



15

Proficiency scale construction

Introduction	276
Development of the described scales	277
Defining the proficiency levels	279
Reporting the results for pisa science	281



INTRODUCTION

This chapter discusses the methodology used to develop the PISA reporting scales, which describe levels of proficiency in the different PISA domains, and presents the outcomes of the development process for science literacy, the major domain in PISA 2015.

The reporting scales are called “proficiency scales” rather than “performance scales” because they describe what students *typically* know and can do at given levels of proficiency, rather than how individuals who were tested *actually* performed on a single test administration. This emphasis reflects the primary goal of PISA, which is to report general population-level results rather than the results for individual students. PISA uses samples of students and items to make estimates about populations. A sample of 15-year-old students is selected to represent all 15-year-olds in a country and a sample of test items from a large pool is administered to each student. Results are then analysed using statistical models that estimate the likely proficiency of the population, based on this sampling.

The PISA test design makes it necessary to use techniques of modern item response modelling (see Chapter 9) to both estimate the ability of all students taking the PISA assessment and the statistical characteristics of all PISA items.

The PISA data are collected using a rotated test design in which students take different but overlapping tasks. The mathematical model employed to analyse the PISA data is implemented through test analysis software that uses iterative procedures to simultaneously estimate the distribution of students along the proficiency dimension assessed by the test, as well as a mathematical function that describes the association of student proficiency and the likelihood of a correct response for each item on the test. The result of these procedures is a set of item parameters that represents, among other things, locations on a proficiency continuum reflecting the domain being assessed. On that continuum, it is possible to estimate the distribution of groups of students, and thereby the average (location) and range (variability) of their skills and knowledge in this domain. This continuum represents the overall PISA scale in the relevant test domain, such as reading, mathematics, or science.

PISA assesses students and uses the outcomes of that assessment to produce estimates of students’ proficiency in relation to the skills and knowledge being assessed in each domain. The skills and knowledge of interest, as well as the kinds of tasks that represent those abilities, are described in the PISA frameworks (OECD, 2017). For each domain, one or more scales are defined, each ranging from very low levels of proficiency to very high levels. Students whose ability estimate places them at a certain point on a PISA proficiency scale would be more likely to be able to successfully complete tasks at or below that point. Those students would be increasingly *more likely* to complete tasks located at progressively lower points on the scale, and increasingly *less likely* to complete tasks located at progressively higher points on the scale. Figure 15.1 depicts a simplified hypothetical proficiency scale, ranging from relatively low levels of proficiency at the bottom of the figure, to relatively high levels towards the top. Six items of varying difficulty are placed along the scale, as are three students of varying ability. The relationship between the students and items at various levels is described in the figure.

In addition to defining the numerical range of the proficiency scale, it is also possible to define the scale by describing the competencies typical of students at particular points along the scale. The distribution of students along this proficiency scale is estimated, and locations of students can be derived from this distribution and their responses on the test. Those location estimates are then aggregated in various ways to generate and report useful information about the proficiency levels of 15-year-old students within and among participating countries.

The development of a method for describing proficiency in PISA reading, mathematical and scientific literacy occurred in the lead-up to the reporting of outcomes of the PISA 2000 survey and was revised in the lead-up to the PISA 2003, 2006, 2009 and 2012 surveys. Although essentially the same methodology has again been used to develop proficiency descriptions for PISA 2015, a more general statistical model compared to previous cycles was used in the scaling procedure (see Chapter 9 for details).

