



Annex A

PISA 2015 TECHNICAL BACKGROUND

All tables in Annex A are available [on line](#)

Annex A1: Indices from the student and school context questionnaire

Annex A2: The PISA target population, the PISA samples and the definition of schools

<http://dx.doi.org/10.1787/888933433129>

Annex A3: Technical notes on analyses in this volume

Annex A4: Quality assurance

Annex A5: Changes in the administration and scaling of PISA 2015 and implications for trends analyses

<http://dx.doi.org/10.1787/888933433162>

Annex A6: The PISA 2015 field Trial mode-effect study

Note regarding B-S-J-G (China)

B-S-J-G (China) refers to the four PISA participating China provinces : Beijing, Shanghai, Jiangsu, Guangdong.

Note regarding CABA (Argentina)

CABA (Argentina) refers to the Ciudad Autónoma de Buenos Aires, Argentina.

Note regarding FYROM

FYROM refers to the Former Yugoslav Republic of Macedonia.

Notes regarding Cyprus

Note by Turkey: The information in this document with reference to “Cyprus” relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the “Cyprus issue”.

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

A note regarding Israel

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.



ANNEX A1

INDICES FROM THE STUDENT AND SCHOOL CONTEXT QUESTIONNAIRE

Explanation of the indices

This section explains the indices derived from the PISA 2015 student and school context questionnaires used in this volume.

Several PISA measures reflect indices that summarise responses from students, their parents, teachers or school representatives (typically principals) to a series of related questions. The questions were selected from a larger pool of questions on the basis of theoretical considerations and previous research. The *PISA 2015 Assessment and Analytical Framework* (OECD, 2016) provides an in-depth description of this conceptual framework. Structural equation modelling was used to confirm the theoretically expected behaviour of the indices and to validate their comparability across countries. For this purpose, a model was estimated separately for each country and collectively for all OECD countries. For a detailed description of other PISA indices and details on the methods, see the *PISA 2015 Technical Report* (OECD, forthcoming).

There are three types of indices: simple indices, new scale indices, and trend scale indices.

Simple indices are the variables that are constructed through the arithmetic transformation or recoding of one or more items in exactly the same way across assessments. Here, item responses are used to calculate meaningful variables, such as the recoding of the four-digit ISCO-08 codes into “Highest parents’ socio-economic index (HISEI)” or teacher-student ratio based on information from the school questionnaire.

New and trend scale indices are the variables constructed through the scaling of multiple items. Unless otherwise indicated, the index was scaled using a two-parameter item response model (a generalised partial credit model was used in the case of items with more than two categories) and values of the index correspond to Warm likelihood estimates (WLE) (Warm, 1985). For details on how each scale index was constructed, see the *PISA 2015 Technical Report* (OECD, forthcoming). In general, the scaling was done in three stages:

1. The item parameters were estimated from equally-weighted samples of students from all countries and economies; only cases with a minimum number of three valid responses to items that are part of the index were included. In the case of **trend indices**, a common calibration linking procedure was used: countries/economies that participated in both PISA 2006 and PISA 2015 contributed both samples to the calibration of item parameters; each cycle, and, within each cycle, each country/economy contributed equally to the estimation.
2. The estimates were computed for all students and all schools by anchoring the item parameters obtained in the preceding step.
3. For **new scale indices**, the Warm likelihood estimates were then standardised so that the mean of the index value for the OECD student population was zero and the standard deviation was one (countries being given equal weight in the standardisation process). **Trend indices** were equated so that the mean and standard deviation across OECD countries of rescaled PISA 2006 estimates and of the original estimates included in the PISA 2006 database matched. Trend indices are therefore reported on the same scale as used originally in PISA 2006, so that values can be directly compared to those included in the PISA 2006 database.

Sequential codes were assigned to the different response categories of the questions in the sequence in which the latter appeared in the student, school or parent questionnaires. Where indicated in this section, these codes were inverted for the purpose of constructing indices or scales. Negative values for an index do not necessarily imply that students responded negatively to the underlying questions. A negative value merely indicates that the respondents answered less positively than all respondents did on average across OECD countries. Likewise, a positive value on an index indicates that the respondents answered more favourably, or more positively, on average, than respondents in OECD countries did. Terms enclosed in brackets < > in the following descriptions were replaced in the national versions of the student, school and parent questionnaires by the appropriate national equivalent. For example, the term <qualification at ISCED level 5A> was translated in the United States into “Bachelor’s degree, post-graduate certificate program, Master’s degree program or first professional degree program”. Similarly the term <classes in the language of assessment> in Luxembourg was translated into “German classes” or “French classes”, depending on whether students received the German or French version of the assessment instruments.

In addition to simple and scaled indices described in this annex, there are a number of variables from the questionnaires that were used in this volume and correspond to single items not used to construct indices. These non-recoded variables have prefix of “ST” for the questionnaire items in the student questionnaire and “SC” for the items in the school questionnaire. All the context questionnaires, and the PISA international database, including all variables, are available through www.oecd.org/pisa.



Student-level simple indices

Student age

The age of a student (AGE) was calculated as the difference between the year and month of the testing and the year and month of a student's birth. Data on student's age were obtained from both the questionnaire (ST003) and the student tracking forms. If the month of testing was not known for a particular student, the median month for that country was used in the calculation.

Parents' level of education

Students' responses on questions ST005, ST006, ST007, and ST008 regarding parental education were classified using ISCED 1997 (OECD, 1999). Indices on parental education were constructed by recoding educational qualifications into the following categories: (0) None, (1) <ISCED level 1> (primary education), (2) <ISCED level 2> (lower secondary), (3) <ISCED level 3B or 3C> (vocational/pre-vocational upper secondary), (4) <ISCED level 3A> (general upper secondary) and/or <ISCED level 4> (non-tertiary post-secondary), (5) <ISCED level 5B> (vocational tertiary) and (6) <ISCED level 5A> and/or <ISCED level 6> (theoretically oriented tertiary and post-graduate). Indices with these categories were provided for a student's mother (MISCED) and father (FISCED). In addition, the index of highest education level of parents (HISCED) corresponds to the higher ISCED level of either parent. The index of highest education level of parents was also recoded into estimated number of years of schooling (PARED). The correspondence between education levels and years of schooling is available in the *PISA 2015 Technical Report* (OECD, forthcoming).

Parents' highest occupational status

Occupational data for both the student's father and the student's mother were obtained from responses to open-ended questions. The responses were coded to four-digit ISCO codes (ILO, 2007) and then mapped to the international socio-economic index of occupational status (ISEI) (Ganzeboom & Treiman, 2003). In PISA 2015, as in PISA 2012, the new ISCO and ISEI in their 2008 version were used rather than the 1988 versions that had been applied in the previous four cycles (Ganzeboom, 2010). Three indices were calculated based on this information: father's occupational status (BFMJ2); mother's occupational status (BMMJ1); and the highest occupational status of parents (HISEI) which corresponds to the higher ISEI score of either parent or to the only available parent's ISEI score. For all three indices, higher ISEI scores indicate higher levels of occupational status.

Immigrant background

The PISA database contains three country-specific variables relating to the students' country of birth, their mother and father (COBN_S, COBN_M, and COBN_F). The items ST019Q01TA, ST019Q01TB and ST019Q01TC were recoded into the following categories: (1) country of birth is the same as country of assessment and (2) other. The index of immigrant background (IMMIG) was calculated from these variables with the following categories: (1) non-immigrant students (those students who had at least one parent born in the country), (2) second-generation immigrant students (those born in the country of assessment but whose parent(s) were born in another country) and (3) first-generation immigrant students (those students born outside the country of assessment and whose parents were also born in another country). Students with missing responses for either the student or for both parents were assigned missing values for this variable.

Language spoken at home

Students indicated what language they usually speak at home (ST022), and the database includes a derived variable (LANGN) containing a country-specific code for each language. In addition, an internationally comparable variable (ST022Q01TA) was derived from this information with the following categories: (1) language at home is the same as the language of assessment for that student and (2) language at home is another language.

Grade repetition

The grade repetition variable (REPEAT) was computed by recoding variables ST127Q01TA, ST127Q02TA, and ST127Q03TA. REPEAT took the value of "1" if the student had repeated a grade in at least one ISCED level and the value of "0" if "no, never" was chosen at least once, given that none of the repeated grade categories were chosen. The index is assigned a missing value if none of the three categories were ticked in any levels.

Study programme

PISA collects data on study programmes available to 15-year old students in each country. This information is obtained through the student tracking form and the student questionnaire. In the final database, all national programmes are included in a separate derived variable (PROGN) where the first six digits represent the National Centre code, and the last two digits are the nationally specific programme code. All study programmes were classified using the International Standard Classification of Education (ISCED) (OECD, 1999). The following indices were derived from the data on study programmes:

- Programme level (ISCEDL) indicates whether students were at the lower or upper secondary level (ISCED 2 or ISCED 3).
- Programme designation (ISCEDD) indicates the designation of the study programme (A = general programmes designed to give access to the next programme level, B = programmes designed to give access to vocational studies at the next programme level, C = programmes designed to give direct access to the labour market, M = modular programmes that combine any or all of these characteristics).

- Programme orientation (ISCEDO) indicates whether the programme's curricular content was general, pre-vocational or vocational.

Science-related career expectations

In PISA 2015, students were asked to answer a question (ST114) about “what kind of job [they] expect to have when [they] are about 30 years old”. Answers to this open-ended question were coded to four-digit ISCO codes (ILO, 2007), in variable OCOD3. This variable was used to derive the index of science-related career expectations.

Science-related career expectations are defined as those career expectations whose realisation requires further engagement with the study of science beyond compulsory education, typically in formal tertiary education settings. The classification of careers into science-related and non-science-related is based on the four-digit ISCO-08 classification of occupations.

Only professionals (major ISCO group 2) and technicians/associate professionals (major ISCO group 3) were considered to fit the definition of science-related career expectations. In a broad sense, several managerial occupations (major ISCO group 1) are clearly science-related: these include research and development managers, hospital managers, construction managers, and other occupations classified under production and specialised services managers (submajor group 13). However, it was considered that when science-related experience and training is an important requirement of a managerial occupation, these are not entry-level jobs and 15-year-old students with science-related career expectations would not expect to be in such a position by age 30.

Several skilled agriculture, forestry and fishery workers (major ISCO group 6) could also be considered to work in science-related occupations. The United States O*NET OnLine (2016) classification of science, technology, engineering and mathematics (STEM) occupations indeed include these occupations. These, however, do not typically require formal science-related training or study after compulsory education. On these grounds, only major occupation groups that require ISCO skill levels 3 and 4 were included among science-related occupational expectations.

Among professionals and technicians/associate professionals, the boundary between science-related and non-science related occupations is sometimes blurred, and different classifications draw different lines.

The classification used in this report includes four groups of jobs:¹

1. **Science and engineering professionals:** All science and engineering professionals (submajor group 21), except product and garment designers (2163), graphic and multimedia designers (2166).
2. **Health professionals:** All health professionals in submajor group 22 (e.g. doctors, nurses, veterinarians), with the exception of traditional and complementary medicine professionals (minor group 223).
3. **ICT professionals:** All information and communications technology professionals (submajor group 25).
4. **Science technicians and associate professionals,** including:
 - physical and engineering science technicians (minor group 311)
 - life science technicians and related associate professionals (minor group 314)
 - air traffic safety electronic technicians (3155)
 - medical and pharmaceutical technicians (minor group 321), except medical and dental prosthetic technicians (3214)
 - telecommunications engineering technicians (3522).

How this classification compares to existing classifications

When three existing classifications of 15-year-olds' science career expectations, all based on the International Standard Classification of Occupations (ISCO), 1988 edition (ISCO-88), are compared to the present classification, based on ISCO-08, a few differences emerge. Some are due to the updated version of occupational codings (as discussed in the next section); the remaining differences are summarised in Table A1.1.

