ANNEX A5
CHANGES IN THE ADMINISTRATION AND SCALING OF PISA 2015 AND IMPLICATIONS FOR TRENDS ANALYSES

Comparing science, reading and mathematics performance across PISA cycles

The PISA 2006, 2009, 2012 and 2015 assessments use the same science performance scale, which means that score points on this scale are directly comparable over time. The same is true for the reading performance scale used since PISA 2000 and the mathematics performance scale used since PISA 2003. Comparisons of scores across time are possible because some items are common across assessments and because an equating procedure aligns performance scales that are derived from different calibrations of item parameters to each other.

All estimates of statistical quantities are associated with statistical uncertainty, and this is also true for the transformation parameters used to equate PISA scales over time. A link error that reflects this uncertainty is included in the estimate of the standard error for estimates of PISA performance trends and changes over time. (For more details concerning link errors, see the sections below.)

The uncertainty in equating scales is the product of changes in the way the test is administered (e.g. differences related to the test design) and scaled (e.g. differences related to the calibration samples) across the years. It also reflects the evolving nature of assessment frameworks. PISA revisits the framework for science, reading and mathematics every nine years, according to a rotating schedule, in order to capture the most recent understanding of what knowledge and skills are important for 15-year-olds to acquire in order to participate fully in tomorrow's societies.

Changes in test administration and design can influence somewhat how students respond to test items. Changes in samples and the models used for the scaling produce different estimates of item difficulty. As a consequence, there is some uncertainty when results from one cycle are reported on the scale based on a previous cycle. All cycles of PISA prior to 2015, for instance, differed from each other in the following three ways:

- *The assessment design.*[1] The assessment design can influence how students respond in several ways. For example, students might not perceive the same reading item as equally difficult when it is presented at the beginning of a test, as was mostly the case in PISA 2000, as when it is presented across different places in the test, as was the case in later assessments. Similarly, students may not invest the same effort when the item is part of a 30-minute "reading" sequence in the middle of a mathematics and science test, as was mostly the case when reading was the minor domain in 2003, 2006 and 2012, compared to when reading is the major domain. In PISA, these effects are unsystematic and are typically small, but they are part of the uncertainty in the estimates.

- *The calibration samples.* In PISA cycles prior to 2015, item difficulty was estimated using only the responses of students who participated in the most recent assessment. In PISA 2009 and PISA 2012, the calibration sample was a random subset of 500 students per country/economy. In PISA 2000, 2003 and 2006, the calibration sample included only students from OECD countries (500

per country) (OECD, 2009). This implies that each trend item had as many (independent) estimates of item difficulty as there were cycles in which it was used. These estimates were not identical, and the variability among these estimated item difficulties contributes to the uncertainty of comparisons over PISA cycles. The use of only a subsample of the PISA student data per country further increases this uncertainty, and was justified by the limited computational power available at the time of early PISA cycles.

- *The set and the number of items common to previous assessments.* Just as the uncertainty around country mean performance and item parameters is reduced by including more schools and students in the sample, so the uncertainty around the link between scales is reduced by retaining more items included in previous assessments for the purpose building this link. For the major domain (e.g. science in 2015), the items that are common to prior assessments are a subset of the total number of items that make up the assessment because PISA progressively renews its pool of items in order to reflect the most recent frameworks. The frameworks are based on the current understanding of the reading, mathematics and science competencies that are required of 15-year-olds to be able to thrive in society.

PISA 2015 introduced several improvements in the test design and scaling procedure aimed at reducing the three sources of uncertainty highlighted above. In particular, the assessment design for PISA 2015 reduced or eliminated the difference in construct coverage across domains and students' perception of certain domains as "major" or "minor". In the most frequently implemented version of the test (the computer-based version in countries that assessed collaborative problem solving), for example, 86% of students were tested in two domains only, for one hour each (33% in science and reading, 33% in science and mathematics, and 22% in science and collaborative problem solving, with the order inversed for half of each group) (see OECD [forthcoming] for details). The number of items that are common to previous assessments was also greatly increased for all domains, and most obviously for minor domains. For example, when reading was a minor domain (in 2003 and 2006), only a number of items equivalent to one hour of testing time, or two 30-minute clusters, was used to support the link with PISA 2000; when mathematics was the major domain for the second time in 2012, the number of items linking back to 2003 was equivalent to one-and-a-half hours of testing time. In 2015, science (the major domain), reading and mathematics all use the equivalent of three hours of testing time to support the link with existing scales.

The scaling procedure was also improved by forming the calibration sample based on all student responses from the past four cycles of the assessment. This includes, for all domains, one assessment in which it was the major domain; for the major domain, the sample goes back to the previous cycle in which the domain was major. For the next PISA cycle (2018) the calibration sample will overlap by up to about 75% with the 2015 cycle. As a consequence, the uncertainty due to the re-estimation of item parameters in scaling will be reduced considerably compared to cycles up to 2012.