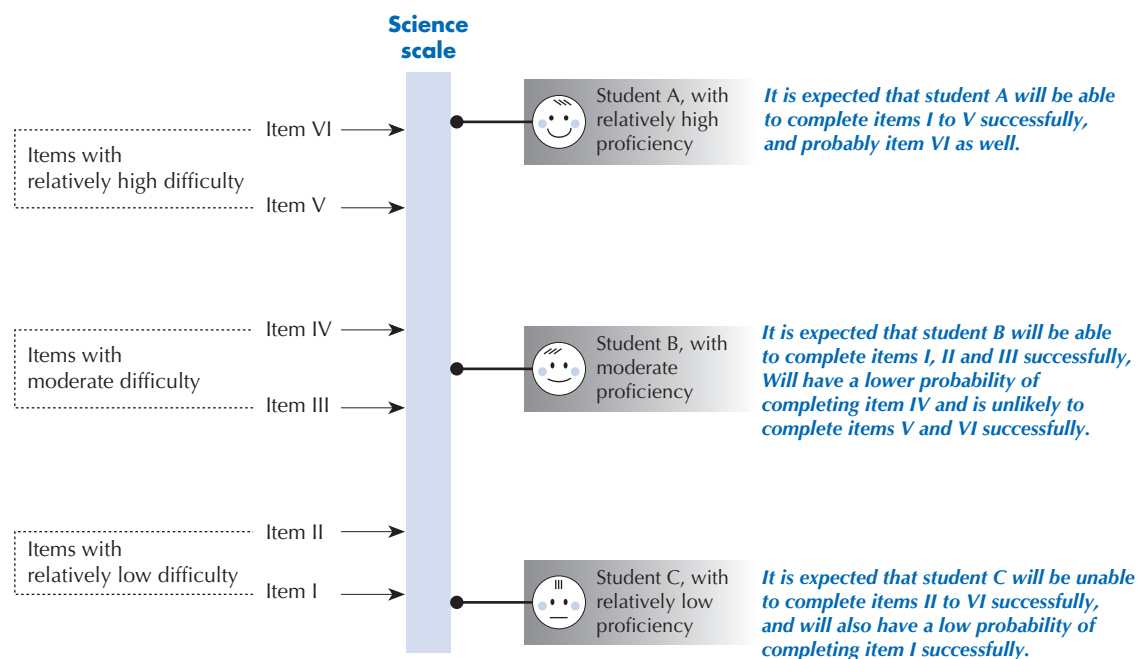
The proficiency descriptions that had been developed for the mathematics domain in PISA 2012, for reading in 2009 and for financial literacy in 2012 were used again to report the 2015 results.

Reporting for science, the major domain in 2015, was linked back to the 2006 proficiency scale and was based on the detailed proficiency level descriptions developed in 2006, the last cycle in which science was the major domain. These proficiency level descriptors were reviewed and revised based on the 2015 data in order to incorporate the new science framework developed for this cycle and the performance of the new computer-based items, including the interactive simulation tasks.



■ Figure 15.1 ■

Simplified relationship between items and students on a proficiency scale



The science expert group worked with the PISA international contractor to review and revise the sets of described proficiency scales and subscales for PISA science. Similarly, the international contractor worked with the collaborative problem solving expert group to develop the described proficiency scale for that domain.

DEVELOPMENT OF THE DESCRIBED SCALES

The development of described proficiency scales for PISA has been carried out through a process that typically involves a number of tasks conducted by the expert groups and the item development team. The process of developing the described scales involved several iterations as the data were collected and analysed during the 2015 cycle. It should be noted that, as each PISA cycle builds upon the work implemented in previous cycles, the same tasks are not completed for every domain in every cycle. The following description of the development process focuses on the development of described proficiency scales for science and collaborative problem solving.

Classification of items

As part of new item development for science and collaborative problem solving, test developers classified all items based on the specifications provided in the framework for each domain. Item classifications for the trend science items were also revised to reflect the 2015 framework. All classifications were reviewed by each of the expert groups and revised as needed.

Defining the overall proficiency scale

As part of its work in developing the framework for science, the expert group drafted initial descriptors of the levels of scientific literacy, based on the knowledge and competency dimensions defined therein. These descriptors, presented as an initial hypothesis, were shared as part of the framework to allow item developers to design items representing the increase in skills and ability reflected across the levels.

Final item parameters were estimated for the trend and new science items based on analysis of the Main Survey data. Using this information on item performance, the science expert group met over several days and reviewed each of the items and discussed key characteristics that differentiated performance along the proficiency scale. As part of that review process, the initial draft descriptors for each level in the overall proficiency scale were refined and finalised.