Developing a comparable classification for ISCO-88

The same open-ended question was also included in the PISA 2006 questionnaire (ID in 2006: ST30), but students' answers were coded in the PISA 2006 database according to ISCO-88. It is not possible to ensure a strictly comparable classification. To report changes over time, the correspondence described in Table A1.2 was used to derive a similar classification based on PISA 2006 data.



Table A1.1 ■ Differences in the definition of science-related career expectations

	This classification	OECD (2007)	Sikora and Pokropek (2012)	Kjærnsli and Lie (2011)
Science-related managerial jobs	out	in	in	out
Psychologists	out	in	in	out
Sociologists and social work professionals	out	in	out	out
Photographers and image and sound recording equipment operators, broadcasting and telecommunications equipment operators	out	in	in	out
Statistical, mathematical and related associate professionals	out	out	in	out
Aircraft controllers (e.g. pilots, air traffic controllers)	out	in	in	out
Ship controllers (Ships' desk officers, etc.)	out	out	in	out
Medical assistants, dental assistants, veterinary assistants, nursing and midwifery associate professionals	out	in	in	out
Computer assistants, computer equipment operators and industrial robot controllers	out	out	out	in
Air traffic safety electronic technicians	in	in	in	out
Pharmaceutical technicians and assistants	in	in	in	out
Dieticians and nutritionists	in	in	in	out

Table A1.2 ■ ISCO-08 to ISCO-88 correspondence table for science-related career expectations

Group	ISCO-08	ISCO-88
<i>Science and engineering professionals</i>	21xx (except 2163 and 2166)	21xx (except 213x), 221x
<i>Health professionals</i>	22xx (except 223x)	22xx (except 221x), 3223, 3226
<i>ICT professionals</i>	25xx	213x
<i>Science technicians and associate professionals</i>	311x, 314x, 3155, 321x (except 3214), 3522	311x, 3133, 3145, 3151, 321x, 3228

The main differences between ISCO-88 and ISCO-08, for the purpose of deriving the index of science-related career expectations, are the following:

- Medical equipment operators (ISCO-88: 3133) correspond to medical imaging and therapeutic equipment technicians in ISCO-08; air traffic safety technicians (ISCO-88: 3145) correspond to air traffic safety electronics technicians in ISCO-08; building and fire inspectors (ISCO-88: 3151) mostly correspond to civil engineering technicians in ISCO-08.
- Dieticians and nutritionists (ISCO-88: 3223) are classified among professionals in ISCO-08. For consistency, this ISCO-88 occupation was classified among health professionals.
- Physiotherapists and related associate professionals (ISCO-88: 3226) form two distinct categories in ISCO-08, with physiotherapists classified among professionals. Given that students who expect to work as physiotherapists far outnumber those who expect to work as related associate professionals, this ISCO-88 occupation was classified among health professionals.
- Several health-related occupations classified as “modern health associate professionals” in ISCO-88 are included among health professionals in ISCO-08 (e.g. speech therapist, ophthalmic opticians). While health professionals are, in general, included among science-related careers, health associate professionals are not included among science-related careers. In applying the classification to ISCO-88, the entire code was excluded from science-related careers.
- Telecommunications engineering technicians (ISCO-08: 3522) do not form a separate occupation in ISCO-88, where they can be found among electronics and telecommunications engineering technicians (ISCO-88: 3114).
- Information and communications technology professionals form a distinct submajor group (25) in ISCO-08 but are classified among physical, mathematical and engineering science professionals in ISCO-88.

Student-level scale indices

New scale indices

Interest in science

The index of broad interest in science topics (INTBRSCI) was constructed using students' responses to a new question developed for PISA 2015 (ST095). Students reported on a five-point Likert scale with the categories "not interested", "hardly interested", "interested", "highly interested", and "I don't know what this is", their interest in the following topics: biosphere (e.g. ecosystem services, sustainability); motion and forces (e.g. velocity, friction, magnetic and gravitational forces); energy and its transformation (e.g. conservation, chemical reactions); the Universe and its history; how science can help us prevent disease. The last response category ("I don't know what this is") was recoded as a missing for the purpose of deriving the index INTBRSCI. Higher values on the index reflect greater levels of agreement with these statements.

Epistemic beliefs about science

The index of epistemic beliefs about science (EPIST) was constructed using students' responses to a new question developed for PISA 2015 about students' views on scientific approaches (ST131). Students reported, on a four-point Likert scale with the answering categories "strongly disagree", "disagree", "agree", and "strongly agree", their agreement with the following statements: A good way to know if something is true is to do an experiment; Ideas in <broad science> sometimes change; Good answers are based on evidence from many different experiments; It is good to try experiments more than once to make sure of your findings; Sometimes <broad science> scientists change their minds about what is true in science; and The ideas in <broad science> science books sometimes change. Higher levels on the index correspond to greater levels of agreement with these statements.

Trend scale indices

Enjoyment of science

The index of enjoyment of science (JOYSCIE) was constructed based on a trend question (ST094) from PISA 2006 (ID in 2006: ST16), asking students on a four-point Likert scale with the categories "strongly agree", "agree", "disagree", and "strongly disagree" about their agreement with the following statements: I generally have fun when I am learning <broad science> topics; I like reading about <broad science>; I am happy working on <broad science> topics; I enjoy acquiring new knowledge in <broad science>; and I am interested in learning about <broad science>. The derived variable JOYSCIE was equated to the corresponding scale in the PISA 2006 database, thus allowing for a trend comparison between PISA 2006 and PISA 2015. Higher values on the index reflect greater levels of agreement with these statements.

Science self-efficacy

The index of science self-efficacy (SCIEEFF) was constructed based on a trend question (ST129) that was taken from PISA 2006 (ID in 2006: ST17). Students were asked, using a four-point answering scale with the categories "I could do this easily", "I could do this with a bit of effort", "I would struggle to do this on my own", and "I couldn't do this", to rate how they would perform in the following science tasks: recognise the science question that underlies a newspaper report on a health issue; explain why earthquakes occur more frequently in some areas than in others; describe the role of antibiotics in the treatment of disease; identify the science question associated with the disposal of garbage; predict how changes to an environment will affect the survival of certain species; interpret the scientific information provided on the labelling of food items; discuss how new evidence can lead you to change your understanding about the possibility of life on Mars; and identify the better of two explanations for the formation of acid rain. Responses were reverse-coded so that higher values of the index correspond to higher levels of science self-efficacy. The derived variable SCIEEFF was equated to the corresponding scale in the PISA 2006 database, thus allowing for a trend comparison between PISA 2006 and PISA 2015.

Science activities

The index of science activities (SCIEACT) was constructed based on a trend question (ST146) from PISA 2006 (ID in 2006: ST19). Students were asked to report on a four-point scale with the answering categories "very often", "regularly", "sometimes", and "never or hardly ever" how often they engaged in the following science-related activities: watch TV programmes about <broad science>; borrow or buy books on <broad science> topics; visit web sites about <broad science> topics; read <broad science> magazines or science articles in newspapers; attend a <science club>; simulate natural phenomena in computer programs/virtual labs; simulate technical processes in computer programs/virtual labs; visit web sites of ecology organisations; and follow news of science, environmental, or ecology organizations via blogs and microblogging. Responses were reverse-coded so that higher values of the index correspond to higher levels of students' science activities. The derived variable SCIEACT was equated to the corresponding scale in the PISA 2006 database, thus allowing for a trend comparison between PISA 2006 and PISA 2015.

Instrumental motivation to learn science

The index of instrumental motivation to learn science (INSTSCIE) was constructed based on a trend question (ST113) from PISA 2006 (ID in 2006: ST35). Students reported on a four-point Likert scale with the categories "strongly agree", "agree", "disagree", and "strongly disagree" about their agreement with the statements: Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on; What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on; Studying my <school science> subject(s) is worthwhile for me because



what I learn will improve my career prospects; and Many things I learn in my <school science> subject(s) will help me to get a job. Responses were reverse-coded so that higher values of the index correspond to higher levels of instrumental motivation. The derived variable INSTSCIE was equated to the corresponding scale in the PISA 2006 database, thus allowing for a trend comparison between PISA 2006 and PISA 2015.

Scaling of indices related to the PISA index of economic social and cultural status

The PISA index of economic, social and cultural status (ESCS) was derived, as in previous cycles, from three variables related to family background: parents' highest level of education (PARED), parents' highest occupation status (HISEI), and home possessions (HOMEPOS), including books in the home. PARED and HISEI are simple indices, described above. HOMEPOS is a proxy measure for family wealth.

Household possessions

In PISA 2015, students reported the availability of 16 household items at home (ST011) including three country-specific household items that were seen as appropriate measures of family wealth within the country's context. In addition, students reported the amount of possessions and books at home (ST012, ST013).

HOMEPOS is a summary index of all household and possession items (ST011, ST012 and ST013). The home possessions scale for PISA 2015 was computed differently than in the previous cycles, to align the IRT model to the one used for all cognitive and non-cognitive scales. Categories for the number of books in the home are unchanged in PISA 2015. The ST011-Items (1="yes", 2="no") were reverse-coded so that a higher level indicates the presence of the indicator.

Computation of ESCS

For the purpose of computing the PISA index of economic, social and cultural status (ESCS), values for students with missing PARED, HISEI or HOMEPOS were imputed with predicted values plus a random component based on a regression on the other two variables. If there were missing data on more than one of the three variables, ESCS was not computed and a missing value was assigned for ESCS.

The PISA index of economic, social and cultural status was derived from a principal component analysis of standardised variables (each variable has an OECD mean of zero and a standard deviation of one), taking the factor scores for the first principal component as measures of the PISA index of economic, social and cultural status. All countries and economies (both OECD and partner countries/economies) contributed equally to the principal component analysis, while in previous cycles, the principal component analysis was based on OECD countries only. However, for the purpose of reporting the ESCS scale has been transformed with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries.

Principal component analysis was also performed for each participating country or economy separately, to determine to what extent the components of the index operate in similar ways across countries or economy.

Computation of a trend-ESCS index

While an index of economic, cultural and social status (ESCS) was included in all past PISA databases, the components of ESCS and the scaling model changed over cycles, meaning that ESCS scores are not comparable across cycles directly. In order to enable a trends study, in PISA 2015 the ESCS was computed for the current cycle and also recomputed for the earlier cycles using a similar methodology.²

Before trend scores could be estimated, slight adjustments to the three components had to be made:

- As in PISA 2012, the occupational coding scheme involved in the process of forming HISEI changed from ISCO-88 to ISCO-08, the occupational codes for previous cycles were mapped from the former to the current scheme (see also PISA 2012 Technical Report, Chapter 3).
- In order to make the PARED component comparable across cycles, the same ISCED to PARED mapping scheme was employed for all the cycles.
- To make the HOMEPOS component more comparable across cycles, the variable *Books in the home* (ST013Q01TA) was recoded into a four-level categorical variable (fewer than or equal to 25 books, 26-100 books, 101-500 books, more than 500 books). The trend HOMEPOS scale was constructed in three steps. In the first step, international item parameters for all items (except country-specific items, i.e. ST011Q17NA, ST011Q18NA and ST011Q19NA) administered in PISA 2015 were obtained from a concurrent calibration of the 2015 data. Except for the recoding of variable ST013Q01TA, this step is identical with the regular scaling of HOMEPOS in PISA 2015 (see above). In the second step, unique items from all previous cycles (i.e., 2000-2012) were scaled, fixing most items administered in 2015 to their 2015 parameters, while allowing a limited set of item parameters to be freely estimated but constrained to be equal across countries within cycles. National items (i.e. ST011Q17NA, ST011Q18NA and ST011Q19NA) received unique (country- and cycle- specific) parameters throughout. In the third and final step, index values (WLEs) were generated for all students from previous cycles (2000-2012). Because 17 out of 27 items involved in the computation of the trend HOMEPOS have the same item parameters across cycles, the trend HOMEPOS scores can be regarded to be on a joint scale, allowing for comparisons of countries across cycles and thus allowing to be used in the calculation of trend ESCS.