While these improvements can be expected to result in reductions in the link error between 2015 and future cycles, they may add to the uncertainty reflected in link errors between 2015 and past cycles, because past cycles had a different test design and followed a different scaling procedure.

In addition, PISA 2015 introduced further changes in test administration and scaling:

- Change in the assessment mode. Computer-based delivery became the main mode of administration of the PISA test in 2015. All trend items used in PISA 2015 were adapted for delivery on computer. The equivalence between the paper- and computer-based versions of trend items used to measure student proficiency in science, reading and mathematics was assessed on a diverse population of students from all countries/economies that participated in the PISA 2015

assessment as part of in an extensive field trial, conducted in all countries/economies that participated in the PISA 2015 assessment. The results of this mode-effect study, concerning the level of equivalence achieved by items ("scalar" equivalence or "metric" equivalence; see e.g. Meredith, 1993; Davidov, Schmidt, and Billiet, 2011) informed the scaling of student responses in the main study. Parameters of scalar- and metric-invariant items were constrained to be the same for the entire calibration sample, including respondents who took them in paper- and computer-based mode (see the section on "Comparing PISA results across paper- and computer-based administrations" for further details).

- Change in the scaling model. A more flexible statistical model was fitted to student responses when scaling item parameters. This model, whose broadest form is the generalised partial credit model (i.e. a two-parameter item-response-theory model; see Birnbaum 1968; Muraki 1992), includes constraints for trend items so as to retain as many trend items with one-parameter likelihood functions as supported by the data, and is therefore referred to as a "hybrid" model. The one-parameter models on which scaling was based in previous cycles (Rasch 1960; Masters 1982) are a special case of the current model. The main difference between the current hybrid model and previously used one-parameter models is that the hybrid model does not give equal weight to all items when constructing a score, but rather assigns optimal weights to tasks based on their capacity to distinguish between high- and low-ability students. It can therefore better accommodate the diversity of response formats included in PISA tests.

- Change in the treatment of differential item functioning across countries. In tests such as PISA, where items are translated into multiple languages, some items in some countries may function differently from how the item functions in the majority of countries. For example, terms that are harder to translate into a specific language are not always avoidable. The resulting item-by-country interactions are a potential threat to validity. In past cycles, common item parameters were used for all countries, except for a very small number of items that were considered "dodgy" and therefore treated as "not administered" for some countries (typically, less than a handful of items, for instance if careless errors in translation or printing were found only late in the process). In 2015, the calibration allowed for a (limited) number of country-by-cycle-specific deviations from the international item parameters (Glas and Jehangir, 2014; Oliveri and von Davier, 2011; Oliveri and von Davier, 2014).[2] This approach preserves the comparability of PISA scores across countries and time, which is ensured by the existence of a sufficient number of invariant items, while reducing the (limited) dependency of country rankings on the selection of items included in the assessment, and thus increasing fairness. The Technical Report for PISA 2015 provides the number of unique parameters for each country/economy participating in PISA (OECD, forthcoming).

- Change in the treatment of non-reached items. Finally, in PISA 2015, non-reached items (i.e. unanswered items at the end of test booklets) were treated as not administered, whereas in previous PISA cycles they were considered as wrong answers when estimating student proficiency (i.e. in the "scoring" step) but as not administered when estimating item parameters (in the "scaling" step). This change makes the treatment of student responses consistent across the estimation of item parameters and student proficiency, and eliminates potential advantages for countries and test takers who randomly guess answers to multiple-choice questions that they could not complete in time compared to test takers who leave these non-reached items unanswered.[3] However, this new treatment of non-reached items might result in higher scores than would have been estimated in the past for countries with many unanswered items.

Linking PISA 2015 results to the existing reporting scales

This section describes how PISA 2015 results were transformed in order to report the results of PISA 2015 on the existing PISA scales (the reading scale defined in PISA 2000, the mathematics scale defined in PISA 2003, and the science scale defined in PISA 2006).

In the estimation of item parameters for 2015, based on student responses from the 2006, 2009, 2012 and 2015 cycles, these responses were assumed to come from M distinct populations, where M is the total number of countries/economies that participated in PISA multiplied by the number of cycles in which they participated (multigroup model). Each population m_{ij} (where i identifies the country, and j the cycle) is characterised by a certain mean and variation in proficiency.[4] The proficiency means and standard deviations were part of the parameters estimated by the scaling model together with item parameters. (As in previous cycles, individual estimates of proficiency were only imputed in a second step, performed separately for each country/economy. This "scoring" step was required and completed only for the 2015 cycle). The result of the scaling step is a linked scale, based on the assumption of invariance of item functions across the 2006, 2009, 2012, and 2015 cycles, in which the means and standard deviations of countries are directly comparable across time.