Defining the proficiency scale for collaborative problem solving was more challenging because the domain was newly developed for the 2015 cycle. The experts defined a matrix of collaborative problem solving skills in the collaborative problem solving framework that served as the basis for describing performance along the scale. They also set cut-off points along the scale that defined each level of performance. Description and definition of the proficiency scale for collaborative problem solving is also provided in the PISA 2015 Frameworks report (OECD, 2017).

Identifying possible subscales

For each domain in PISA, reporting includes an overall proficiency scale based on the combined results for all items within that domain. In addition, the framework may support subscales based on the various dimensions of the framework. Where subscales are included, they must arise clearly from the domain framework, be meaningful and potentially useful for feedback and reporting purposes, and be defensible with respect to their measurement properties. Thus, the first stage in the process involves having the experts articulate possible reporting subscales based on the most recent framework.

As the major domain in PISA 2015, work on identifying possible subscales for science, in addition to the overall scientific literacy scale, began with a review of the subscales used in the 2006 cycle, when science was last a major domain. The subscales selected for inclusion in the PISA 2006 database were the three competency-based subscales based on the scientific dimensions documented in the framework: *explaining phenomena scientifically*, *identifying scientific issues* and *using scientific evidence*. The 2015 expert group recommended reporting again on the three scientific competencies, as they were defined in the updated framework: *explain phenomena scientifically*, *evaluate and design scientific enquiry*, and *interpret data and evidence scientifically*. In addition, the expert group recommended that two knowledge subscales be reported: *content knowledge* and *procedural/epistemic knowledge*. Procedural and epistemic knowledge were combined into a single reporting subscale due to a limited number of epistemic items in some of the administered forms. Finally, for continuity with previous reporting scales, three systems – *physical*, *living* and *Earth and space* – were recommended as a third reporting scale.

For reading in the PISA 2000 cycle, in addition to the overall reading literacy scale, two main options were considered: subscales based on the type of reading task and subscales based on the form of reading material. For the international report, the first of these was implemented, leading to the development of subscales to describe the types of reading tasks, or “aspects” of reading: a subscale for *retrieving information*, a second subscale for *interpreting texts* and a third for *reflection and evaluation*. The thematic report for PISA 2000, *Reading for Change*, also reported on the development of subscales based on the form of reading material: *continuous texts* and *non-continuous texts* (OECD, 2002). In the 2009 cycle, volume I of the *PISA 2009 Results* included descriptions of both sets of subscales as well as a combined print reading scale (OECD, 2010). The names of the aspect subscales were modified in order to better apply to digital as well as print reading tasks. The modified aspect category names were *access and retrieve* (replacing *retrieving information*), *integrate and interpret* (replacing *interpreting texts*) and *reflect and evaluate* (for *reflection and evaluation*). For digital reading, a separate, single scale was developed based on the digital reading assessment items administered in 19 countries in PISA 2009 as an international option (OECD, 2011). For PISA 2012, when reading reverted to minor domain status, a single print reading scale was reported, along with a single digital reading scale.

In the case of mathematics, a single mathematical literacy scale was developed for PISA 2000. With the additional data available in the 2003 survey cycle, when mathematics was the major test domain, subscales based on the four overarching ideas – *space and shape*, *change and relationships*, *quantity* and *uncertainty* – were reported. In PISA 2006 and PISA 2009, when mathematics was again a minor domain, only a single scale was reported. For PISA 2012, the expert group carried out a comprehensive revision of the framework at the specific behest of the PISA Governing Board that indicated an interest in seeing mathematical process dimensions used as the primary basis for reporting in mathematics. As well as considering ways in which this could be done, the mathematics expert group also had to consider how the addition of the optional computer-based assessment component included in this cycle could be incorporated into the reporting for 2012. The outcome of these considerations was, firstly, a decision that the computer-based items would be used to expand the same mathematical literacy dimension that was expressed through the paper-based items. Secondly, the expert group recommended that three process-based subscales should be reported. These included: *formulating situations mathematically* (or “formulate”), *employing mathematical concepts, facts, procedures and reasoning* (or “employ”), and *interpreting, applying and evaluating mathematical outcomes* (or “interpret”). In addition, for continuity with the PISA 2003 reporting scales, the content-based scales including *space and shape*, *change and relationships*, *quantity*, and *uncertainty and data* (formerly “uncertainty”), were also reported.