The principal component analysis for obtaining trend-ESCS scores was then calculated as described above, except that the calculation was done across all cycles using these three comparable components (trend HISEI, trend PARED, and trend HOMEPOS).

School-level scale indices

School resources

PISA 2015 included a question with eight items about school resources, measuring the school principals' perceptions of potential factors hindering the provision of instruction at school ("Is your school's capacity to provide instruction hindered by any of the following issues?"). The four response categories were "not at all", "very little", "to some extent", to "a lot". A similar question was used in previous cycles, but items were reduced and reworded for 2015 focusing on two derived variables. The index on staff shortage (STAFFSHORT) was derived from the four items: a lack of teaching staff; inadequate or poorly qualified teaching staff; a lack of assisting staff; and inadequate or poorly qualified assisting staff. The index of shortage of educational material (EDUSHORT) was scaled using the following four items: a lack of educational material (e.g. textbooks, IT equipment, library or laboratory material); inadequate or poor quality educational material (e.g. textbooks, IT equipment, library or laboratory material); a lack of physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems); and inadequate or poor quality physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems). Positive values on these indices mean that schools principals view the amount and/or quality of resources in their schools as an obstacle to providing instruction to a greater extent than the OECD average; negative values reflect the perception that the school suffers from a lack or inadequacy of resources to a lesser extent than the OECD average.

Proportion of missing observations for variables used in this volume

Unless otherwise indicated, no adjustment is made for non-response to questionnaires in analyses included in this volume. The reported percentages and estimates based on indices refer to the proportion of the sample with valid responses to the corresponding questionnaire items. Table A1.3, available online, reports the proportion of the sample covered by analyses based on student or school questionnaire variables. Where this proportion shows large variation across countries/economies or across time, caution is required when comparing results on these dimensions.

Table available online

Table A1.3. Weighted share of responding students covered by analyses based on questionnaires (<http://dx.doi.org/10.1787/888933433112>)



Notes

1. In the United Kingdom (excluding Scotland), career expectations were coded to the three-digit level only. As a result, the occupations of product and garment designers (ISCO08: 2163) and graphic and multimedia designers (2166) are included among science and engineering professionals, medical and dental prosthetic technicians (3214) are included among science technicians and associate professionals, while telecommunications engineering technicians (3522) are excluded. These careers represent a small percentage of the students classified as having science-related career expectations, such that results are not greatly affected.

2. As a result of this procedure, two indices exist for 2015 (ESCS and trend-ESCS). The Pearson correlation between the two indices is $r=.989$ across all PISA 2015 participating countries and economies. This includes 22 countries/economies where the correlation was $r>.990$; another 50 countries/economies where the correlation was $r=[.960, .990]$; and another country (Georgia) where it was $r=.946$. In Chapters 6 and 7, in order to maintain consistency across tables, results for 2015 relating to trends in ESCS employ the 2015 ESCS index rather than the 2015 trend-ESCS index.

References

Ganzeboom, H.B.G. (2010), "A new international socio-economic index [ISEI] of occupational status for the International Standard Classification of Occupation 2008 [ISCO-08] constructed with data from the ISSP 2002-2007; with an analysis of quality of occupational measurement in ISSP." Paper presented at Annual Conference of International Social Survey Programme, Lisbon, May 1 2010.

Ganzeboom, H. B.G. and D.J. Treiman (2003), "Three Internationally Standardised Measures for Comparative Research on Occupational Status", pp. 159-193 in J.H.P. Hoffmeyer-Zlotnik and C. Wolf (Eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, Kluwer Academic Press, New York.

Kjærnsli, M. and S. Lie (2011), "Students' Preference for Science Careers: International Comparisons Based on PISA 2006", *International Journal of Science Education*, Vol. 33/1, pp. 121-44, <http://dx.doi.org/10.1080/09500693.2010.518642>.

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

OECD (2016), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264255425-en>.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264040014-en>.

OECD (1999), *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries*, OECD Publishing, Paris.

Sikora, J. and A. Pokropek (2012), "Gender Segregation of Adolescent Science Career Plans in 50 Countries", *Science Education*, Vol. 96/2, pp. 234-64, <http://dx.doi.org/10.1002/sce.20479>.

Warm, T.A. (1985), "Weighted Maximum Likelihood Estimation of Ability in Item Response Theory with Tests of Finite Length", Technical Report CGI-TR-85-08, U.S. Coast Guard Institute, Oklahoma City.

O*NET OnLine (n.d), "All STEM disciplines", webpage, www.onetonline.org/find/stem?t=0, (accessed 4 October 2016).



ANNEX A2

THE PISA TARGET POPULATION, THE PISA SAMPLES AND THE DEFINITION OF SCHOOLS

Definition of the PISA target population

PISA 2015 provides an assessment of the cumulative outcomes of education and learning at a point at which most young adults are still enrolled in initial education.

A major challenge for an international survey is to ensure that international comparability of national target populations is guaranteed.

Differences between countries in the nature and extent of pre-primary education and care, the age at entry into formal schooling and the institutional structure of education systems do not allow for a definition of internationally comparable grade levels. Consequently, international comparisons of performance in education typically define their populations with reference to a target age group. Some previous international assessments have defined their target population on the basis of the grade level that provides maximum coverage of a particular age cohort. A disadvantage of this approach is that slight variations in the age distribution of students across grade levels often lead to the selection of different target grades in different countries, or between education systems within countries, raising serious questions about the comparability of results across, and at times within, countries. In addition, because not all students of the desired age are usually represented in grade-based samples, there may be a more serious potential bias in the results if the unrepresented students are typically enrolled in the next higher grade in some countries and the next lower grade in others. This would exclude students with potentially higher levels of performance in the former countries and students with potentially lower levels of performance in the latter.

In order to address this problem, PISA uses an age-based definition for its target population, i.e. a definition that is not tied to the institutional structures of national education systems. PISA assesses students who were aged between 15 years and 3 (complete) months and 16 years and 2 (complete) months at the beginning of the assessment period, plus or minus a 1-month allowable variation, and who were enrolled in an educational institution with grade 7 or higher, regardless of the grade level or type of institution in which they were enrolled, and regardless of whether they were in full-time or part-time education. Educational institutions are generally referred to as schools in this publication, although some educational institutions (in particular, some types of vocational education establishments) may not be termed schools in certain countries. As expected from this definition, the average age of students across OECD countries was 15 years and 9 months. The range in country means was 2 months and 18 days (0.20 years), from the minimum country mean of 15 years and 8 months to the maximum country mean of 15 years and 10 months.

Given this definition of population, PISA makes statements about the knowledge and skills of a group of individuals who were born within a comparable reference period, but who may have undergone different educational experiences both in and outside school. In PISA, these knowledge and skills are referred to as the outcomes of education at an age that is common across countries. Depending on countries' policies on school entry, selection and promotion, these students may be distributed over a narrower or a wider range of grades across different education systems, tracks or streams. It is important to consider these differences when comparing PISA results across countries, as observed differences between students at age 15 may no longer appear later on as/if students' educational experiences converge over time.

If a country's scores in science, reading or mathematics are significantly higher than those in another country, it cannot automatically be inferred that the schools or particular parts of the education system in the first country are more effective than those in the second. However, one can legitimately conclude that the cumulative impact of learning experiences in the first country, starting in early childhood and up to the age of 15, and embracing experiences in school, home and beyond, have resulted in higher outcomes in the literacy domains that PISA measures.

The PISA target population does not include residents attending schools in a foreign country. It does, however, include foreign nationals attending schools in the country of assessment.

To accommodate countries that requested grade-based results for the purpose of national analyses, PISA 2015 provided a sampling option to supplement age-based sampling with grade-based sampling.

Population coverage

All countries and economies attempted to maximise the coverage of 15-year-olds enrolled in education in their national samples, including students enrolled in special-education institutions. As a result, PISA 2015 reached standards of population coverage that are unprecedented in international surveys of this kind.

The sampling standards used in PISA permitted countries to exclude up to a total of 5% of the relevant population either by excluding schools or by excluding students within schools. All but 12 countries – the United Kingdom (8.22%), Luxembourg (8.16%), Canada (7.49%), Norway (6.75%), New Zealand (6.54%), Sweden (5.71%), Estonia (5.52%), Australia (5.31%),



Montenegro (5.17%), Lithuania (5.12%), Latvia (5.07%), and Denmark (5.04%) – achieved this standard, and in 29 countries and economies, the overall exclusion rate was less than 2%. When language exclusions were accounted for (i.e. removed from the overall exclusion rate), Denmark, Latvia, New Zealand and Sweden no longer had an exclusion rate greater than 5%. For details, see www.pisa.oecd.org.

Exclusions within the above limits include:

- At the school level: schools that were geographically inaccessible or where the administration of the PISA assessment was not considered feasible; and schools that provided teaching only for students in the categories defined under “within-school exclusions”, such as schools for the blind. The percentage of 15-year-olds enrolled in such schools had to be less than 2.5% of the nationally desired target population (0.5% maximum for the former group and 2% maximum for the latter group). The magnitude, nature and justification of school-level exclusions are documented in the *PISA 2015 Technical Report* (OECD, forthcoming).
- At the student level: students with an intellectual disability; students with a functional disability; students with limited assessment language proficiency; other (a category defined by the national centres and approved by the international centre); and students taught in a language of instruction for the main domain for which no materials were available. Students could not be excluded solely because of low proficiency or common disciplinary problems. The percentage of 15-year-olds excluded within schools had to be less than 2.5% of the nationally desired target population.

Table A2.1 describes the target population of the countries participating in PISA 2015. Further information on the target population and the implementation of PISA sampling standards can be found in the *PISA 2015 Technical Report* (OECD, forthcoming).

- **Column 1** shows the total number of 15-year-olds according to the most recent available information, which in most countries means the year 2014 as the year before the assessment.
- **Column 2** shows the number of 15-year-olds enrolled in schools in grade 7 or above (as defined above), which is referred to as the “eligible population”.
- **Column 3** shows the national desired target population. Countries were allowed to exclude up to 0.5% of students a priori from the eligible population, essentially for practical reasons. The following a priori exclusions exceed this limit but were agreed with the PISA Consortium: Belgium excluded 0.21% of its population for a particular type of student educated while working; Canada excluded 1.22% of its population from Territories and Aboriginal reserves; Chile excluded 0.04% of its students who live in Easter Island, Juan Fernandez Archipelago and Antarctica; and the United Arab Emirates excluded 0.04% of its students who had no information available. The adjudicated region of Massachusetts in the United States excluded 13.11% of its students, and North Carolina excluded 5.64% of its students. For these two regions, the desired target populations cover 15-year-old students in grade 7 or above in public schools only. The students excluded from the desired population are private school students.
- **Column 4** shows the number of students enrolled in schools that were excluded from the national desired target population, either from the sampling frame or later in the field during data collection.
- **Column 5** shows the size of the national desired target population after subtracting the students enrolled in excluded schools. This is obtained by subtracting Column 4 from Column 3.
- **Column 6** shows the percentage of students enrolled in excluded schools. This is obtained by dividing Column 4 by Column 3 and multiplying by 100.
- **Column 7** shows the number of students participating in PISA 2015. Note that in some cases this number does not account for 15-year-olds assessed as part of additional national options.
- **Column 8** shows the weighted number of participating students, i.e. the number of students in the nationally defined target population that the PISA sample represents.
- Each country attempted to maximise the coverage of PISA’s target population within the sampled schools. In the case of each sampled school, all eligible students, namely those 15 years of age, regardless of grade, were first listed. Sampled students who were to be excluded had still to be included in the sampling documentation, and a list drawn up stating the reason for their exclusion. Column 9 indicates the total number of excluded students, which is further described and classified into specific categories in Table A2.2.
- **Column 10** indicates the weighted number of excluded students, i.e. the overall number of students in the nationally defined target population represented by the number of students excluded from the sample, which is also described and classified by exclusion categories in Table A2.2. Excluded students were excluded based on five categories: students with an intellectual disability (the student has a mental or emotional disability and is cognitively delayed such that he/she cannot perform in the PISA testing situation); students with a functional disability (the student has a moderate to severe permanent physical disability such that he/she cannot perform in the PISA testing situation); students with limited proficiency in the assessment language (the student is unable to read or speak any of the languages of the assessment in the country and would be unable to overcome the language barrier in the testing situation – typically a student who has received less than one year of instruction in the languages of assessment may be excluded); other (a category defined by the national centres and approved by the international centre); and students taught in a language of instruction for the main domain for which no materials were available.



