Annex A7. Comparing mathematics, reading and science performance across PISA assessments

The methodology underpinning the analysis of trends in performance in international studies of education is complex. 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 different assessment cycles must be similarly representative of the target population; only results from samples that meet the strict standards set by PISA can be compared over time. Even though some countries and economies took part in successive PISA assessments, some of them cannot compare all their PISA results over time.

Comparisons over time can 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 form 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 assessments 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 statistical significance of trend estimates (see the section on "link errors" below).

Changes in enrolment rates do not affect the representative nature of the PISA sample with regards to its target population (15-year-olds enrolled in Grade 7 or above), nevertheless, such changes may affect the interpretation of trends.

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 2018 (and earlier) and PISA 2022 results in mathematics 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 2012 and the PISA 2022 calibrations of the mathematics scale). Standard errors for estimates of changes in performance and trends across PISA assessments take this uncertainty into account.

Similarly to past assessments, only the uncertainty around the location of scores from past PISA assessments on the 2022 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 2022 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 assessment. This approach for estimating the link errors was used for the first time in PISA 2015 (OECD, 2017[1]). The number of observations used for the computation of each link error equals the number of countries with results in both assessments. 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. Robust link error for comparisons of performance between PISA 2022 and previous assessments

Comparison	Reading	Mathematics	Science
PISA 2000 to 2022	6.67		
PISA 2003 to 2022	5.25	5.54	
PISA 2006 to 2022	8.56	4.09	3.68
PISA 2009 to 2022	4.66	4.28	5.92
PISA 2012 to 2022	6.01	3.58	5.20
PISA 2015 to 2022	3.63	2.74	1.38
PISA 2018 to 2022	1.47	2.24	1.61

Note: Comparisons between PISA 2022 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 2022, for example, are not possible.
Source: PISA 2022 Technical Report *(OECD, forthcoming)*

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?

Link error for regression-based trends in performance

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 presents the magnitude of the link error associated with the estimation of the average decennial trend (see below for a definition of the average decennial trend).

The estimation of the link errors for regression-based trends 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, in Table 12.8 of the PISA 2015 Technical Report and Table 12.8 of the PISA 2018 Technical Report (OECD, 2014[3]; OECD, 2017[1]; OECD, 2020[4]). 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 decennial trend between PISA 2022 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.

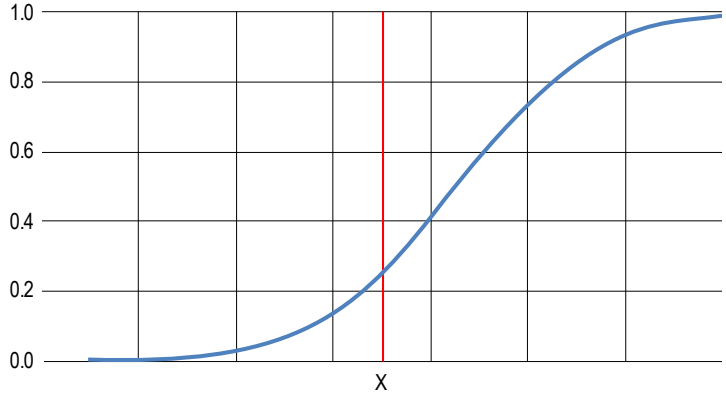
Link error for non-linear transformations of scores

While in previous assessments the link error for comparisons based on non-linear transformations of scores (i.e. proficiency levels) were estimated by simulation of the link error used to compare two PISA assessments, in PISA 2022 the link error is estimated using a parametric approximation of the distribution of student proficiency (the normal distribution), together with the "delta method".

The computation of the link errors using the delta method can be illustrated by taking the percentage of students below proficiency Level 2 as the variable of interest. However, this method applies to any generic non-linear transformation of PISA scores.

In this illustration, the variable of interest is a value in a cumulative normal distribution (Figure I.A7.1). Values on the PISA scale (including the scale link error) are placed on the x-axis; the "proportion below" a particular value on the PISA scale (X) can be read on the y-axis (about .25, or 25%, in this example); and scale errors will be translated to errors on the y-axis as a function of the slope of the curve around the value of X. As the figure makes clear, the link error on the x-axis will affect the error on the y-axis differently, depending where the value of interest (X) is located on the x-axis. In regions where the slope is steeper, an error on the x-axis will translate into a larger error on the y-axis; where the slope flattens out (at the far tails of the distribution), an error on the x-axis will translate to a small error on the y-axis.

