-year trend. It is presented in Table I.A7.3.

The average three-year trend is a more robust measure of a country's/economy's progress in education outcomes as it is based on information available from all assessments. It is thus less sensitive to abnormal measurements that may alter comparisons based on only two assessments. The average three-year trend is calculated as the best-fitting line throughout a country's/economy's participation in PISA. PISA scores are regressed on the year the country participated in PISA (measured in three-year units of time). The average three-year trend also takes into account the fact that, for some countries and economies, the period between PISA assessments is less than three years. This is the case for those countries and economies that participated in PISA 2000 or PISA 2009 as part of PISA+. They conducted the assessment in 2001, 2002 or 2010 instead of 2000 or 2009.¹

Curvilinear trends are estimated in a similar way, by fitting a quadratic regression function to the PISA results for country i across assessments indexed by t :

$$\text{Equation I.A7.5} \quad PISA_{i,t} = \beta_2 + \beta_3 year_t + \beta_4 year_t^2 + \varepsilon_{i,t}$$

where $year_t$ is a variable measuring time in years since 2018 and $year_t^2$ is equal to the square of year t . Because $year$ is scaled such that it is equal to zero in 2018, β_3 indicates the estimated annual rate of change in 2018 and β_4 the acceleration/deceleration of the trend. If β_4 is positive, it indicates that the observed trend is U-shaped, and rates of change in performance observed in years closer to 2018 are higher (more positive) than those observed in earlier years. If β_4 is negative, the observed trend has

an inverse-U shape, and rates of change in performance observed in years closer to 2018 are lower (more negative) than those observed in earlier years. Just as a link error is added in the estimation of the standard errors for the average three-year trend, the standard errors for β_3 and β_4 also include a link error (Table I.A7.4). Curvilinear trends are only estimated for countries/economies that can compare their performance across five assessments at least, to avoid over-fitting the data.

ADJUSTED TRENDS

PISA maintains its technical standards over time. Although this means that trends can be calculated over populations defined in a consistent way, the share of the 15-year-old population that this represents, and/or the demographic characteristics of 15-year-old students can also be subject to change, for example because of migration.

Because trend analyses illustrate the pace of progress of successive cohorts of students, in order to draw reliable conclusions from such results, it is important to examine the extent to which they are driven by changes in the coverage rate of the sample and in the demographic characteristics of students included in the sample. Three sets of trend results were therefore developed: unadjusted trends, adjusted trends accounting for changes in enrolment, and adjusted trends accounting for changes in the demographic characteristics of the sample. Adjusted trends represent trends in performance estimated after neutralising the impact of concurrent changes in the demographic characteristics of the sample.

Adjusted trends accounting for changes in enrolment

To neutralise the impact of changes in enrolment rates on trends in median performance and on performance at higher percentiles (or, more precisely, the impact of changes in the coverage rate of the PISA sample with respect to the total population of 15-year-olds; see Coverage Index 3 in Annex A2), the assumption was made that the 15-year-olds not covered by the assessment would all perform below the percentile of interest across all 15-year-olds. With this assumption, the median score across all 15-year-olds (for countries where the coverage rate of the sample is at least 50%) and higher percentiles could be computed without the need to specify the level of performance of the 15-year-olds who were not covered (note that the assumption made is more demanding for the median than for higher percentiles, such as the 75th percentile).

In practice, the estimation of adjusted trends accounting for changes in enrolment first requires that a single case by country/economy be added to the database, representing all 15-year-olds not covered by the PISA sample. The final student weight for this case is computed as the difference between the total population of 15-year-olds (see Table I.A2.2) and the sum of final student weights for the observations included in the sample (the weighted number of participating students). Similarly, each replicate weight for this case is computed as the difference between the total population of 15-year-olds and the sum of the corresponding replicate weights. Any negative weights resulting from this procedure are replaced by 0. A value below any of the plausible values in the PISA sample is entered for the performance variables of this case.

In a second step, the median and upper percentiles of the distribution are computed on the augmented sample. In a few cases where the coverage rate is below 50%, the estimate for the adjusted median is reported as missing.

Adjusted trends accounting for changes in the demographic characteristics of the sample

A re-weighting procedure, analogous to post-stratification, is used to adjust the sample characteristics of past samples to the observed composition of the PISA 2018 sample.