A corrigendum has been issued for this page. See: <http://www.oecd.org/about/publishing/Corrigenda-PISA2015-Volumel.pdf>

[Part 1/2]

Table A2.2 Exclusions

	Student exclusions (unweighted)					
	Number of excluded students with functional disability (Code 1)	Number of excluded students with intellectual disability (Code 2)	Number of excluded students because of language (Code 3)	Number of excluded students for other reasons (Code 4)	Number of excluded students because of no materials available in the language of instruction (Code 5)	School-level exclusion rate (%)
	(1)	(2)	(3)	(4)	(5)	(6)
OECD						
Australia	85	528	68	0	0	681
Austria	8	15	61	0	0	84
Belgium	4	18	17	0	0	39
Canada	156	1 308	366	0	0	1 830
Chile	6	30	1	0	0	37
Czech Republic	2	9	14	0	0	25
Denmark	18	269	156	70	1	514
Estonia	17	93	6	0	0	116
Finland	2	90	17	8	7	124
France	5	21	9	0	0	35
Germany	4	25	25	0	0	54
Greece	3	44	11	0	0	58
Hungary	3	13	9	30	0	55
Iceland	9	66	47	9	0	131
Ireland	25	57	55	60	0	197
Israel	22	68	25	0	0	115
Italy	78	147	21	0	0	246
Japan	0	2	0	0	0	2
Korea	3	17	0	0	0	20
Latvia	7	47	16	0	0	70
Luxembourg	4	254	73	0	0	331
Mexico	4	23	3	0	0	30
Netherlands	1	13	0	0	0	14
New Zealand	23	140	167	0	3	333
Norway	11	253	81	0	0	345
Poland	11	20	0	3	0	34
Portugal	4	99	2	0	0	105
Slovak Republic	7	71	2	34	0	114
Slovenia	33	36	45	0	0	114
Spain	9	144	47	0	0	200
Sweden	154	0	121	0	0	275
Switzerland	8	42	57	0	0	107
Turkey	1	23	7	0	0	31
United Kingdom	77	690	102	0	1	870
United States	16	120	44	13	0	193
Partners						
Albania	0	0	0	0	0	0
Algeria	0	0	0	0	0	0
Argentina	10	10	1	0	0	21
Brazil	20	99	0	0	0	119
B-S-J-G (China)	6	25	2	0	0	33
Bulgaria	39	6	4	0	0	49
Colombia	3	4	2	0	0	9
Costa Rica	3	1	0	9	0	13
Croatia	2	75	9	0	0	86
Cyprus*	12	164	52	0	0	228
Dominican Republic	1	3	0	0	0	4
FYROM	7	1	0	0	0	8
Georgia	3	25	7	0	0	35
Hong Kong (China)	0	35	1	0	0	36
Indonesia	0	0	0	0	0	0
Jordan	43	17	10	0	0	70
Kazakhstan	0	0	0	0	0	0
Kosovo	9	13	27	0	0	50
Lebanon	0	0	0	0	0	0
Lithuania	12	213	2	0	0	227
Macao (China)	0	0	0	0	0	0
Malaysia	10	22	9	0	0	41
Malta	8	27	6	0	0	41
Moldova	12	8	1	0	0	21
Montenegro	14	23	5	0	258	300
Peru	4	9	0	0	0	13
Qatar	76	110	7	0	0	193
Romania	1	1	1	0	0	3
Russia	3	10	0	0	0	13
Singapore	3	15	7	0	0	25
Chinese Taipei	3	19	0	0	0	22
Thailand	1	19	2	0	0	22
Trinidad and Tobago	0	0	0	0	0	0
Tunisia	0	0	3	0	0	3
United Arab Emirates	16	24	23	0	0	63
Uruguay	2	4	0	0	0	6
Viet Nam	0	0	0	0	0	0

Exclusion codes:

Code 1: Functional disability – student has a moderate to severe permanent physical disability.

Code 2: Intellectual disability – student has a mental or emotional disability and has either been tested as cognitively delayed or is considered in the professional opinion of qualified staff to be cognitively delayed.


Code 3: Limited assessment language proficiency – student is not a native speaker of any of the languages of the assessment in the country and has been resident in the country for less than one year.

Code 4: Other reasons defined by the national centres and approved by the international centre.

Code 5: No materials available in the language of instruction.

Note: For a full explanation of the details in this table please refer to the *PISA 2015 Technical Report* (OECD, forthcoming).

* See note at the beginning of this Annex.

StatLink  <http://dx.doi.org/10.1787/888933433129>

[Part 2/2]

Table A2.2 Exclusions

	Student exclusion (weighted)					
	Weighted number of excluded students with functional disability (Code 1)	Weighted number of excluded students with intellectual disability (Code 2)	Weighted number of excluded students because of language (Code 3)	Weighted number of excluded students for other reasons (Code 4)	Weighted number of excluded students because of no materials available in the language of instruction (Code 5)	Total weighted number of excluded students (Code 1-5)
	(7)	(8)	(9)	(10)	(11)	(12)
OECD						
Australia	932	6 011	793	0	0	7 736
Austria	74	117	675	0	0	866
Belgium	33	192	185	0	0	410
Canada	1 901	18 018	5 421	0	0	25 340
Chile	194	1 190	9	0	0	1 393
Czech Republic	40	140	188	0	0	368
Denmark	122	1 539	551	421	11	2 644
Estonia	29	176	13	0	0	218
Finland	18	858	156	67	58	1 157
France	562	2 144	914	0	0	3 620
Germany	423	2 562	2 357	0	0	5 342
Greece	43	729	193	0	0	965
Hungary	57	284	114	554	0	1 009
Iceland	9	67	47	9	0	132
Ireland	213	526	516	570	0	1 825
Israel	349	1 070	384	0	0	1 803
Italy	3 316	5 199	880	0	0	9 395
Japan	0	318	0	0	0	318
Korea	291	1 515	0	0	0	1 806
Latvia	21	115	38	0	0	174
Luxembourg	4	254	73	0	0	331
Mexico	842	4 802	1 165	0	0	6 810
Netherlands	33	469	0	0	0	502
New Zealand	233	1 287	1 568	0	24	3 112
Norway	105	2 471	790	0	0	3 366
Poland	876	1 339	0	203	0	2 418
Portugal	29	818	13	0	0	860
Slovak Republic	44	567	12	288	0	912
Slovenia	84	71	92	0	0	247
Spain	511	7 662	2 720	0	0	10 893
Sweden	2 380	0	1 944	0	0	4 324
Switzerland	91	540	726	0	0	1 357
Turkey	43	4 094	1 222	0	0	5 359
United Kingdom	2 724	27 808	4 001	0	214	34 747
United States	7 873	67 816	26 525	7 366	0	109 580
Partners						
Albania	0	0	0	0	0	0
Algeria	0	0	0	0	0	0
Argentina	579	770	18	0	0	1 367
Brazil	1 743	11 800	0	0	0	13 543
B-S-J-G (China)	438	2 970	201	0	0	3 609
Bulgaria	347	51	35	0	0	433
Colombia	181	309	17	0	0	507
Costa Rica	22	5	0	71	0	98
Croatia	13	501	75	0	0	589
Cyprus*	16	212	65	0	0	292
Dominican Republic	24	82	0	0	0	106
FYROM	15	4	0	0	0	19
Georgia	19	170	41	0	0	230
Hong Kong (China)	0	363	11	0	0	374
Indonesia	0	0	0	0	0	0
Jordan	656	227	122	0	0	1 006
Kazakhstan	0	0	0	0	0	0
Kosovo	28	37	104	0	0	174
Lebanon	0	0	0	0	0	0
Lithuania	40	1 000	10	0	0	1 050
Macao (China)	0	0	0	0	0	0
Malaysia	663	1 100	580	0	0	2 344
Malta	8	27	6	0	0	41
Moldova	66	51	1	0	0	118
Montenegro	27	38	6	0	261	332
Peru	224	520	0	0	0	745
Qatar	76	110	7	0	0	193
Romania	31	63	26	0	0	120
Russia	425	2 044	0	0	0	2 469
Singapore	22	115	43	0	0	179
Chinese Taipei	78	568	0	0	0	647
Thailand	114	1 830	163	0	0	2 107
Trinidad and Tobago	0	0	0	0	0	0
Tunisia	0	0	61	0	0	61
United Arab Emirates	30	75	47	0	0	152
Uruguay	10	22	0	0	0	32
Viet Nam	0	0	0	0	0	0

Exclusion codes:

Code 1: Functional disability – student has a moderate to severe permanent physical disability.

Code 2: Intellectual disability – student has a mental or emotional disability and has either been tested as cognitively delayed or is considered in the professional opinion of qualified staff to be cognitively delayed.


Code 3: Limited assessment language proficiency – student is not a native speaker of any of the languages of the assessment in the country and has been resident in the country for less than one year.

Code 4: Other reasons defined by the national centres and approved by the international centre.

Code 5: No materials available in the language of instruction.

Note: For a full explanation of the details in this table please refer to the *PISA 2015 Technical Report* (OECD, forthcoming).

* See note at the beginning of this Annex.

StatLink  <http://dx.doi.org/10.1787/888933433129>



- **Column 11** shows the percentage of students excluded within schools. This is calculated as the weighted number of excluded students (Column 10), divided by the weighted number of excluded and participating students (Column 8 plus Column 10), then multiplied by 100.
- **Column 12** shows the overall exclusion rate, which represents the weighted percentage of the national desired target population excluded from PISA either through school-level exclusions or through the exclusion of students within schools. It is calculated as the school-level exclusion rate (Column 6 divided by 100) plus within-school exclusion rate (Column 11 divided by 100) multiplied by 1 minus the school-level exclusion rate (Column 6 divided by 100). This result is then multiplied by 100.
- **Column 13** presents an index of the extent to which the national desired target population is covered by the PISA sample. Australia, Canada, Denmark, Estonia, Latvia, Lithuania, Luxembourg, Montenegro, New Zealand, Norway, Sweden and the United Kingdom were the only countries where the coverage is below 95%.
- **Column 14** presents an index of the extent to which 15-year-olds enrolled in schools are covered by the PISA sample. The index measures the overall proportion of the national enrolled population that is covered by the non-excluded portion of the student sample. The index takes into account both school-level and student-level exclusions. Values close to 100 indicate that the PISA sample represents the entire education system as defined for PISA 2015. The index is the weighted number of participating students (Column 8) divided by the weighted number of participating and excluded students (Column 8 plus Column 10), times the nationally defined target population (Column 5) divided by the eligible population (Column 2) (times 100).
- **Column 15** presents an index of the coverage of the 15-year-old population. This index is the weighted number of participating students (Column 8) divided by the total population of 15-year-old students (Column 1).