For both collaborative problem solving and the optional assessment of financial literacy in PISA 2015, proficiency descriptions on a single overall reporting scale were developed.

Developing an item map

Based on item performance in the main survey, the test items in the study can be ordered from easiest to most difficult and this range of difficulty can be described using an item map. The item map contains a brief description of a selected number of released items along with their scale values. These descriptions explain the specific skills each item is designed to assess and are linked to the descriptions of performance at each level for the overall scale. As a result, the item map provides some insight into the range of skills and knowledge required of students and the proficiencies they need to demonstrate at various points along the scale.

DEFINING THE PROFICIENCY LEVELS

The proficiency levels for each of the PISA domains were defined in previous cycles when each was first a major domain. The goal of that process was to decide how to divide the proficiency continuum up into levels that might have some utility. And, having defined those levels, decisions needed to be made about how to decide on the level to which a particular student should be assigned.

The relationship between the observed responses, on the one hand, and student proficiency and item characteristics, on the other hand, is probabilistic. That is, there is some probability that a particular student can correctly solve any particular item and each item can be differentially responsive to the proficiency being measured.

One of the basic tenets of the measurement of human skills or proficiencies is this: If a student's proficiency level exceeds the item's demands, the probability that the student can successfully complete that item is relatively high, and if the student's proficiency is lower than that required by the item, the probability of success for that student on that item is relatively low. The rate of change of the probability of success across the range of proficiency for each item is also affected by the sensitivity of the item to the proficiency scale.

This leads to the question as to the precise criterion that should be used to locate a student on the same scale as that on which the items are located. How can we assign a location that represents student proficiency in meaningful ways? When placing a student at a particular point on the scale, what probability of success should we deem sufficient in relation to items located at the same point on the scale? If a student were given a test comprising a large number of items, each with the same item characteristics, what proportion of those items would we expect the student to successfully complete? Or, thinking of it in another way, if a large number of students of equal ability were given a single test item with a specified item characteristic, about how many of those students would we expect to successfully complete the item?

The answers to these questions depend on assumptions about how items differ in their characteristics or how items function, as well as on what level of probability is deemed a *sufficient probability of success*. In order to define and report PISA outcomes in a consistent manner, an approach is needed to define performance levels and to associate students with those levels. The methodology that was developed and used for previous cycles of PISA was essentially retained for PISA 2015, except that a more general statistical model was used to estimate item parameters, including difficulties (see Chapter 9 for details).

Defining proficiency levels for PISA 2000 progressed in two broad phases. The first, which came after the development of the described scales, was based on a substantive analysis of PISA items in relation to the aspects of literacy that underpinned each test domain. This produced descriptions of increasing proficiency that reflected observations of student performance and a detailed analysis of the cognitive demands of PISA items. The second phase involved decisions about where to set cut-off points for levels and how to associate students with each level in order to lay out how a *sufficient probability of success* plays out in these levels. This is both a technical and a very practical matter of interpreting what it means to be at a level, and has significant consequences for reporting national and international results.

Several principles were considered in developing and establishing a useful meaning of being at a level, and therefore for determining an approach to locating cut-off points between levels and associating students with them. For the levels to provide useful information to PISA stakeholders, it is important to develop a common understanding of what performance at each of those levels means.