To align the scale established in the scaling step with the existing numerical scale used for reporting PISA results from prior cycles, a linear transformation was applied to the results. The intercept and slope parameters for this transformation were defined by comparing the country/economy means and standard deviations, estimated during the scaling step in the logit scale, to the corresponding means and standard deviations in the PISA scale, obtained in past cycles and published in PISA reports. Specifically, the transformation for science was based on the comparison of the OECD average mean score and (within-country) standard deviation to the OECD average mean score and (within-country) standard deviation in 2006. This transformation preserves the meaning of the PISA scale as "having a mean of 500 and a standard deviation of 100, across OECD countries, the first time a domain is the major domain". A similar procedure was used for mathematics (matching average means and standard deviations for OECD countries to the last cycle in which it was the major domain, i.e. 2012) and reading (matching re-estimated results to the 2009 reported results).

Assessing the impact on trends of changes in the scaling approach introduced in 2015

It is possible to estimate what the past country means would have been if the current approach to scaling student responses were applied to past cycles. This section reports on the comparison between the means published in past PISA reports (e.g. OECD, 2014b) and the country/economy means obtained from the 2015 scaling step.

Table A5.1 shows the correlations between two sets of country means for 2006, 2009, 2012 and 2015: those reported in the tables included in Annex B and discussed throughout this report, and the mean estimates, based on the same data, but produced, under the 2015 scaling approach, as a result of the multiple group model described above. The differences in the means may result from the use of larger calibration samples that pool data from multiple cycles; from the new treatment of differential item functioning across countries and of non-reached items; or from the use of a hybrid item-response-theory model in lieu of the one-parameter models used in past cycles. The column referring to 2015 illustrates the magnitude of differences due to the imputation of scores during the scoring step, which is negligible.

The high correlations reported in this table for the years 2006, 2009 and 2012 (all higher than 0.993, with the exception of reading in 2006, for which the correlation is 0.985) indicate that the relative position of countries on the PISA scale is hardly affected by the changes introduced in 2015 in the scaling approach. The magnitude of these correlations across estimates derived under different methodologies is also larger

than the magnitude of correlations of mean scores across consecutive PISA assessments, and much larger than the magnitude of correlations of mean scores between two major cycles for the same domain (at intervals of nine years).[5] This means that changes in methodology can, at best, account for only a small part of the changes and trends reported in PISA.

Table A5.1. Correlation of country means under alternative scaling approaches

	2006	2009	2012	2015
Science	0.9941	0.9961	0.9966	0.9997
Reading	0.9850	0.9949	0.9934	0.9992
Mathematics	0.9953	0.9974	0.9973	0.9995

Comparing country means under a consistent scaling approach

Once the country means produced during the scaling of item parameters are transformed in the way described in the previous section, they can be used to assess, for each country, the sensitivity of the trends reported in the main text and in tables included in Annex B to changes in the scaling approach and in the calibration samples introduced in 2015.[6] These transformed means are reported in Table A5.3 for science, Table A5.4 for reading and Table A5.5 for mathematics.

For a large majority of countries/economies, the differences between the mean scores reported in Annex B and the mean scores reported in Tables A5.3, A5.4 and A5.5 are well within the confidence interval associated with the link error (see below). However, there are some noteworthy exceptions (Figures A5.1, A5.2 and A5.3). In particular, when focusing on changes between 2015 and the last time a domain was major, the following observations emerge:

- Science
  - The improvement in mean science performance reported for Colombia is almost entirely due to changes in the approach to scaling. The increase in mean score would have been only three points (not significant) had the 2015 approach and calibration sample been used to scale 2006 results. To a lesser extent, the non-significant increases in mean scores reported for Chile, Brazil, Indonesia and Uruguay are also due to the changes in the calibration sample and in the approach to scaling. These four countries would have had less positive trends (but most likely, still not significant) had the past mean scores been reported based on the PISA 2015 scaling approach. It is not possible to identify with certainty which differences between the original scaling of PISA 2006 data and the PISA 2015 re-scaling produced these results. However, a likely cause for these differences is the new treatment of non-reached items. In all these countries, many students did not reach the items placed at the end of the test booklets or forms.
  - The United States shows a non-significant improvement (of seven score points) in science between 2006 and 2015. The improvement would have been somewhat larger, and most likely reported as significant (+15 points), had the 2015 approach and calibration sample been used to scale 2006 results. While larger than the reported change, the change observed under the 2015 scaling approach is nevertheless included in the confidence interval for the reported change.