This high level of coverage contributes to the comparability of the assessment results. For example, even assuming that the excluded students would have systematically scored worse than those who participated, and that this relationship is moderately strong, an exclusion rate on the order of 5% would likely lead to an overestimation of national mean scores of less than 5 score points (on a scale with an international mean of 500 score points and a standard deviation of 100 score points). This assessment is based on the following calculations: if the correlation between the propensity of exclusions and student performance is 0.3, resulting mean scores would likely be overestimated by 1 score point if the exclusion rate is 1%, by 3 score points if the exclusion rate is 5%, and by 6 score points if the exclusion rate is 10%. If the correlation between the propensity of exclusions and student performance is 0.5, resulting mean scores would be overestimated by 1 score point if the exclusion rate is 1%, by 5 score points if the exclusion rate is 5%, and by 10 score points if the exclusion rate is 10%. For this calculation, a model was used that assumes a bivariate normal distribution for performance and the propensity to participate. For details, see the *PISA 2015 Technical Report* (OECD, forthcoming).

Sampling procedures and response rates

The accuracy of any survey results depends on the quality of the information on which national samples are based as well as on the sampling procedures. Quality standards, procedures, instruments and verification mechanisms were developed for PISA that ensured that national samples yielded comparable data and that the results could be compared with confidence.

Most PISA samples were designed as two-stage stratified samples (where countries applied different sampling designs, these are documented in the *PISA 2015 Technical Report* [OECD, forthcoming]). The first stage consisted of sampling individual schools in which 15-year-old students could be enrolled. Schools were sampled systematically with probabilities proportional to size, the measure of size being a function of the estimated number of eligible (15-year-old) students enrolled. At least 150 schools were selected in each country (where this number existed), although the requirements for national analyses often required a somewhat larger sample. As the schools were sampled, replacement schools were simultaneously identified, in case a sampled school chose not to participate in PISA 2015.

In the case of Iceland, Luxembourg, Macao (China), Malta and Qatar, all schools and all eligible students within schools were included in the sample.

Experts from the PISA Consortium performed the sample selection process for most participating countries and monitored it closely in those countries that selected their own samples. The second stage of the selection process sampled students within sampled schools. Once schools were selected, a list of each sampled school's 15-year-old students was prepared. From this list, 42 students were then selected with equal probability (all 15-year-old students were selected if fewer than 42 were enrolled). The number of students to be sampled per school could deviate from 42, but could not be less than 20.

Data-quality standards in PISA required minimum participation rates for schools as well as for students. These standards were established to minimise the potential for response biases. In the case of countries meeting these standards, it was likely that any bias resulting from non-response would be negligible, i.e. typically smaller than the sampling error.

A minimum response rate of 85% was required for the schools initially selected. Where the initial response rate of schools was between 65% and 85%, however, an acceptable school-response rate could still be achieved through the use of replacement schools.



This procedure brought with it a risk of increased response bias. Participating countries were, therefore, encouraged to persuade as many of the schools in the original sample as possible to participate. Schools with a student participation rate between 25% and 50% were not regarded as participating schools, but data from these schools were included in the database and contributed to the various estimations. Data from schools with a student participation rate of less than 25% were excluded from the database.

PISA 2015 also required a minimum participation rate of 80% of students within participating schools. This minimum participation rate had to be met at the national level, not necessarily by each participating school. Follow-up sessions were required in schools in which too few students had participated in the original assessment sessions. Student participation rates were calculated over all original schools, and also over all schools, whether original sample or replacement schools, and from the participation of students in both the original assessment and any follow-up sessions. A student who participated in the original or follow-up cognitive sessions was regarded as a participant. Those who attended only the questionnaire session were included in the international database and contributed to the statistics presented in this publication if they provided at least a description of their father's or mother's occupation.

Table A2.3 shows the response rates for students and schools, before and after replacement.

- **Column 1** shows the weighted participation rate of schools before replacement. This is obtained by dividing Column 2 by Column 3.
- **Column 2** shows the weighted number of responding schools before school replacement (weighted by student enrolment).
- **Column 3** shows the weighted number of sampled schools before school replacement (including both responding and non-responding schools, weighted by student enrolment).
- **Column 4** shows the unweighted number of responding schools before school replacement.
- **Column 5** shows the unweighted number of responding and non-responding schools before school replacement.
- **Column 6** shows the weighted participation rate of schools after replacement. This is obtained by dividing Column 7 by Column 8.
- **Column 7** shows the weighted number of responding schools after school replacement (weighted by student enrolment).
- **Column 8** shows the weighted number of schools sampled after school replacement (including both responding and non-responding schools, weighted by student enrolment).
- **Column 9** shows the unweighted number of responding schools after school replacement.
- **Column 10** shows the unweighted number of responding and non-responding schools after school replacement.
- **Column 11** shows the weighted student participation rate after replacement. This is obtained by dividing Column 12 by Column 13.
- **Column 12** shows the weighted number of students assessed.
- **Column 13** shows the weighted number of students sampled (including both students who were assessed and students who were absent on the day of the assessment).
- **Column 14** shows the unweighted number of students assessed. Note that any students in schools with student-response rates of less than 50% were not included in these rates (both weighted and unweighted).
- **Column 15** shows the unweighted number of students sampled (including both students that were assessed and students who were absent on the day of the assessment). Note that any students in schools where fewer than half of the eligible students were assessed were not included in these rates (neither weighted nor unweighted).

Definition of schools

In some countries, subunits within schools were sampled instead of schools, and this may affect the estimation of the between-school variance components. In Austria, the Czech Republic, Germany, Hungary, Japan, Romania and Slovenia, schools with more than one study programme were split into the units delivering these programmes. In the Netherlands, for schools with both lower and upper secondary programmes, schools were split into units delivering each programme level. In the Flemish community of Belgium, in the case of multi-campus schools, implantations (campuses) were sampled, whereas in the French community, in the case of multi-campus schools, the larger administrative units were sampled. In Australia, for schools with more than one campus, the individual campuses were listed for sampling. In Argentina and Croatia, schools that had more than one campus had the locations listed for sampling. In Spain, the schools in the Basque region with multi-linguistic models were split into linguistic models for sampling. In Luxembourg, a school on the border with Germany was split according to the country in which the students resided. In addition, the International schools in Luxembourg were split into the students who were instructed in any of the three official languages, and those in the part of the schools that was excluded because no materials were available in the languages of instruction. The United Arab Emirates had schools split by curricula, and sometimes by gender, with other schools remaining whole. Because of reorganisation, some of Sweden's schools were split into parts, with each part having one principal. In Portugal, schools were reorganised into clusters, with teachers and the principal shared by all units in the school cluster.

Grade levels

Students assessed in PISA 2015 are at various grade levels. The percentage of students at each grade level is presented by country in Table A2.4a and by gender within each country in Table A2.4b.


[Part 1/1]