First, it is important to understand that the skills measured in each PISA domain fall along a continuum: There are no natural breaking points to mark borderlines between stages along this continuum. Dividing the continuum into levels, though useful for communication about students' development, is essentially arbitrary. Like the definition of units on, for example, a scale of length, there is no fundamental difference between 1 metre and 1.5 metres – it is a matter of degree. It is useful, however, to define stages, or levels along the continua, because they enable us to communicate about the proficiency of students in terms other than continuous numbers. This is a rather common concept, an approach we all know from categorising shoes or shirts by size (S, M, L, XL, etc.).

The approach adopted for PISA 2000 was that it would only be useful to regard students as having attained a particular level if this would mean that we can have certain expectations about what these students are capable of, in general, when they are said to be at that level. It was thus decided that this expectation would have to mean, at a minimum, that students at a particular level would be more likely than not to successfully complete tasks at that level. By implication, it must be expected that they would succeed on at least half of the items on a test composed of items uniformly spread across that level. This definition of being “at a level” is useful in helping to interpret the proficiency of students at different points across the proficiency range defined at each level.

For example, the expectation is that students located at the bottom of a level would complete at least 50% of tasks correctly on a test set at the level, while students at the middle and top of each level would be expected to achieve a higher success rate. At the top border of a level would be the students who have mastered that level. These students would be likely to solve a high proportion of the tasks at that level. But, being at the top border of that level, they would also be at the bottom border of the next highest level where, according to the reasoning here, they should have at least a 50% likelihood of solving any tasks defined to be at that higher level.

Furthermore, the meaning of being at a level for a given scale should be more or less consistent for each level and, indeed, also for scales from the different domains. In other words, to the extent possible within the substantively based definition and description of levels, cut-off points should create levels of more or less constant breadth. Some small variation may be appropriate, but for interpretation and definition of cut-off points and levels to be consistent, the levels have to be about equally broad within each scale. The exception would be the highest and lowest proficiency levels, which are unbounded.

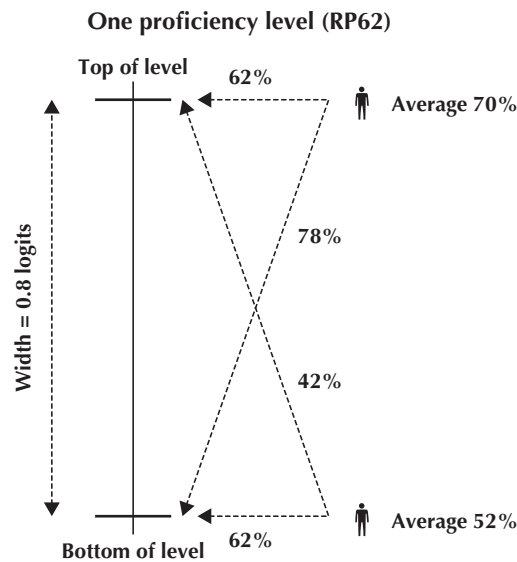
Thus, a consistent approach should be taken to defining levels for the different scales. Their breadth may not be exactly the same for the proficiency scales in different domains, but the same kind of interpretation should be possible for each scale that is developed. This approach links the two variables mentioned in the preceding paragraphs, and third related variable. The three variables can be expressed as follows:

- the expected success of a student at a particular level on a test containing items at that level (proposed to be set at a minimum that is near 50% for the student at the bottom of the level and greater for students who are higher in the level)
- the width of the levels in that scale (determined largely by substantive considerations of the cognitive demands of items at the level and data related to student performance on the items)
- the probability that a student in the middle of a level would correctly answer an item of average difficulty for that level (in fact, the probability that a student at any particular level would get an item at the same level correct), sometimes referred to as the “RP value” for the scale, where “RP” indicates “response probability”.