Figure I.A7.1. Normal cumulative distribution function



By assuming that the distribution of PISA scores is approximately normal, it is possible to compute the "slope" factor which affects the translation of link errors from PISA scale to percentage scale used for reporting values of the cumulative distribution (e.g. the "percentage of students below proficiency Level 2").

Comparisons of performance: Difference between two assessments and average decennial trend

To evaluate how performance evolved over time, analyses report the change in performance between two assessment cycles and the average decennial trend in performance. When at least 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 2022 or the change in performance of a subgroup) are calculated as:

$$\Delta_{2022-t} = PISA_{2022} - PISA_t \qquad \text{Equation I.A7.1}$$

where Δ_{2022-t} is the difference in performance between PISA 2022 and a previous PISA assessment, $PISA_{2022}$ is the mathematics, reading or science score observed in PISA 2022, 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 2022 and PISA 2000 are not possible, nor are comparisons of science performance between PISA 2022 and PISA 2000 or PISA 2003).

The standard error of the change in performance $\sigma(\Delta_{2022-t})$ is:

$$\sigma(\Delta_{2022-t}) = \sqrt{\sigma_{2022}^2 + \sigma_t^2 + error_{2022,t}^2} \qquad \text{Equation I.A7.2}$$

where σ_{2022} is the standard error observed for $PISA_{2022}$, σ_t is the standard error observed for $PISA_t$ and $error_{2022,t}^2$ is the link error for comparisons of mathematics, reading or science performance between the PISA 2022 assessment and a previous (t) assessment. The value for $error_{2022,t}^2$ is shown in Table I.A7.1.

A second set of analyses reported in this volume relates to the average decennial trend in performance. The average decennial trend is the average rate of change observed through a country's/economy's participation in PISA per 10-year period. Thus, a positive average decennial trend of x points indicates that the country/economy has improved in performance by x points per 10-year period since its earliest comparable PISA results. The average decennial trend in performance is calculated through a regression of the form:

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

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 10-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 10-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:

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

where $\sigma_{s,i}^2(\beta_1)$ is the sampling and imputation error associated with the estimation of β_1 and $\sigma_l^2(\beta_1)$ is the link error associated with the average 10-year trend. It is presented in Table I.A7.2.

The average 10-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 10-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 10-year units of time).

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*:

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

where $year_t$ is a variable measuring time in years since 2022 and $year_t^2$ is equal to the square of year t. Because year is scaled such that it is equal to zero in 2022, β_3 indicates the estimated annual rate of change in 2022 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 2022 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 2022 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 10-year trend, the standard errors for β_3 and β_4 also include a

link error (Table I.A7.3). Curvilinear trends are only estimated for countries/economies that can compare their performance across at least five assessments 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 can also be subject to change.

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. Two sets of trend results were therefore developed: unadjusted trends and adjusted trends accounting for changes in enrolment.

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.A2.1) 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.

Comparing the OECD average across PISA assessments

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-35” includes only 35 OECD countries that have non-missing observations for all assessments since PISA 2012; 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.

Table I.A7.4. Tables comparing performance across PISA assessments

	Table I.A7.1.	Link errors for comparisons between PISA 2022 and previous assessments
WEB	Table I.A7.2.	Link errors for the linear trend between previous assessments and PISA 2022
WEB	Table I.A7.3.	Link errors for the curvilinear trend between previous assessments and PISA 2022

StatLink https://stat.link/48f0zo

References

OECD (2020), *PISA 2018 Technical Report*, OECD Publishing, Paris. [4]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris. [1]

OECD (2014), *PISA 2012 Technical Report*, OECD Publishing, Paris. [3]

Rousseeuw, P. and C. Croux (1993), "Alternatives to the Median Absolute Deviation", *Journal of the American Statistical Association*, Vol. 88/424, pp. 1273-1283, https://doi.org/10.1080/01621459.1993.10476408. [2]