ANNEX A1
The construction of proficiency scales and of indices from the student context questionnaire

PROFICIENCY SCALES FOR READING, MATHEMATICS AND SCIENCE

Proficiency scores in reading, mathematics and science are based on student responses to items that represent the assessment framework for each domain (see Chapter 2). While different students saw different questions, the test design, which ensured a significant overlap of items across different forms, made it possible to construct proficiency scales that are common to all students for each domain. In general, the PISA frameworks assume that a single continuous scale can be used to report overall proficiency in a domain; but this assumption is further verified during scaling (see below).

PISA proficiency scales are constructed using item-response-theory models, in which the likelihood that the test-taker responds correctly to any question is a function of the question's characteristics (see below) and of the test-taker's position on the scale. In other words, the test-taker's proficiency is associated with a particular point on the scale that indicates the likelihood that he or she responds correctly to any question. Higher values on the scale indicate greater proficiency, which is equivalent to a greater likelihood of responding correctly to any question. A description of the modelling technique used to construct proficiency scales can be found in the *PISA 2018 Technical Report* (OECD, forthcoming[1]).

In the item-response-theory models used in PISA, the task characteristics are summarised by two parameters that represent task difficulty and task discrimination. The first parameter, task difficulty, is the point on the scale where there is at least a 50% probability of a correct response by students who score at or above that point; higher values correspond to more difficult items. For the purpose of describing proficiency levels that represent mastery, PISA often reports the difficulty of a task as the point on the scale where there is at least a 62% probability of a correct response by students who score at or above that point.[1]

The second parameter, task discrimination, represents the rate at which the proportion of correct responses increases as a function of student proficiency. For an idealised highly discriminating item, close to 0% of students respond correctly if their proficiency is below the item difficulty, and close to 100% of students respond correctly as soon as their proficiency is above the item difficulty. In contrast, for weakly discriminating items, the probability of a correct response still increases as a function of student proficiency, but only gradually.

A single continuous scale can therefore show both the difficulty of questions and the proficiency of test-takers (see Figure I.2.1). By showing the difficulty of each question on this scale, it is possible to locate the level of proficiency in the domain that the question demands. By showing the proficiency of test-takers on the same scale, it is possible to describe each test-taker's level of skill or literacy by the type of tasks that he or she can perform correctly most of the time.

Estimates of student proficiency are based on the kinds of tasks that students are expected to perform successfully. This means that students are likely to be able to successfully answer questions located at or below the level of difficulty associated with their own position on the scale. Conversely, they are unlikely to be able to successfully answer questions above the level of difficulty associated with their position on the scale.[2]

The higher a student's proficiency level is located above a given test question, the more likely is he or she to be able to answer the question successfully. The discrimination parameter for this particular test question indicates how quickly the likelihood of a correct response increases. The further the student's proficiency is located below a given question, the less likely is he or she to be able to answer the question successfully. In this case, the discrimination parameter indicates how fast this likelihood decreases as the distance between the student's proficiency and the question's difficulty increases.

How many scales per domain? Assessing the dimensionality of PISA domains

PISA frameworks for reading, mathematics and science assume that a single continuous scale can summarise performance in each domain for all countries. This assumption is incorporated in the item-response-theory model used in PISA. Violations of this assumption therefore result in model misfit, and can be assessed by inspecting fit indices.

After the field trial, initial estimates of model fit for each item, and for each country and language group, provide indications about the plausibility of the uni-dimensionality assumption and about the equivalence of scales across countries. These initial estimates are used to refine the item set used in each domain: problematic items are sometimes corrected (e.g. if a translation error is

detected); and coding and scoring rules can be amended (e.g. to suppress a partial-credit score that affected coding reliability, or to combine responses to two or more items when the probability of a correct response to one question appears to depend on the correct answer to an earlier question). Items can also be deleted after the field trial. Deletions are carefully balanced so that the set of retained items continues to provide a good balance of all aspects of the framework.

After the main study, the estimates of model fit are mainly used to refine the scaling model (some limited changes to the scoring rules and item deletions can also be considered). In response to earlier criticisms (Kreiner and Christensen, 2013[2]; Oliveri and von Davier, 2013[3]) and to take advantage of the increased computational resources available, PISA, in its 2015 cycle, moved to a more flexible item-response-theory model. This model allows items to vary not only in difficulty, but in their ability to discriminate between high and low performance. It also assigns country- and language-specific characteristics to items that do not fit the model for the particular item and language (see Annex A6 and OECD, forthcoming[1]). This "tailoring" of the measurement model makes it possible to improve model fit considerably, while retaining the desired level of comparability across countries and the interpretation of scales through a single set of proficiency descriptors.

With the 2015 assessment, PISA also introduced an additional test of dimensionality to confirm that "trend" and "new" items can be reported on the same scale. Using the international dataset, this test compares fit statistics for a model assuming uni-dimensionality with fit statistics for a model that assumes that "trend" and "new" items represent two distinct continuous traits. In 2015, for science, and then again in 2018, for reading, this test confirmed that a uni-dimensional model for "trend" and "new" items fits the data almost as well as a two-dimensional model, and that a uni-dimensional scale is more reliable than separate scales for "trend" and "new" items. This evidence was interpreted as showing that a single coherent scale can represent the constructs of science and reading in 2018 (OECD, forthcoming[1]).

Despite the evidence in favour of a uni-dimensional scale, for the "major" domain (i.e. reading in PISA 2018) PISA nevertheless provides multiple estimates of performance, in addition to the overall scale, through so-called "subscales". Subscales represent different framework dimensions and provide a more nuanced picture of performance in a domain. Subscales within a domain are usually highly correlated across students (thus supporting the assumption that a coherent overall scale can be formed by combining items across subscales). Despite this high correlation, interesting differences in performance across subscales can often be observed at aggregate levels (across countries, across education systems within countries, or between boys and girls).