- Reading

  - The negative change between PISA 2009 and PISA 2015 reported for Korea (-22 score points) is, to a large extent, due to the difference in the scaling approach. Had the PISA 2009 results for reading been scaled with the PISA 2015 calibration sample and the PISA 2015 approach to scaling, the difference in results for Korea would have been only -9 points, and most likely would not have been reported as significant. According to the PISA 2015 scaling model, past results in reading for Korea are somewhat over-reported. It is not possible to identify with certainty, from these results, which aspect of the PISA 2015 approach is responsible for the difference. However, a likely cause is the new treatment of differential item functioning. Indeed, most items exhibiting a moderate level of differential item functioning for Korea, and thus receiving country-specific parameters in the PISA 2015 calibration, are items in which the success of students in Korea in past PISA cycles was greater than predicted by the international parameters. To a lesser extent, Thailand shows a similar pattern. The reported negative change (-12 points) would have been reported as not significant (-3 points), had the comparison be made with rescaled 2009 results.

  - Denmark shows a non-significant improvement (of five points) between PISA 2009 and PISA 2015. However, under the PISA 2015 approach, the improvement would have been 15 points, and most likely be reported as significant.

  - Estonia shows a significant improvement of 18 points, but the improvement would have been of only 10 points had the PISA 2009 results been derived using the PISA 2015 scaling model.

  - The Netherlands shows a non-significant deterioration (of five points) between PISA 2009 and PISA 2015. However, under the PISA 2015 approach, the Netherlands would have seen an increase by 4 points (most likely not significant).

  - The improvement in mean reading performance reported for Colombia, Trinidad and Tobago and Uruguay is most likely due to changes in the approach to scaling. The change in mean score would have been close to 0 (and reported as not significant) had the 2015 approach and calibration sample been used to scale 2009 results. Similarly, the increase in the mean score for Peru and Moldova would have only been 15 points and 21 points, respectively (compared to a reported increase of 28 points), under a constant scaling approach. A likely cause for these differences is the new treatment of non-reached items. In all these countries, many students did not reach the items placed at the end of the test booklets or forms.

- Mathematics

  - The negative changes between PISA 2012 and PISA 2015 reported for Chinese Taipei (-18 score points) and Viet Nam (-17 score points) are, to a large extent, due to the use of a different scaling approach. Had the PISA 2012 results for mathematics been scaled with the PISA 2015 calibration sample and the PISA 2015 approach to scaling, the differences in results for Chinese Taipei and Viet Nam would have been only -3 points and -4 points, respectively, and most likely would not have been reported as significant. The new treatment of differential item functioning may be the main reason for these differences.

  - The reported change for Turkey between PISA 2012 and PISA 2015 (-28 score points) would have been only -18 score points had all results been generated under the 2015 scaling

approach. While the reported trend amplifies the magnitude of the change, the direction and the significance of the change are similar under the two sets of results.

- The increase in the mathematics mean score for Albania between PISA 2012 and PISA 2015 (+19 score points) would have been smaller and most likely be reported as not significant (+7 points) had all results been generated under a consistent scaling approach. A likely cause for this difference is the new treatment of non-reached items. Similarly, the non-significant increase reported for Uruguay (+9 points) would have been even closer to zero (+1 point) under a consistent scaling approach.

- Singapore shows a deterioration of mean performance of 9 points, which, given the reduced sampling error for this country, is reported as significant. Had the PISA 2012 results been derived using the PISA 2015 scaling model, however, they would have been seven points below the published results; as a result, the difference from PISA 2015 results under a consistent scaling approach would have been of only -2 points.

All other differences between reported changes and changes based on applying the PISA 2015 approach to scaling to past PISA assessments are smaller than the differences expected given the linking errors provided in the following sections of this annex.

Figure A5.1. Changes in science performance between 2006 and 2015, based on originally scaled and on rescaled results

Figure A5.2. Changes in reading performance between 2009 and 2015, based on originally scaled and on rescaled results

Figure A5.3. Changes in mathematics performance between 2012 and 2015, based on originally scaled and on rescaled results

Comparing PISA results across paper- and computer-based administrations

The equivalence of link items, assessed at the international level, was established in the extensive mode-effect study that was part of the field trial for PISA 2015. These results provide strong support for the assertion that results can be reported on the same scale across modes. In addition, the possibility of country-by-cycle-specific parameters can, to some extent, account for national deviations from the international norm.

The equivalence of link items was first assessed during the field trial (in 2014) on equivalent populations created by random assignment within schools. More than 40 000 students from the countries and economies that were planning to conduct the PISA 2015 assessment on computers were randomly allocated to the computer- or paper-based mode within each school, so that the distribution of student ability was comparable across the two modes. As a result, it was possible to attribute any differences across modes in students' response patterns, particularly differences that exceeded what could be expected due to random variations alone, to an impact of mode of delivery on the item rather than to students' ability to use the mode of delivery. The field trial was designed to examine mode effects at the international level, but not for each national sample or for sub-samples with a country.