Table A2.4a Percentage of students at each grade level

	All students											
	7th grade		8th grade		9th grade		10th grade		11th grade		12th grade and above	
	%	S.E.	%	S.E.	%	S.E.	%	S.E.	%	S.E.	%	S.E.
OECD												
Australia	0.0	(0.0)	0.1	(0.0)	11.2	(0.3)	74.6	(0.4)	14.0	(0.4)	0.1	(0.0)
Austria	0.0	(0.0)	2.0	(0.6)	20.8	(0.9)	71.2	(1.0)	5.9	(0.3)	0.0	(0.0)
Belgium	0.6	(0.1)	6.4	(0.5)	30.7	(0.7)	61.0	(0.9)	1.3	(0.1)	0.0	(0.0)
Canada	0.1	(0.0)	0.7	(0.1)	10.8	(0.5)	87.6	(0.6)	0.8	(0.1)	0.0	(0.0)
Chile	1.7	(0.3)	4.1	(0.6)	24.0	(0.7)	68.1	(1.0)	2.1	(0.2)	0.0	(0.0)
Czech Republic	0.5	(0.1)	3.9	(0.3)	49.4	(1.2)	46.2	(1.2)	0.0	(0.0)	0.0	c
Denmark	0.2	(0.1)	16.4	(0.6)	81.9	(0.7)	1.4	(0.5)	0.0	c	0.0	c
Estonia	0.8	(0.2)	21.3	(0.6)	76.6	(0.6)	1.3	(0.3)	0.0	c	0.0	(0.0)
Finland	0.5	(0.1)	13.6	(0.4)	85.7	(0.4)	0.0	(0.0)	0.2	(0.1)	0.0	c
France	0.0	(0.0)	1.0	(0.2)	23.1	(0.6)	72.5	(0.7)	3.2	(0.2)	0.1	(0.1)
Germany	0.5	(0.1)	7.7	(0.4)	47.3	(0.8)	43.1	(0.8)	1.5	(0.5)	0.0	(0.0)
Greece	0.2	(0.1)	0.7	(0.2)	3.8	(0.8)	95.3	(0.9)	0.0	c	0.0	c
Hungary	1.7	(0.3)	8.5	(0.5)	75.8	(0.7)	14.0	(0.5)	0.0	c	0.0	c
Iceland	0.0	c	0.0	c	0.0	c	100.0	c	0.0	c	0.0	c
Ireland	0.0	(0.0)	1.8	(0.2)	60.6	(0.7)	26.5	(1.1)	11.1	(0.9)	0.0	c
Israel	0.0	c	0.1	(0.0)	16.4	(0.9)	82.7	(0.9)	0.9	(0.3)	0.0	c
Italy	0.1	(0.0)	1.0	(0.2)	15.2	(0.6)	77.2	(0.7)	6.6	(0.3)	0.0	c
Japan	0.0	c	0.0	c	0.0	c	100.0	(0.0)	0.0	c	0.0	c
Korea	0.0	c	0.0	c	9.1	(0.8)	90.4	(0.8)	0.5	(0.1)	0.0	c
Latvia	0.9	(0.2)	11.7	(0.5)	84.4	(0.6)	2.9	(0.3)	0.0	(0.0)	0.0	c
Luxembourg	0.3	(0.1)	7.9	(0.1)	50.9	(0.1)	40.3	(0.1)	0.6	(0.0)	0.0	c
Mexico	2.3	(0.3)	4.8	(0.4)	31.9	(1.4)	60.3	(1.6)	0.5	(0.1)	0.2	(0.0)
Netherlands	0.1	(0.0)	2.8	(0.3)	41.6	(0.6)	54.8	(0.6)	0.8	(0.2)	0.0	(0.0)
New Zealand	0.0	c	0.0	c	0.0	(0.0)	6.2	(0.3)	88.8	(0.5)	5.0	(0.5)
Norway	0.0	c	0.0	c	0.6	(0.1)	99.3	(0.2)	0.1	(0.1)	0.0	c
Poland	0.6	(0.1)	4.9	(0.3)	93.8	(0.4)	0.6	(0.2)	0.0	c	0.0	c
Portugal	3.2	(0.3)	8.4	(0.5)	22.9	(0.9)	65.1	(1.2)	0.4	(0.1)	0.0	c
Slovak Republic	2.2	(0.4)	4.6	(0.4)	42.6	(1.3)	50.6	(1.2)	0.1	(0.0)	0.0	c
Slovenia	0.0	c	0.3	(0.1)	4.8	(0.3)	94.6	(0.4)	0.3	(0.1)	0.0	c
Spain	0.1	(0.0)	8.6	(0.5)	23.4	(0.6)	67.9	(0.9)	0.1	(0.1)	0.0	c
Sweden	0.1	(0.1)	3.1	(0.4)	94.9	(0.8)	1.8	(0.7)	0.1	(0.1)	0.0	c
Switzerland	0.5	(0.1)	11.8	(0.7)	61.3	(1.2)	25.9	(1.3)	0.5	(0.1)	0.0	(0.0)
Turkey	0.6	(0.1)	2.6	(0.4)	20.7	(1.0)	72.9	(1.2)	3.0	(0.3)	0.1	(0.0)
United Kingdom	0.0	c	0.0	c	0.0	c	1.6	(0.3)	97.4	(0.4)	1.0	(0.3)
United States	0.0	(0.0)	0.5	(0.3)	9.6	(0.7)	72.4	(0.9)	17.3	(0.6)	0.1	(0.0)
Partners												
Albania	0.2	(0.1)	1.0	(0.2)	35.8	(2.3)	61.7	(2.3)	1.2	(0.7)	0.0	(0.0)
Algeria	18.8	(1.0)	23.5	(1.1)	35.1	(1.5)	19.4	(2.1)	3.2	(0.7)	0.0	c
Brazil	3.5	(0.2)	6.4	(0.4)	12.5	(0.5)	35.9	(0.9)	39.2	(0.8)	2.5	(0.2)
B-S-J-G (China)	1.1	(0.2)	9.2	(0.7)	52.7	(1.7)	34.6	(2.0)	2.2	(0.5)	0.1	(0.0)
Bulgaria	0.5	(0.2)	3.0	(0.6)	92.2	(0.8)	4.3	(0.4)	0.0	c	0.0	c
Colombia	5.3	(0.4)	12.3	(0.6)	22.7	(0.6)	40.2	(0.7)	19.5	(0.6)	0.0	c
Costa Rica	6.2	(0.7)	14.0	(0.7)	33.0	(1.2)	46.5	(1.6)	0.2	(0.1)	0.1	(0.1)
Croatia	0.0	c	0.2	(0.2)	79.2	(0.5)	20.6	(0.4)	0.0	c	0.0	c
Cyprus*	0.0	c	0.3	(0.0)	5.8	(0.1)	93.1	(0.1)	0.7	(0.1)	0.0	c
Dominican Republic	7.1	(0.8)	13.8	(1.2)	20.6	(0.8)	41.9	(1.1)	14.2	(0.7)	2.4	(0.3)
FYROM	0.1	(0.1)	0.1	(0.1)	70.2	(0.2)	29.7	(0.2)	0.0	c	0.0	c
Georgia	0.1	(0.0)	0.8	(0.2)	22.0	(0.8)	76.0	(0.9)	1.1	(0.3)	0.0	c
Hong Kong (China)	1.1	(0.1)	5.6	(0.4)	26.0	(0.7)	66.7	(0.7)	0.6	(0.5)	0.0	c
Indonesia	2.1	(0.3)	8.1	(0.7)	42.1	(1.5)	45.5	(1.6)	2.3	(0.4)	0.0	(0.0)
Jordan	0.2	(0.1)	0.6	(0.1)	6.6	(0.4)	92.6	(0.4)	0.0	c	0.0	c
Kosovo	0.0	(0.1)	0.6	(0.1)	24.9	(0.8)	72.4	(0.9)	2.1	(0.2)	0.0	c
Lebanon	3.7	(0.5)	8.3	(0.8)	16.6	(1.1)	62.3	(1.4)	9.0	(0.8)	0.1	(0.1)
Lithuania	0.1	(0.0)	2.6	(0.2)	86.3	(0.4)	11.0	(0.4)	0.0	(0.0)	0.0	c
Macao (China)	2.9	(0.1)	12.2	(0.2)	29.7	(0.2)	54.5	(0.1)	0.6	(0.1)	0.0	c
Malta	0.0	c	0.0	c	0.3	(0.1)	6.1	(0.2)	93.6	(0.1)	0.1	(0.0)
Moldova	0.2	(0.1)	7.6	(0.5)	84.5	(0.8)	7.5	(0.8)	0.0	(0.0)	0.0	c
Montenegro	0.0	c	0.0	c	83.7	(0.1)	16.3	(0.1)	0.0	c	0.0	c
Peru	2.5	(0.3)	6.6	(0.4)	15.9	(0.5)	50.2	(0.8)	24.8	(0.8)	0.0	c
Qatar	0.9	(0.1)	3.5	(0.1)	16.3	(0.1)	60.7	(0.1)	18.0	(0.1)	0.6	(0.0)
Romania	1.4	(0.3)	8.9	(0.5)	74.8	(0.9)	14.9	(0.7)	0.0	c	0.0	c
Russia	0.2	(0.1)	6.6	(0.3)	79.7	(1.5)	13.4	(1.5)	0.1	(0.0)	0.0	c
Singapore	0.0	(0.0)	1.9	(0.3)	7.9	(0.8)	90.0	(1.0)	0.1	(0.0)	0.1	(0.0)
Chinese Taipei	0.0	c	0.0	c	35.4	(0.7)	64.6	(0.7)	0.0	c	0.0	c
Thailand	0.2	(0.1)	0.6	(0.2)	23.8	(1.0)	72.9	(1.0)	2.4	(0.4)	0.0	c
Trinidad and Tobago	3.3	(0.2)	10.8	(0.3)	27.3	(0.3)	56.5	(0.3)	2.2	(0.2)	0.0	c
Tunisia	4.3	(0.3)	10.6	(0.8)	19.6	(1.3)	60.9	(1.7)	4.6	(0.4)	0.0	c
United Arab Emirates	0.6	(0.1)	2.5	(0.3)	10.6	(0.7)	53.4	(0.8)	31.4	(0.8)	1.5	(0.1)
Uruguay	7.5	(0.6)	9.7	(0.5)	20.7	(0.7)	61.3	(1.2)	0.8	(0.1)	0.0	c
Viet Nam	0.3	(0.1)	1.7	(0.4)	7.7	(1.8)	90.4	(2.2)	0.0	(0.0)	0.0	c
Argentina**	1.6	(0.4)	9.7	(0.8)	27.4	(1.2)	58.5	(1.6)	2.8	(0.3)	0.0	c
Kazakhstan**	0.1	(0.1)	2.7	(0.3)	60.4	(1.7)	36.2	(1.8)	0.6	(0.1)	0.0	c
Malaysia**	0.0	c	0.0	c	3.2	(0.6)	96.4	(0.7)	0.4	(0.3)	0.0	c

* See note at the beginning of this Annex.

** Coverage is too small to ensure comparability (see Annex A4).

StatLink  <http://dx.doi.org/10.1787/888933433129>

ANNEX A3

TECHNICAL NOTES ON ANALYSES IN THIS VOLUME

Methods and definitions

Relative risk

The relative risk is a measure of the association between an antecedent factor and an outcome factor. The relative risk is simply the ratio of two risks, i.e. the risk of observing the outcome when the antecedent is present and the risk of observing the outcome when the antecedent is not present. Figure A3.1 presents the notation that is used in the following.

Figure A3.1 ■ Labels used in a two-way table

P_{11}	P_{12}	$P_{1.}$
P_{21}	P_{22}	$P_{2.}$
$P_{.1}$	$P_{.2}$	$P_{..}$

P_{ij} represents the probabilities for each cell and is equal to the number of observations in a particular cell divided by the total number of observations. $P_{i.}$, $P_{.j}$ respectively represent the marginal probabilities for each row and for each column. The marginal probabilities are equal to the marginal frequencies divided by the total number of students.

Assuming that rows represent the antecedent factor, with the first row for “having the antecedent” and the second row for “not having the antecedent”, and that the columns represent the outcome: the first column for “having the outcome” and the second column for “not having the outcome”. The relative risk is then equal to:

$$RR = \frac{(P_{11}/P_{1.})}{(P_{21}/P_{2.})}$$

Odds ratio

The same notation can be used to define the odds ratio, another measure of the relative likelihood of a particular outcome across two groups. The odds ratio for observing the outcome when an antecedent is present is simply

$$OR = \frac{(P_{11}/P_{12})}{(P_{21}/P_{22})}$$

where P_{11}/P_{12} represents the “odds” of observing the outcome when the antecedent is present, and P_{21}/P_{22} represents the “odds” of observing the outcome when the antecedent is not present.

Logistic regression can be used to estimate the odds ratio: the exponentiated logit coefficient for a binary variable is equivalent to the odds ratio. A “generalised” odds ratio, after accounting for other differences across groups, can be estimated by introducing control variables in the logistic regression.

Statistics based on multilevel models

Statistics based on multilevel models include variance components (between- and within-school variance), the index of inclusion derived from these components, and regression coefficients where this has been indicated. Multilevel models are generally specified as two-level regression models (the student and school levels), with normally distributed residuals, and estimated with maximum likelihood estimation. Where the dependent variable is science, reading or mathematics performance, the estimation uses ten plausible values for each student’s performance on the mathematics scale. Models were estimated using the Stata® (version 14.1) “mixed” module.

In multilevel models, weights are used at both the student and school levels. The purpose of these weights is to account for differences in the probabilities of students being selected in the sample. Since PISA applies a two-stage sampling procedure, these differences are due to factors at both the school and the student levels. For the multilevel models, student final weights (W_FSTUWT) were used. Within-school weights correspond to student final weights, rescaled to amount to the sample size within each school. Between-school weights correspond to the sum of final student weights (W_FSTUWT) within each school. The definition of between-school weights is the same as in PISA 2012 initial reports.



The index of inclusion is defined and estimated as:

$$100 * \frac{\sigma_w^2}{\sigma_w^2 + \sigma_b^2}$$

where σ_w^2 and σ_b^2 , respectively, represent the within- and between-variance estimates.

The results in multilevel models, and the between-school variance estimate in particular, depend on how schools are defined and organised within countries and by the units that were chosen for sampling purposes. For example, in some countries, some of the schools in the PISA sample were defined as administrative units (even if they spanned several geographically separate institutions, as in Italy); in others they were defined as those parts of larger educational institutions that serve 15-year-olds; in still others they were defined as physical school buildings; and in others they were defined from a management perspective (e.g. entities having a principal). The *PISA 2015 Technical Report* (OECD, forthcoming) and Annex A2 provide an overview of how schools are defined. In Slovenia, for example, the primary sampling unit is defined as a group of students who follow the same study programme within a school (an education track within a school). So in this case, the between-school variation is actually the within-school, between-track difference. The use of stratification variables in the selection of schools may also affect the estimate of the between-school variation, particularly if stratification variables are associated with between-school differences.

Because of the manner in which students were sampled, the within-school variation includes variation between classes as well as between students.

Effect sizes

Sometimes it is useful to compare differences in an index between groups, such as boys and girls, across countries. A problem that may occur in such instances is that the distribution of the index varies across groups or countries. One way to resolve this is to calculate an effect size that accounts for differences in the distributions. An effect size measures the difference between, say, the self-efficacy in science of male and female students in a given country, relative to the average variation in self-efficacy in science among all students in the country.

In accordance with common practices, Table I.3.6 reports effect sizes of less than 0.20 as small, effect sizes on the order of 0.50 as medium, and effect sizes greater than 0.80 as large.

The effect size between two subgroups is calculated as:

$$\frac{m_1 - m_2}{\sqrt{\sigma^2}}$$

where m_1 and m_2 , respectively, represent the mean values for the subgroups 1 and 2 and σ^2 represents the overall (between and within-group) variance.

Concentration indices

Index of current concentration

The country/economy-level index of current concentration of immigrant students in schools (or current concentration index) corresponds to the minimum share of students, both immigrant and non-immigrant, who would have to be relocated from one school to another if all schools were to have an identical share of immigrant students and, consequently, an identical share of non-immigrant students. It is defined as

$$CC = \frac{\sum_{i=1}^l N_i |p_i - p|}{N}$$

with N_i equal to the number of students in school i , N equal to the number of students in the population, l equal to the number of schools. $p_i = A_i/N_i$ is the share of immigrant students in school i and $p = A/N$ is the share of immigrant students in the population.