Figure 15.2 summarises the relationship among these three mathematically linked variables under a particular scenario. The vertical line represents a segment of the proficiency scale, with marks delineating the “top of level” and “bottom of level” for any level one might want to consider, with a width of 0.8 logits between the boundaries of the level (noting that this width can vary somewhat for different domains). The RP62 indicates that students will be located on the scale at a point that gives them a 62% chance of getting a typical item at that same level correct.¹ The student represented near the top of the level shown has a 62% chance of getting an item correct that is located at the top of the level, and similarly the student represented at the bottom of the level has the same chance of correctly answering a question at the bottom of the level. A student at the bottom of the level will have an average score of about 52% correct on a set of items spread uniformly across the level. Of course, that student will have a higher likelihood (62%) of getting an item at the bottom of the level correct, and a lower likelihood (about 42%) of getting an item at the top of the level correct. A student at the top of the level will have an average score of about 70% correct on a set of items spread uniformly across the level. That student will have a higher likelihood (about 78%) of getting a typical item at the bottom of the level correct and a lower likelihood (62%) of getting an item at the top of the level correct.



■ Figure 15.2 ■
Calculating the RP values used to define PISA proficiency levels



PISA 2000 implemented the following solution: Start with the range of described abilities for each bounded level in each scale (the desired band breadth); then determine the highest possible RP value that will be common across domains potentially having bands of slightly differing breadth that would give effect to the broad interpretation of the meaning of being at a level (an expectation of correctly responding to a minimum of 50% of the items in a test comprising items spread uniformly across that level). The value $RP = 0.62$ is a probability value that satisfied the logistic equations for typical items in that level through which the scaling model is defined, subject to the two constraints mentioned earlier (a width per level of about 0.8 logits and the expectation that a student would get at least half of the items correct on a hypothetical test composed of items spread evenly across the level). In fact, $RP=0.62$ satisfied the requirements for any scales having band widths up to about 0.97 logits.

The highest and lowest levels are unbounded. For a certain high point on the scale and below a certain low point, the proficiency descriptions could, arguably, cease to be applicable. At the high end of the scale, this is not such a problem since extremely proficient students could reasonably be assumed to be capable of at least the achievements described for the highest level. At the other end of the scale, however, the same argument does not hold. A lower limit therefore needs to be determined for the lowest described level, below which no meaningful description of proficiency is possible. It was proposed that the floor of the lowest described level be set so that it was the same breadth as the other described levels. Student performance below this level is lower than that which PISA can reliably assess and, more importantly, describe.

REPORTING THE RESULTS FOR PISA SCIENCE

In this section, the ways in which levels of scientific literacy are defined, described and reported will be discussed. This will be illustrated using a subset of items from the PISA 2015 assessment.

Building an item map for science

The data from the PISA science assessment were analysed to estimate a set of item characteristics for the 184 items included in the main survey.² During the process of item development, each item was classified to reflect the scientific competency and type of knowledge it required. In addition, items were classified based on specific content knowledge, or systems (physical systems, living systems or Earth and space systems), as well as their context (personal, local/national or global). Following data analysis, the items were associated with their difficulty estimates and framework classifications. Figure 15.3 shows the item map, which includes this information along with a brief qualitative description for the released items from the PISA 2015 test. Each row in Figure 15.3 represents an individual item. The selected items have

been ordered according to their difficulty, with the most difficult at the top, and the least difficult at the bottom. The difficulty estimate for each item is given, along with the associated classifications and descriptions.

When an item map such as this is prepared, it becomes possible to look for factors that are associated with item difficulty. This can be done by referring to the ways in which scientific literacy is associated with questions located at different points ranging from the bottom to the top of the scale. For example, the item map in Figure 15.3 shows that the easiest items tend to require the application of everyday content knowledge and the ability to recognise aspects of simple scientific phenomena. The most difficult items, by contrast, draw on a range of interrelated scientific ideas and concepts and require the application of sophisticated procedural and epistemic knowledge to offer explanatory hypotheses of novel scientific phenomena, events and processes.