The mode-effects study asked two main questions:

- Do the items developed in prior PISA cycles for delivery in paper-based mode measure the same skills when delivered on computer? For instance, do all the science items that were adapted for

computer delivery measure science skills only, or do they measure a mixture of science and computer skills?

- Is the difficulty of the paper-based versions of these items the same as that of computer-based versions?

Only if an item measured the same skills and was equally difficult across the two modes was it considered to be fully equivalent (i.e. scalar invariant) and to support meaningful comparisons of performance across modes. This analysis of test equivalence was based on pooled data from all countries/economies using explanatory item-response-theory (IRT) models. In these models, two distinct sets of parameters estimate how informative student responses are about proficiency on the intended scale, and what level of proficiency they indicate. The analysis identified three groups of items:

- Group 1: Items that had the same estimated difficulty and discrimination parameters in both modes and were therefore found to be fully equivalent on paper and computer (*scalar invariance*).

- Group 2: Items that had the same discrimination parameter but distinct difficulty parameter (*metric invariance*). Success on these items did say something about proficiency in the domain, in general; but the difficulty of items varied depending on the mode, often because of interface issues, such as answer formats that required free-hand drawing or the construction of equations. Several items proved to be more difficult on computers, and a few items were easier on computers.

- Group 3: Items for which field trial estimates indicated that they measured different skills, depending on the mode (no *metric invariance*).

Items in Group 3 were not used in the computer-based test in the main study (two items in mathematics were used in the paper-based test only). Items from Group 1 and 2 were used, and the stability of item parameters across cycles and mode was further probed during scaling operations for the main study. In the end, the data supported the full (scalar) equivalence across modes for up to 61 items in science, 51 items in mathematics and 65 items in reading.[7] These items function as anchor items or link items for scaling purposes and are the basis for comparisons of performance across modes and across time. For the remaining trend items included in the PISA 2015 main study (24 in science, 38 in reading and 30 in mathematics), metric equivalence was confirmed, but each of these items received a mode-specific difficulty parameter. When comparing students who sat the PISA test in different modes, this subset of metric-invariant items only provides information about the ranking of students' proficiencies within a given mode (and therefore contributes to the measurement precision), but does not provide information to rank students and countries across different modes. Items that reached scalar equivalence have identical item parameters for PBA (paper-based assessment) and CBA (computer-based assessment) in Tables C2.1, C2.3 and C2.4; items that only reached metric equivalence have the same slope parameters, but different difficulty parameters.

The full equivalence of link items across modes, assessed on a population representing all students participating in PISA who took the test on computers, ensures that results can be compared across paper- and computer-based modes, and that the link between these sets of results is solid. It implies, among other things, that if all students who took the PISA 2015 test on computer had taken the same test on paper, their mean score, as well as the proportion of students at the different levels of proficiency, would not have been significantly different.

Annex A6 provides further information on the exploratory analysis of mode-by-group interactions that was carried out on field trial data. While the results of this analysis, in particular with respect to mode-by-gender interactions, are encouraging, the limitations of field-trial data for this type of exercise must be borne in mind when interpreting results.

Assessing the comparability of new science items and trend items

New science items were developed for PISA 2015 to reflect changes in the PISA framework for assessing science and in the main mode of delivery. Framework revisions that coincide with the development of new items occur periodically in PISA: the reading framework was revised in 2009, and the mathematics framework in 2012. The development of new items in science was guided by the need to provide balanced coverage of all framework aspects, particularly aspects that were refined or given greater emphasis in the PISA 2015 framework compared with the PISA 2006 framework. These include the distinction between epistemic and procedural knowledge, which was only implicit in the prior framework, and the more active component of science literacy. The latter is reflected in the new way science literacy is organised around the competencies to "evaluate and design scientific enquiry" and to "interpret data and evidence scientifically" (along with "explain phenomena scientifically"). These competencies are related to, but clearly do not overlap perfectly with, what was previously described as "identifying scientific issues" and "using scientific evidence".

After the 2015 main study, the possibility of reporting results on the existing science scale, established in 2006, was tested through an assessment of dimensionality. When new and existing science items were treated as related to distinct latent dimensions, the median correlation (across countries/language groups) between these dimensions was 0.92, a relatively high value (similar to the correlation observed among subscales from a same domain). Model-fit statistics confirmed that a unidimensional model fits the data better than a two-dimensional model, supporting the conclusion that new and existing science items form a coherent unidimensional scale with good reliability. Further details on scaling outcomes can be found in the *PISA 2015 Technical Report* (OECD, forthcoming).