The current concentration index S is related to the segregation index developed by Gorard and Taylor (2002), which corresponds to the percentage of immigrant students who would have to be relocated from one school to another if all schools were to have an identical share of immigrant students, given the initial size of the schools. Gorard and Taylor's segregation index is defined as:

$$S = 0.5 \times \sum_{i=1}^l \left| \frac{A_i}{A} - \frac{N_i}{N} \right|$$

The current concentration index can be directly derived from the segregation index as $CC = 2pS$. Gorard and Taylor's segregation index is highly dependent on the percentage of immigrants in the population. If the country has very few immigrants and if these immigrants are mostly enrolled in one international school, then the percentage of immigrants to be moved would be close to 100%. The current concentration index is less sensitive to this extreme case, but remains sensitive to the overall percentage of immigrants in the population.



When the current concentration index is computed from a representative sample it is important to take sampling weights and sampling error into account. The current concentration index can be rewritten as an average across students,

$$\frac{\sum_{i=1}^I N_i |p_i - p|}{N} = \frac{1}{N} \sum_{i=1}^I N_i |p_i - p| = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{N_i} |p_i - p|.$$

It can therefore be readily generalised to weighted samples, simply by replacing the latter expression by a weighted average:

$$\frac{\sum_{i=1}^I \sum_{j=1}^{n_i} w_{ij} |p_i - p|}{\sum_{i=1}^I \sum_{j=1}^{n_i} w_{ij}}.$$

The current concentration index can then be computed, at the student level, as the absolute difference between the school percentage of immigrants and the national percentage of immigrants weighted by the final student weight. Standard errors for the index are obtained by replacing the final weight by the 80 weight replicates in the computation.

Index of maximum concentration

The index of maximum concentration is a theoretical maximum of the concentration of immigrant students in schools, given the size of schools and the number of immigrants in a country. It corresponds to the minimum share of students, both immigrant and non-immigrant, who would have to be relocated from one school to another if all schools were to have an identical share of immigrant students, in the counterfactual situation in which all immigrant students were located in the largest schools to begin with. In this hypothetical scenario, the concentration is maximal in the sense that immigrant students are present only in the smallest possible number of schools (given the size of schools and the immigrant population).

The computation of the index requires, first, to sort schools in each country in descending order by their respective school weight (computed as the sum of the final student weights in that school). In a second step, all immigrant students are allocated to the schools according to this sorting, up to the weighted sum of immigrant students in that particular country. The concentration index defined above is then computed. Standard errors for the index are obtained by replacing in the computation the final weight by the 80 weight replicates.

Definition of low- and high-concentration schools

The classification of schools as having either a low or a high concentration of immigrant students is based on a cutpoint that is specific to each country/economy, so that the number of low- and high-concentration schools is not dependent on the share of immigrant students in each education system. The cutpoint is defined as the (weighted) median of the distribution of shares of immigrant students across schools. In each country, approximately 50% of students are in high-concentration schools and 50% of students are in low-concentration schools.

Standard errors and significance tests

The statistics in this report represent estimates of national performance based on samples of students, rather than values that could be calculated if every student in every country had answered every question. Consequently, it is important to measure the degree of uncertainty of the estimates. In PISA, each estimate has an associated degree of uncertainty, which is expressed through a standard error. The use of confidence intervals provides a way to make inferences about the population means and proportions in a manner that reflects the uncertainty associated with the sample estimates. From an observed sample statistic and assuming a normal distribution, it can be inferred that the corresponding population result would lie within the confidence interval in 95 out of 100 replications of the measurement on different samples drawn from the same population.

In many cases, readers are primarily interested in whether a given value in a particular country is different from a second value in the same or another country, e.g. whether girls in a country perform better than boys in the same country. In the tables and charts used in this report, differences are labelled as statistically significant when a difference of that size smaller or larger in absolute value would be observed less than 5% of the time, if there were actually no difference in corresponding population values.

Throughout the report, significance tests were undertaken to assess the statistical significance of the comparisons made.

Gender differences and differences between subgroup means

Gender differences in student performance or other indices were tested for statistical significance. Positive differences indicate higher scores for boys while negative differences indicate higher scores for girls. Generally, differences marked in bold in the tables in this volume are statistically significant at the 95% confidence level.



Similarly, differences between other groups of students (e.g. non-immigrant students and students with an immigrant background) or categories of schools (e.g. advantaged and disadvantaged schools) were tested for statistical significance. The definitions of the subgroups can, in general, be found in the tables and the text accompanying the analysis. Socio-economically (dis) advantaged school are defined as schools in the (bottom) top quarter of the distribution of the average PISA index of economic, social and cultural status (ESCS) across schools within each country/economy. All differences marked in bold in the tables presented in Annex B of this report are statistically significant at the 95% level.

Differences between subgroup means, after accounting for other variables

For many tables, subgroup comparisons were performed both on the observed difference (“before accounting for other variables”) and after accounting for other variables, such as the PISA index of economic, social and cultural status of students. The adjusted differences were estimated using linear regression and tested for significance at the 95% confidence level. Significant differences are marked in bold.

Performance differences between the top and bottom quartiles of PISA indices and scales

Differences in average performance between the top and bottom quarters of the PISA indices and scales were tested for statistical significance. Figures marked in bold indicate that performance between the top and bottom quarters of students on the respective index is statistically significantly different at the 95% confidence level.

Change in the performance per unit of the index

For many tables, the difference in student performance per unit on the index shown was calculated. Figures in bold indicate that the differences are statistically significantly different from zero at the 95% confidence level.

Relative risk and odds ratio

Figures in bold in the data tables presented in Annex B of this report indicate that the relative risk/odds ratio is statistically significantly different from 1 at the 95% confidence level. To compute statistical significance around the value of 1 (the null hypothesis), the relative-risk/odds-ratio statistic is assumed to follow a log-normal distribution, rather than a normal distribution, under the null hypothesis.

For many tables, “generalised” odds ratios (after accounting for other variables) are also presented. These odds ratios were estimated using logistic regression and tested for significance against the null hypothesis of an odds ratio equal to 1 (i.e. equal likelihoods, after accounting for other variables).

Range of ranks

To calculate the range of ranks for countries, data are simulated using the mean and standard error of the mean for each relevant country to generate a distribution of possible values. Some 10 000 simulations are implemented and, based on these values, 10 000 possible rankings for each country are produced. For each country, the counts for each rank are aggregated from largest to smallest until they equal 9 500 or more. Then the range of ranks per country is reported, including all the ranks that have been aggregated. This means that there is at least 95% confidence about the range of ranks, and it is safe to assume unimodality in this distribution of ranks. This method has been used in all cycles of PISA since 2003, including PISA 2015.

The main difference between the range of ranks (e.g. Figure I.2.14) and the comparison of countries’ mean performance (e.g. Figure I.2.13) is that the former takes account of the multiple comparisons involved in ranking countries/economies, while the latter does not. Therefore, sometimes there is a slight difference between the range of ranks and counting the number of countries above a given country, based on pairwise comparisons of the selected countries’ performance. For instance, Beijing, Shanghai, Jiangsu, Guangdong (China) (hereafter B-S-J-G [China]) and Korea have similar mean performance and the same set of countries whose mean score is not statistically different from theirs, based on Figure I.2.13; but the rank for Korea can be restricted to be, with 95% confidence, between 9th and 14th, while the range of ranks for B-S-J-G (China) is wider (between 8th and 16th) (Figure I.2.14). Since it is safe to assume that the distribution of rank estimates for each country has a single mode (unimodality), the results of range of ranks for countries should be used when examining countries’ rankings.

Standard errors in statistics estimated from multilevel models

For statistics based on multilevel models (such as the estimates of variance components and regression coefficients from two-level regression models) the standard errors are not estimated with the usual replication method, which accounts for stratification and sampling rates from finite populations. Instead, standard errors are “model-based”: their computation assumes that schools, and students within schools, are sampled at random (with sampling probabilities reflected in school and student weights) from a theoretical, infinite population of schools and students which complies with the model’s parametric assumptions.

The standard error for the estimated index of inclusion is calculated by deriving an approximate distribution for it from the (model-based) standard errors for the variance components, using the delta method.



Standard errors in trend analyses of performance: link error

Standard errors for comparisons of performance across time account for the uncertainty in the equating procedure that allows scores in different PISA assessments to be expressed on the same scale. This additional source of uncertainty results in more conservative standard errors (larger than standard errors that were estimated before the introduction of this link error) (see Annex A5 for a technical discussion of the link error).

Figures in bold in the data tables for performance trends or changes presented in Annex B of this report indicate that the change in performance for that particular group is statistically significantly different from 0 at the 95% confidence level. The standard errors used to calculate the statistical significance of the reported performance trend or change include the link error.

References

Gorard, S. and C. Taylor (2002), "What is segregation ? A comparison of measures in terms of 'strong' and 'weak' compositional invariance", *Sociology*, Vol.36/4, pp. 875-895, <http://dx.doi.org/10.1177/003803850203600405>.

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



ANNEX A4

QUALITY ASSURANCE

Quality assurance procedures were implemented in all parts of PISA 2015, as was done for all previous PISA surveys. The PISA 2015 Technical Standards (www.oecd.org/pisa/) specify the way in which PISA must be implemented in each country, economy and adjudicated region. International contractors monitor the implementation in each of these and adjudicate on their adherence to the standards.

The consistent quality and linguistic equivalence of the PISA 2015 assessment instruments were facilitated by assessing the ease with which the original English version could be translated. Two source versions of the assessment instruments, in English and French were prepared (except for the financial literacy assessment and the operational manuals, which were provided only in English) in order for countries to conduct a double translation design, i.e. two independent translations from the source language(s), and reconciliation by a third person. Detailed instructions for the localisation (adaptation, translation and validation) of the instruments for the field trial and for their review for the main survey, and translation/adaptation guidelines were supplied. An independent team of expert verifiers, appointed and trained by the PISA Consortium, verified each national version against the English and/or French source versions. These translators' mother tongue was the language of instruction in the country concerned, and the translators were knowledgeable about education systems. For further information on PISA translation procedures, see the *PISA 2015 Technical Report* (OECD, forthcoming).

The survey was implemented through standardised procedures. The PISA Consortium provided comprehensive manuals that explained the implementation of the survey, including precise instructions for the work of school co-ordinators and scripts for test administrators to use during the assessment sessions. Proposed adaptations to survey procedures, or proposed modifications to the assessment session script, were submitted to the PISA Consortium for approval prior to verification. The PISA Consortium then verified the national translation and adaptation of these manuals.

To establish the credibility of PISA as valid and unbiased and to encourage uniformity in administering the assessment sessions, test administrators in participating countries were selected using the following criteria: it was required that the test administrator not be the science, reading or mathematics instructor of any students in the sessions he or she would conduct for PISA; and it was considered preferable that the test administrator not be a member of the staff of any school in the PISA sample. Participating countries organised an in-person training session for test administrators.

Participating countries and economies were required to ensure that test administrators worked with the school co-ordinator to prepare the assessment session, including reviewing and updating the Student Tracking Form; completing the Session Attendance Form, which is designed to record students' attendance and instruments allocation; completing the Session Report Form, which is designed to summarise session times, any disturbance to the session, etc.; ensuring that the number of test booklets and questionnaires collected from students tallied with the number sent to the school (paper-based assessment countries) or ensuring that the number of USB sticks used for the assessment were accounted for (computer-based assessment countries); and sending the school questionnaire, student questionnaires, parent and teacher questionnaires (if applicable), and all test materials (both completed and not completed) to the national centre after the testing.