How reporting scales are set and linked across multiple assessments

The reporting scale for each domain was originally established when the domain was the major focus of assessment in PISA for the first time: PISA 2000 for reading, PISA 2003 for mathematics and PISA 2006 for science.

The item-response-theory models used in PISA describe the relationship between student proficiency, item difficulty and item discrimination, but do not set a measurement unit for any of these parameters. In PISA, this measurement unit is chosen the first time a reporting scale is established. The score of "500" on the scale is defined as the average proficiency of students across OECD countries; "100 score points" is defined as the standard deviation (a measure of the variability) of proficiency across OECD countries.[3]

To enable the measurement of trends, achievement data from successive assessments are reported on the same scale. It is possible to report results from different assessments on the same scale because in each assessment PISA retains a significant number of items from previous PISA assessments. These are known as trend items. All items used to assess mathematics and science in 2018, and a significant number of items used to assess reading (72 out of 244), were developed and already used in earlier assessments (see Tables I.A5.1 and I.A5.3). Their difficulty and discrimination parameters were therefore already estimated in previous PISA assessments.

The answers to the trend questions from students in earlier PISA cycles, together with the answers from students in PISA 2018, were both considered when scaling PISA 2018 data to determine student proficiency, item difficulty and item discrimination. In particular, when scaling PISA 2018 data, item parameters for new items were freely estimated, but item parameters for trend items were initially fixed to their PISA 2015 values, which, in turn, were based on a concurrent calibration involving response data from multiple cycles (OECD, 2017[4]). All constraints on trend item parameters were evaluated and, in some cases, released in order to better describe students' response patterns. See the *PISA 2018 Technical Report* (OECD, forthcoming[1]) for details.

The extent to which the item characteristics estimated during the scaling of PISA 2018 data differ from those estimated in previous calibrations is summarised in the "link error", a quantity (expressed in score points) that reflects the uncertainty in comparing PISA results over time. A link error of zero indicates a perfect match in the parameters across calibrations, while a non-zero link error indicates that the relative difficulty of certain items or the ability of certain items to discriminate between high and low achievers has changed over time, introducing greater uncertainty in trend comparisons.

INDICES FROM THE STUDENT CONTEXT QUESTIONNAIRE

In addition to scale scores representing performance in reading, mathematics and science, this volume uses indices derived from the PISA student questionnaires to contextualise PISA 2018 results or to estimate trends that account for demographic changes over time. The following indices and database variables are used:

- Student age (database variable: AGE)
- Student gender (ST004)
- Immigrant background (IMMIG)
- Language spoken at home (ST022)
- The PISA index of economic, social and cultural status (ESCS)

For a description of how these indices were constructed, see Annex A1 in *PISA 2018 Results (Volume II): Where all Students Can Succeed* (OECD, 2019[5]) and the *PISA 2018 Technical Report* (OECD, forthcoming[1]).

Chapter 1 also reports changes, over time, in time spent using the Internet (2012 and 2018), in the proportion of students having access to various digital devices, in the number of devices available at home, and in students' reading habits and attitudes towards reading (2009 and 2018).

Most of these analyses report proportions of particular answer categories in the student questionnaire or in the ICT familiarity questionnaire, which was optional for countries. In a few cases, some answer categories were combined (e.g. "agree" and "strongly agree") prior to conducting the analysis; these simple recodes are indicated in column headers and in notes under Tables I.B1.54-I.B1.59.

In addition, three indices were used for the analysis of time spent using the Internet, in Tables I.B1.51-I.B1.53. The indices of time spent using the Internet were constructed from students' answers to the following questions, which were included in the optional ICT familiarity questionnaire:

- During a typical weekday, for how long do you use the Internet at school? (IC005)
- During a typical weekday, for how long do you use the Internet outside of school? (IC006)
- On a typical weekend day, for how long do you use the Internet outside of school? (IC007)

Students were allowed to respond in intervals of: no time; between 1-30 minutes per day; between 31-60 minutes per day; between 1 hour and 2 hours per day; between 2 hours and 4 hours per day; between 4 hours and 6 hours per day; and more than 6 hours per day. To build the indices of time spent using the Internet, these responses were converted to the smallest number of minutes in the interval (0, 1, 31, 61, 121, 241 or 361, respectively). As such, the indices represent lower bounds of the time spent on the Internet reported by each student.

Notes

1. This definition of task difficulty, referred to as RP62 in the *PISA 2018 Technical Report* (OECD, forthcoming[1]), is used in particular to classify assessment items into proficiency levels (see Chapter 5). The choice of a probability of 62%, rather than of 50%, sets the bar for mastery of a particular level of proficiency significantly above chance levels, including for simple multiple-choice response formats. In the typical parametrisation of the two-parameters IRT-model used by PISA, RP62 values depend on both model parameters.
2. "Unlikely", in this context, refers to a probability below 62%.
3. The standard deviation of 100 score points corresponds to the standard deviation in a pooled sample of students from OECD countries, where each national sample is equally weighted.

References

Kreiner, S. and **K. Christensen** (2013), "Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy", *Psychometrika*, Vol. 79/2, pp. 210-231, http://dx.doi.org/10.1007/s11336-013-9347-z. [2]

OECD (2019), *PISA 2018 Results (Volume II): Where All Students Can Succeed*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b5fd1b8f-en. [5]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, http://www.oecd.org/pisa/data/2015-technical-report/ (accessed on 31 July 2017). [4]

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

Oliveri, M. and **M. von Davier** (2013), "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. [3]