Quantifying the uncertainty of scale comparability in the link error

Standard errors for estimates of changes in performance and trends across PISA cycles take into account the uncertainty introduced by the linking of scales produced under separate calibrations. These more conservative standard errors (larger than standard errors that were estimated before the introduction of the linking error) reflect not only the measurement precision and sampling variation as for the usual PISA results, but also the linking error provided in Table A5.2. For PISA 2015, the linking error reflects not only the uncertainty due to the selection of link items, but also the uncertainty due to the changes in the scaling methodology introduced in 2015.

As in past cycles, only the uncertainty around the location of scores from past PISA cycles on the 2015 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 2015 were estimated based on the comparison of rescaled country/economy means per domain (e.g. those reported in Tables A5.3, A5.4 and A5.5) with the corresponding means derived from public use files and produced under the original scaling of each cycle. This new approach for estimating the link errors was used for the first time in PISA 2015. 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).

Table A5.2. Link errors for comparisons between PISA 2015 and previous assessments

Link error for other types of comparisons of student 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 A5.6 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 A5.7 presents the magnitude of the link error associated with the estimation 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 A5.2. From this distribution, 500 errors are drawn and added to the first plausible value of each country's/economy's 2015 students, to represent the 500 possible scenarios in which the only source of differences with respect to 2015 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 in 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 A5.2. 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 A5.2. To simulate data from multiple PISA assessments, 2000 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 A5.2 as well as by those between previous PISA reporting scales, published in Table 12.31 of the *PISA 2012 Technical Report* (OECD, 2014a). These draws represent 2000 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 2015 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. For reading, where up to six data points are available, curvilinear trend trajectories are also estimated.

1.

Comparisons between two assessments (e.g. a country's/economy's change in performance between PISA 2006 and PISA 2015 or the change in performance of a subgroup) are calculated as:

$$\Delta_{2015-t} = PISA_{2015} - PISA_t$$

where $\Delta_{2015\text{-}t}$ is the difference in performance between PISA 2015 and a previous PISA assessment (comparisons are only possible to when the subject first became a major domain or later assessment cycles; as a result, comparisons of mathematics performance between PISA 2015 and PISA 2000 are not possible, nor are comparisons in science performance between PISA 2015 and PISA 2000 or PISA 2003.) $PISA_{2015}$ is the mathematics, reading or science score observed in PISA 2015, and $PISA_t$ is the mathematics, reading or science score observed in a previous assessment. The standard error of the change in performance $\sigma(\Delta_{2015\text{-}t})$ is:

$$\delta(\Delta_{2015-t}) = \sqrt{\delta_{2015}^2 + \delta_t^2 + error_{2015,t}^2}$$

where σ_{2015} is the standard error observed for $PISA_{2015}$, σ_t is the standard error observed for $PISA_t$ and $error_{2015,t}$ is the link error for comparisons of science, reading or mathematics performance between the PISA 2015 assessment and a previous (t) assessment. The value for $error_{2015,t}$ is shown in Table A5.2 for most of the comparisons and Table A5.6 for comparisons of proficiency levels.

A second set of analyses reported in PISA 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 2012 and PISA 2015, the average three-year trend is equal to the difference between the two assessments.[8]

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

$$PISA_{i,t} = \beta_0 + \beta_1 time_t + \varepsilon_{i,t}$$

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:

$$\sigma(\beta_1) = \sqrt{{\sigma_{s,i}}^2(\beta_1) + {\sigma_l}^2(\beta_1)}$$

where $\sigma_{s,i}(\beta_1)$ 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 three-year trend. It is presented in Table A5.7.

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.

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

$$PISA_{i,t} = \beta_2 + \beta_3 year_t + \beta_4 year_t^2 + \varepsilon_{i,t}$$

where $year_t$ is a variable measuring time in years since 2015 and $year_t^2$ is equal to the square of $year_t$. Because *year* is scaled such that it is equal to zero in 2015, β_3 indicates the estimated annual rate of change in 2015 and β_2 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 2012 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 2012 are lower (more negative) than those observed in earlier years. Just as a link error is added when 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 A5.8). Curvilinear trends are only estimated for reading, and 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 (or, more precisely, 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 median level for all 15-year-olds. With this assumption, the median score among 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.

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.6.1 and Annex A2) 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 2015 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.[9]

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 2015 sample. Estimates of the mean and distribution of student performance are then performed 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 2015 in terms of the variables used in this re-weighting procedure.

Table A5.9 provides, for each country/economy, the number of cells used for post-stratification, as well as, for each cycle, the number of observations excluded from trends accounting for changes in the demographic characteristics of the sample.

Table A5.10 provides, for each country/economy, the means of the background variables used for the adjustment.