The PISA Consortium responsible for overseeing survey operations implemented all phases of the PISA Quality Monitor (PQM) process: interviewing and hiring PQM candidates in each of the countries, organising their training, selecting the schools to visit, and collecting information from the PQM visits. PQMs are independent contractors located in participating countries who are hired by the international survey operations contractor. They visit a sample of schools to observe test administration and to record the implementation of the documented field-operations procedures in the main survey.

Typically, two or three PQMs were hired for each country, and they visited an average of 15 schools in each country. If there were adjudicated regions in a country, it was usually necessary to hire additional PQMs, as a minimum of five schools were observed in adjudicated regions.

All quality-assurance data collected throughout the PISA 2015 assessment were entered and collated in a central data-adjudication database on the quality of field operations, printing, translation, school and student sampling, and coding.



Comprehensive reports were then generated for the PISA Adjudication Group. This group was formed by the Technical Advisory Group and the Sampling Referee. Its role is to review the adjudication database and reports to recommend adequate treatment to preserve the quality of PISA data. For further information, see the *PISA 2015 Technical Report* (OECD, forthcoming).

The results of adjudication and subsequent further examinations showed that the PISA Technical Standards were met in all countries and economies that participated in PISA 2015 except for those countries listed below:

- In Albania, the PISA assessment was conducted in accordance with the operational standards and guidelines of the OECD. However, because of the ways in which the data were captured, it was not possible to match the data in the test with the data from the student questionnaire. As a result, Albania cannot be included in analyses that relate students' responses from the questionnaires to the test results.
- In Argentina, the PISA assessment was conducted in accordance with the operational standards and guidelines of the OECD. However, there was a significant decline in the proportion of 15-year-olds who were covered by the test, both in absolute and relative numbers. There had been a re-structuring of Argentina's secondary schools, except for those in the adjudicated region of Ciudad Autónoma de Buenos Aires, which is likely to have affected the coverage of eligible schools listed in the sampling frame. As a result, Argentina's results may not be comparable to those of other countries or to results for Argentina from previous years.
- In Kazakhstan, the national coders were found to be lenient in marking. Consequently, the human-coded items did not meet PISA standards and were excluded from the international data. Since human-coded items form an important part of the constructs that are tested by PISA, the exclusion of these items resulted in a significantly smaller coverage of the PISA test. As a result, Kazakhstan's results may not be comparable to those of other countries or to results for Kazakhstan from previous years.
- In Malaysia, the PISA assessment was conducted in accordance with the operational standards and guidelines of the OECD. However, the weighted response rate among the initially sampled Malaysian schools (51%) falls well short of the standard PISA response rate of 85%. Therefore, the results may not be comparable to those of other countries or to results for Malaysia from previous years.

Reference

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



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 inverted 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 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. Davidov, Schmidt and Billiet, 2011; Meredith, 1993) 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 (Masters, 1982; Rasch 1960) 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, 2014; Oliveri and von Davier, 2011).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).



A corrigendum has been issued for this page. See: <http://www.oecd.org/about/publishing/Corrigenda-PISA2015-Volumel.pdf>

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, 2014a) 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.

Table A5.1. Correlation of country means under alternative scaling approaches
Across all countries/economies that participated in PISA 2015

	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

Note: This table reports the correlation coefficient between the mean estimates included in Annex B, based on cycle-specific scaling approaches, and the means for posterior distributions produced under the 2015 scaling approach.

StatLink  <http://dx.doi.org/10.1787/888933433162>

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.

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 for science, for reading and 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 5 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 5 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.

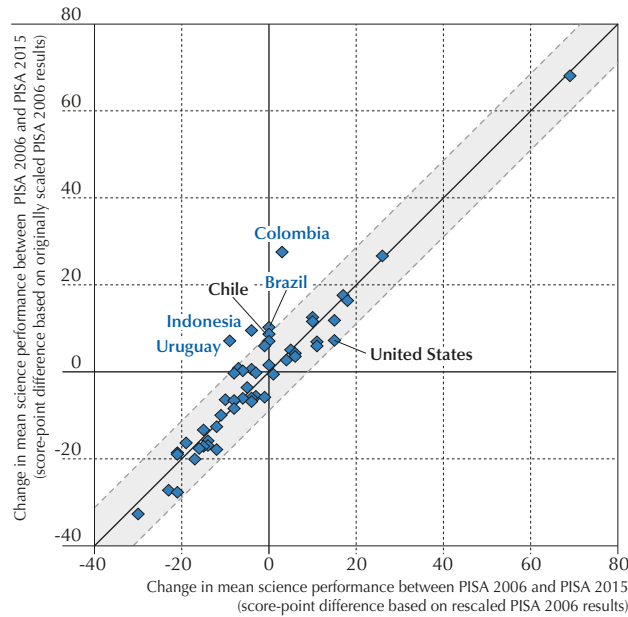
Mathematics

- 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**

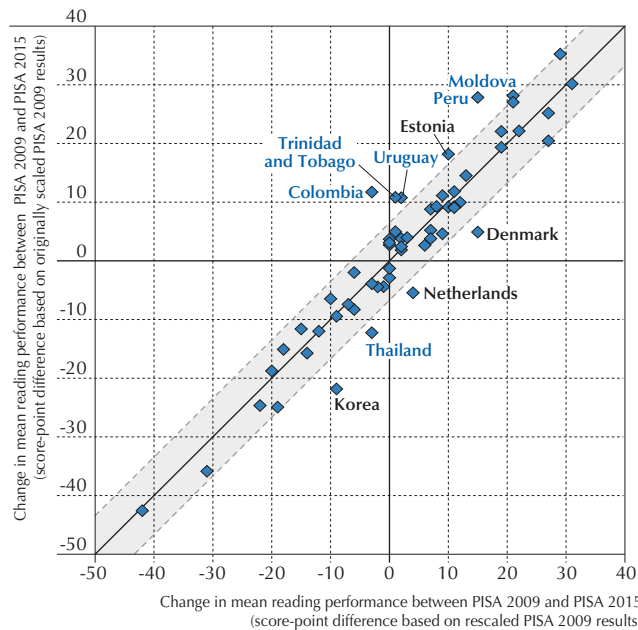


Note: The solid line indicates the diagonal, where both changes are equal. The area shaded in grey indicates the confidence interval of the diagonal, based on the link error for comparisons between originally scaled 2006 results and 2015 results (see Table A5.2).

Source: OECD, PISA 2015 Database, Tables I.2.4a and A5.3.

StatLink <http://dx.doi.org/10.1787/888933433132>

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

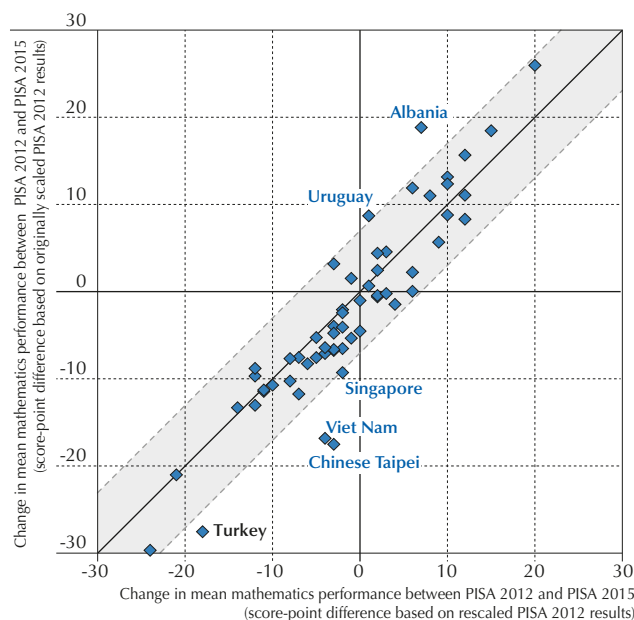


Note: The solid line indicates the diagonal, where both changes are equal. The area shaded in grey indicates the confidence interval of the diagonal, based on the link error for comparisons between originally scaled 2009 results and 2015 results (see Table A5.2).

Source: OECD, PISA 2015 Database, Tables I.4.4a and A5.4.


StatLink <http://dx.doi.org/10.1787/888933433149>

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



Note: The solid line indicates the diagonal, where both changes are equal. The area shaded in grey indicates the confidence interval of the diagonal, based on the link error for comparisons between originally scaled 2012 results and 2015 results (see Table A5.2).

Source: OECD, PISA 2015 Database, Tables I.5.4a and A5.5.

StatLink  <http://dx.doi.org/10.1787/888933433156>

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 subsamples 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 modes 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, 65 items in reading and 51 items in mathematics.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.

A corrigendum has been issued for this page. See: <http://www.oecd.org/about/publishing/Corrigenda-PISA2015-Volumel.pdf>


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 , and) 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

Comparison	Science	Reading	Mathematics
PISA 2000 to 2015		6.8044	
PISA 2003 to 2015		5.3907	5.6080
PISA 2006 to 2015	4.4821	6.6064	3.5111
PISA 2009 to 2015	4.5016	3.4301	3.7853
PISA 2012 to 2015	3.9228	5.2535	3.5462

Note: Comparisons between PISA 2015 scores and previous assessments can only be made when the subject first became a major domain. As a result, comparisons of science performance between PISA 2000 and PISA 2015, for example, are not possible.

StatLink  <http://dx.doi.org/10.1787/888933433162>

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 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 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, 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 A5.2 as well as by those between previous PISA reporting scales, published in Table 12.31 of the *PISA 2012 Technical Report* (OECD, 2014b). 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 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.

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:

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



A corrigendum has been issued for this page. See: <http://www.oecd.org/about/publishing/Corrigenda-PISA2015-Volumel.pdf>

where Δ_{2015-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 $\sigma(\Delta_{2015-t})$ is:

$$\sigma(\Delta_{2015-t}) = \sqrt{\sigma_{2015}^2 + \sigma_t^2 + error_{2015,t}^2}$$

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:

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

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 .

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 (). 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.

A corrigendum has been issued for this page. See: <http://www.oecd.org/about/publishing/Corrigenda-PISA2015-Volumel.pdf>

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 1.6.1 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 reweighting 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.

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.

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 $\Delta_{2015,t}$ 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 $\sigma(\Delta_{2015-2006})$ is:

$$\sigma(\Delta_{2015-2006}) = \sqrt{\sigma_{2015}^2 + \sigma_{2006}^2}$$

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 underestimated. 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 (StatLink <http://dx.doi.org/10.1787/888933433162>)

- 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

Notes

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, 2014a). 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.

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. Jhangir (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 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 (2014b), *PISA 2012 Technical Report*, OECD Publishing, Paris, <http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>.
- 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” *Psychological Test and Assessment Modeling*, Vol. 53/1, pp. 315-33.



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

Rousseuw, 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>.

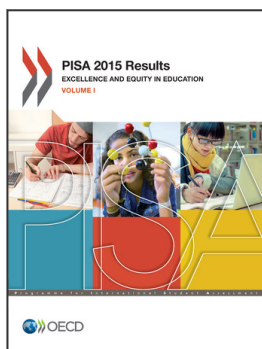


ANNEX A6

THE PISA 2015 FIELD TRIAL MODE-EFFECT STUDY

Available on line only.

It can be found at: www.oecd.org/pisa



From:
PISA 2015 Results (Volume I)
Excellence and Equity in Education

Access the complete publication at:
<https://doi.org/10.1787/9789264266490-en>

Please cite this chapter as:

OECD (2016), "PISA 2015 technical background", in *PISA 2015 Results (Volume I): Excellence and Equity in Education*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/9789264266490-13-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 and any 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.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.