■ Figure 15.3 ■

A map for selected science items

Code	Item Name	Item Difficulty (RP=0.62)	Item Demands	Explain Phenomena	Evaluate & Design Scientific Enquiry	Interpret Data & Evidence	Content	Procedural	Epistemic	Physical	Living	Earth and Space
CS601Q01	Sustainable Fish Farming	740	Use multiple sources of information to evaluate a system in an unfamiliar context and the interaction among elements in that system.	●			●				●	
CS623Q04	Running in Hot Weather	641	Draw on scientific knowledge to explain a biological reason for an outcome observed in a simulated experiment.	●			●				●	
CS656Q02	Bird Migration	630	Identify a factor that could result in an inadequate or inaccurate set of data and explain its effect.		●			●			●	
CS623Q06	Running in Hot Weather	598	Run a simulated experiment manipulating two independent variables. Use those results to hypothesize the outcome of the experiment with a value for one variable that is not available in the simulation. Select data from the experiment supporting that choice and explain how it does so.		●			●			●	
CS637Q05	Slope-Face Investigation	589	Draw on epistemic knowledge and use provided data to identify the appropriate conclusion from an experiment using controls, providing a reason that justifies that choice.			●			●			●
CS623Q05	Running in Hot Weather	592	Given one defined variable, run a simulated experiment to identify the highest level for a second variable before a negative outcome would occur. Select data from the experiment supporting that choice and explain how it does so.		●			●			●	
CS601Q04	Sustainable Fish Farming	585	Go beyond the provided information to identify a procedure that would meet a specified goal.	●			●			●		
CS623Q02	Running in Hot Weather	580	Run a simulated experiment holding two variables constant and identify the effect of varying the third. Select data from the experiment supporting that choice.			●	●				●	
CS656Q04	Bird Migration	574	Identify one or more statements supported by information provided in two moderately complex representations of data.		●			●			●	
CS623Q03	Running in Hot Weather	531	Given one defined variable, run a simulated experiment to identify the impact of a second variable and identify data supporting that choice.		●			●			●	
CS637Q01	Slope-Face Investigation	517	Draw on epistemic knowledge to explain why a simple experimental design includes two independent measures of a phenomenon.		●				●			●
CS656Q01	Bird Migration	501	Draw on knowledge of life science to identify an explanation of a familiar phenomenon.	●			●				●	
CS623Q01	Running in Hot Weather	497	Follow instructions to carry out and interpret the results of a simple simulated experiment involving two independent variables.			●		●			●	
CS641Q01	Meteoroids & Craters	483	Use simple scientific knowledge to identify the effect of Earth's mass on the speed of objects entering the atmosphere.	●			●			●		
CS601Q02	Sustainable Fish Farming	456	Identify one component of a system that will result in a desired outcome, given an explanation of the function performed by each component.			●	●				●	
CS641Q02	Meteoroids & Craters	450	Use simple scientific knowledge to identify the relationship between a planet's atmosphere and the likelihood that meteoroids will burn up before hitting the planet surface.	●			●					●
CS641Q04	Meteoroids & Craters	438	Use familiar and simple scientific knowledge to order three craters by their age from oldest to newest, based on an image showing craters of different sizes.			●	●					●
CS641Q03	Meteoroids & Craters	299	Use everyday scientific knowledge to match the size of a meteoroid with the size of the crater it would create on a planet's surface, based on an image showing three craters of different sizes.			●	●					●

Based on the patterns observed in the science item pool, it was possible to characterise the increasing complexity of competencies measured. This can be done by referring to the ways in which science competencies are associated with items located at different points, ranging from the bottom to the top of the scale. The ascending difficulty of science questions in PISA 2015 is associated with the following attributes, which require all three competencies but shift in



emphasis as students progress from the application of simple everyday knowledge to using more sophisticated content, procedural and epistemic knowledge to develop hypotheses about novel scientific phenomena, events and processes.

The attributes include the following:

- The degree to which the transfer and application of knowledge is required. At the lowest levels the application of knowledge is simple and direct. The requirement can often be fulfilled with simple recall of single facts. At higher levels of the scale, individuals are required to draw on multiple fundamental concepts and combine categories of knowledge in order to respond correctly.
- The degree of cognitive demand required to analyse the presented situation and to synthesise an appropriate answer. The 2015 Scientific Literacy framework defined increasing complexity based on levels of cognitive demand within the assessment of scientific literacy and across all three competencies of the framework. The factors that determine the cognitive demand of items in science include: the number of elements of knowledge and their degree of complexity; the level of familiarity and prior knowledge that students may have of the content, procedural and epistemic knowledge involved; the cognitive operation required by the item (e.g., recall, analysis, evaluation); and the extent to which forming a response is dependent on models or abstract scientific ideas.