Comparing items and non-performance scales across PISA cycles

To gather information about students' and schools' characteristics, PISA asks both students and school principals to complete a background questionnaire. Between PISA 2006 and PISA 2015, several questions remained the same, allowing for a comparison of responses to these questions over time. Questions with subtle word changes or questions with major word changes were not compared across time (unless otherwise noted) because it is impossible to discern whether observed changes in the response are due to changes in the construct they are measuring or to changes in the way the construct is being measured.

Also, as described in Annex A1, questionnaire items in PISA are used to construct indices. Two types of indices are used in PISA: simple indices and scale indices.

Simple indices recode a set of responses to questionnaire items. For trends analyses, the values observed in PISA 2006 are compared directly to PISA 2015, just as simple responses to questionnaire items are. This is the case of indices like student-teacher ratio or immigrant status.

Scale indices, on the other hand, are included as Warm Likelihood Estimates (WLE; Warm, 1989) in the database and are based on a generalised partial credit model (GPCM; see Muraki 1992). Whenever at least part of the questions used in the construction of indices remains intact in PISA 2006 and PISA 2015, scaling of the corresponding index is based on a concurrent calibration with PISA 2006 and PISA 2015 data, followed by a linear transformation to report the resulting scale on the original PISA 2006 scale for the index, which was derived under a partial credit model (PCM; see OECD 2009). This procedure, which is analogous to the procedure used for cognitive scales, ensures that the corresponding index values can be compared.

To evaluate change in these items and scales, analyses report the change in the estimate between two assessments, usually PISA 2006 and PISA 2015. Comparisons between two assessments (e.g. a country's/economy's change index of enjoyment of learning science between PISA 2006 and PISA 2015 or the change in this index for a subgroup) is calculated as:

$$\Delta_{2015,2006} = PISA_{2015} - PISA_{2006}$$

where $\Delta_{2012,t}$ is the difference in the index between PISA 2015 and a previous assessment, $PISA_{2015}$ is the index value observed in PISA 2015, and $PISA_{2006}$ is the index value observed in 2006. The standard error of the change in the index value $\sigma(\Delta_{2015\text{-}2006})$ is:

$$\delta(\Delta_{2015-2006}) = \sqrt{\delta_{2015}^2 + \delta_{2006}^2}$$

where σ_{2015} is the standard error observed for $PISA_{2015}$ and σ_{2006} is the standard error observed for $PISA_{2006}$. Standard errors for changes in index values do not include measurement uncertainty and the uncertainty due to the equating procedure, and are therefore somewhat under-estimated. Standard errors for changes in responses to single items are not subject to measurement or equating uncertainty.

OECD average

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 trends tables and figures, the OECD average is reported on consistent sets of OECD countries. For instance, the "OECD average 33" includes only 33 OECD countries that have non-missing observations for the assessments for which this average itself is non-missing. This restriction allows for valid comparisons of the OECD average over time.

Tables available on line

Table A5.3. **Mean scores in science since 2006 produced with the 2015 approach to scaling**

Table A5.4. **Mean scores in reading since 2006 produced with the 2015 approach to scaling**

Table A5.5. **Mean scores in mathematics since 2006 produced with the 2015 approach to scaling**

Table A5.6. **Link error for comparisons of proficiency levels between PISA 2015 and previous assessments**

Table A5.7. **Link error for comparisons of the average three-year change between PISA 2015 and previous assessments**

Table A5.8. **Link error for the curvilinear trend between PISA 2015 and previous assessments**

Table A5.9. **Cells used to adjust science, reading and mathematics scores to the PISA 2015 samples.**

Table A5.10. **Descriptive statistics for variables used to adjust science, reading and mathematics scores to the PISA 2015 samples.**

References

Birnbaum, A. (1968), *On the Estimation of Mental Ability*, Series Report 15, USAF School of Aviation Medicine, Randolph Air Force Base (TX).

Carstensen, C.H. (2013), "Linking PISA Competencies over Three Cycles – Results from Germany", pp. 199–213 in *Research on PISA*, Springer, Netherlands, http://dx.doi.org/10.1007/978-94-007-4458-5_12.

Davidov, E., P. Schmidt and J. Billiet (eds.) (2011), *Cross-Cultural Analysis: Methods and Applications*. Routledge, New York.

Glas, C. and K. Jehangir (2014), "Modeling Country Specific Differential Item Functioning", in *Handbook of International Large-Scale Assessment*, CRC Press, Boca Raton (FL).

Masters, G.N. (1982), "A Rasch Model for Partial Credit Scoring." *Psychometrika*, Vol.47/2, pp. 149–74, http://dx.doi.org/10.1007/BF02296272.