In a first step, the sample included in each assessment cycle is divided into discrete cells, defined by the students' immigrant status (four categories: non-immigrant, first-generation, second-generation, missing), gender (two categories: boy, girl) and relative age (four categories, corresponding to four three-month periods). The few observations included in past PISA datasets with missing gender or age are deleted. This defines, at most, 32 discrete cells for the entire population. However, whenever the number of observations included in one of these 32 cells is less than 10 for a certain country/economy and PISA assessment, the corresponding cell is combined with another, similar cell, according to a sequential algorithm, until all cells reach a minimum sample size of 10.

In a second step, the cells are reweighted so that the sum of final student weights within each cell is constant across assessments, and equal to the sum of final student weights in the PISA 2018 sample. Estimates of the mean and distribution of student performance are then calculated on these reweighted samples, representing the (counterfactual) performance that would have been observed had the samples from previous years had the same composition of the sample in PISA 2018 in terms of the variables used in this re-weighting procedure.

COMPARING THE OECD AVERAGE ACROSS PISA CYCLES

Throughout this report, the OECD average is used as a benchmark. It is calculated as the average across OECD countries, weighting each country equally. Some OECD countries did not participate in certain assessments; other OECD countries do not have comparable results for some assessments; still others did not include certain questions in their questionnaires or changed

them substantially from assessment to assessment. In trend tables and figures, the OECD average is reported on consistent sets of OECD countries, and multiple averages may be included. For instance, the “OECD average-23” includes only 23 OECD countries that have non-missing observations for all assessments since PISA 2000; other averages include only OECD countries that have non-missing observations for the years for which this average itself is non-missing. This restriction allows for valid comparisons of the OECD average over time and neutralises the effect of changing OECD membership and participation in PISA on the estimated trends.

Tables available on line

<https://doi.org/10.1787/888934028957>

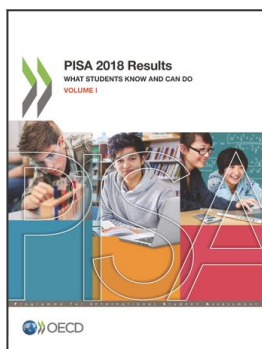
- Table I.A7.2. Link errors for comparisons of proficiency levels between PISA 2018 and previous assessments
- Table I.A7.3. Link errors for the linear trend between previous assessments and PISA 2018
- Table I.A7.4. Link errors for the curvilinear trend between previous assessments and PISA 2018

Notes

1. Countries and economies that participated in the PISA+ projects administered the same assessments as their PISA 2000 or PISA 2009 counterparts, the only difference being that the assessments were conducted one or two years later. These countries/economies' data were adjudicated against the same technical and quality standards as their PISA 2000 and PISA 2009 counterparts. Results from the PISA+ projects appeared originally in OECD/UNESCO Institute for Statistics (2003^[6]) and Walker (2011^[5]), and data from these countries and economies are available as part of the PISA 2000 and PISA 2009 data sets.

References

- OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, <http://www.oecd.org/pisa/data/2015-technical-report/> (accessed on 31 July 2017). [1]
- OECD (2014), *PISA 2012 Technical Report*, OECD Publishing, Paris, <http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf> (accessed on 18 September 2019). [4]
- OECD (forthcoming), *PISA 2018 Technical Report*, OECD Publishing, Paris. [3]
- OECD/UNESCO Institute for Statistics (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*, OECD Publishing, <http://dx.doi.org/10.1787/9789264102873-en>. [6]
- Rousseeuw, P. and C. Croux (1993), “Alternatives to the Median Absolute Deviation”, *Journal of the American Statistical Association*, Vol. 88/424, pp. 1273–83, <http://dx.doi.org/10.1080/01621459.1993.10476408>. [2]
- Walker, M. (2011), *PISA 2009 Plus Results: Performance of 15-year-olds in reading, mathematics and science for 10 additional participants*, ACER Press. [5]



From:
PISA 2018 Results (Volume I)
What Students Know and Can Do

Access the complete publication at:
<https://doi.org/10.1787/5f07c754-en>

Please cite this chapter as:

OECD (2019), “Comparing reading, mathematics and science performance across PISA cycles”, in *PISA 2018 Results (Volume I): What Students Know and Can Do*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/eb6c0071-en>

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at <http://www.oecd.org/termsandconditions>.