Meredith, W. (1993), "Measurement Invariance, Factor Analysis and Factorial Invariance", *Psychometrika*, Vol. 58/4, pp. 525–43, http://dx.doi.org/10.1007/BF02294825.

Muraki, E. (1992), "A Generalized Partial Credit Model: Application of an EM Algorithm" *Applied Psychological Measurement*, Vol. 16/2, pp. 159–76, http://dx.doi.org/10.1177/014662169201600206.

OECD (forthcoming), *PISA 2015 Technical Report*, PISA, OECD Publishing, Paris.

OECD (2014a), *PISA 2012 Technical Report*, OECD Publishing, Paris, http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

OECD (2014b), *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised Edition, February 2014)*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264208780-en.

OECD (2009), *PISA 2006 Technical Report*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264048096-en.

Oliveri, M.E. and M. von Davier (2014), "Toward Increasing Fairness in Score Scale Calibrations Employed in International Large-Scale Assessments" *International Journal of Testing*, Vol. 14/1, pp. 1–21, http://dx.doi.org/10.1080/15305058.2013.825265.

Oliveri, M.E. and M. von Davier (2011), "Investigation of Model Fit and Score Scale Comparability in International Assessments" *Pyschological Test and Assessment Modeling*, Vol. 53/1, pp. 315–33.

Rasch, G (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen & Lydiche, Copenhagen.

Rousseeuw, P.J. 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.

Urbach, D. (2013), "An Investigation of Australian OECD PISA Trend Results", in *Research on PISA*, pp. 165–79, Springer Netherlands http://dx.doi.org/10.1007/978-94-007-4458-5_10.

Warm, T.A. (1989), "Weighted likelihood estimation of ability in item response theory", *Psychometrika*, Vol. 54/3, pp 427-450, http://dx.doi.org/10.1007/BF02294627.

[1] Also see Carstensen (2013) for the influence of test design on trend measurement.

[2] The limited treatment of DIF in past cycles, combined with the cycle-specific calibration sample, has been criticised for leading to trend estimates that are inconsistent with national calibrations using concurrent samples (Urbach, 2013).

[3] The number of not reached items is used in PISA 2015 as a source of background information in the generation of plausible values, so that the correlation of not-reached items and proficiency is modelled and accounted for in the results.

[4] The model allows for some countries/economies to contribute data for fewer than four assessment years.

[5] The correlation of PISA 2009 and PISA 2012 mean scores, for countries/economies that participated in 2015, is 0.985 in science (where both assessments coincide with years in which science was a minor domain, and therefore use the exact same tasks), 0.972 in reading (where PISA 2012 uses only a subset of PISA 2009 tasks) and 0.981 in mathematics (where PISA 2012 coincides with a revision of the framework and a larger set of assessment tasks). PISA 2009 and PISA 2012 are the two cycles with the most similar test design and approach to scaling. The correlation of PISA 2000 and PISA 2009 mean scores in reading (for countries/economies that participated in 2015) is 0.955; the correlation of PISA 2003 and PISA 2012 mean scores in mathematics is 0.953; and the correlation of PISA 2006 and PISA 2015 mean scores in science is 0.947 (0.944 based on results in Table A5.3, derived under a consistent approach to scaling).

[6] The country means produced during scaling are those that would have been observed based only on students who have response data on the domains. However, because PISA imputes data for all students in all domains assessed in a country/economy, whether a student has received a booklet that contains units for a domain or not, the model-based mean scores produced during scaling may differ from the mean scores reported in Annex B. However, the effect of imputed scores on means is negligible, as can be seen by comparing the results for 2015 between the estimates, based on the scaling mode, reported in Tables A5.3, A5.4 and A5.5, and the estimates, based on the full population model, reported in Tables I.2.3, I.4.3 and I.5.3.

[7] When examining results for a particular country or economy, these numbers must be interpreted as an upper bound on the actual number of scalar invariant items, because of the possibility of country-and-cycle-specific deviations from the international norm.

[8] The average three-year trend is related to what was referred to, in previous PISA reports, as the "annualised change" (OECD, 2014b). The average three-year trend can be obtained by multiplying the annualised change by three.

[9] Samples are always first separated by immigrant status (unless this would result in groups with fewer than 10 observations), then, within groups defined by immigrant status, by gender (unless this would result in groups with fewer than 10 observations), and finally by age groups. At any stage, if there are groups with fewer than 10 observations, the following mergers are done; within each stage, the sequence of mergers stops as soon as all groups reach a minimum size of 10. Step 1 (immigrant status, within language groups defined previously): merge missing and non-immigrant; merge "first generation" and "second generation"; merge all categories. Step 2 (gender, within immigrant groups defined previously): merge boys and girls. Step 3 (age, within immigrant/gender groups defined previously): merge first and second quarter; merge third and fourth quarter; merge all categories.