For example, items with low cognitive complexity typically involve carrying out a one-step procedure, for example, recalling a fact, term, principle, or concept, or locating a single point of information from a graph or table. Items with medium cognitive complexity require the use and application of conceptual knowledge to describe or explain phenomena, select appropriate procedures involving two or more steps, organise or display data, or interpret or use simple data sets or graphs. Finally, items with high cognitive demand require students to analyse complex information or data, synthesise or evaluate evidence, reason given various sources, or develop a plan or sequence of steps to approach a problem.

- The degree of analysis needed to answer a question is also an important driver of difficulty. This includes the demands arising from the requirement to discriminate among issues presented in the situation under analysis, identify the appropriate knowledge domain, and use appropriate evidence for claims or conclusions. The analysis may include the extent to which the scientific demands of the situation are clearly apparent or whether students must differentiate among components of the situation to clarify the scientific issues as opposed to other non-salient or non-scientific issues.
- The degree of synthesis required may impact item complexity. Synthesis may range from a single piece of evidence where no real construction of justification or argument is required to situations requiring students to apply multiple sources of evidence and compare competing lines of evidence and different explanations to adequately argue a position.

Defining levels of scientific literacy

The reporting approach used by the OECD has been defined in previous cycles of PISA and is based on the definition of a number of levels of proficiency. Descriptions were developed to characterise typical student performance at each level. The levels were used to summarise the performance of students, to compare performances across subgroups of students, and to compare average performances among groups of students, in particular among the students from different participating countries. A similar approach has been used here to analyse and report PISA 2015 outcomes for science.

For PISA 2006 science, student scores were transformed to the PISA scale, with a mean of 500 and a standard deviation of 100, and levels of proficiency were defined and described. In accordance with the approach taken for the other PISA domains, the science scale has been extended to describe one level below the lowest previously-described level. Thus the PISA 2015 science scale has seven described levels instead of the six defined for PISA 2006. The previously-named Level 1 was renamed Level 1a and the level defined below this was named Level 1b.

The level definitions on the PISA scale are given in Table 15.1.

Table 15.1 Scientific literacy performance band definitions on the PISA scale

Level	Score points on the PISA scale
6	Higher than 707.93
5	Higher than 633.33 and less than or equal to 707.93
4	Higher than 558.73 and less than or equal to 633.33
3	Higher than 484.14 and less than or equal to 558.73
2	Higher than 409.54 and less than or equal to 484.14
1a	Higher than 334.94 and less than or equal to 409.54
1b	260.54 to less than or equal to 334.94

Information about the items in each level is used to develop summary descriptions of the kinds of scientific literacy associated with different levels of proficiency. These summary descriptions can then be used to encapsulate typical science proficiency of students associated with each level. As a set, they describe development in scientific literacy.

PISA is administered once every three years, with each of the three core domains the major focus in turn. Science was the major domain in PISA 2006. PISA 2015, therefore, had a set of level descriptors upon which to build. The new items that were developed for PISA 2015 were considered in relation to the existing level descriptions and in relation to the preliminary descriptions that were included in the 2015 framework for scientific literacy. The focus was first on the descriptions for the overall science scale, presented here in Figure 15.4.

■ Figure 15.4 ■

Summary descriptions of the seven proficiency levels on the scientific literacy scale

Level	What students can typically do
6	At Level 6, students can draw on a range of interrelated scientific ideas and concepts from the physical, life and Earth and space sciences and use procedural and epistemic knowledge in order to offer explanatory hypotheses of novel scientific phenomena, events and processes that require multiple steps or to make predictions. In interpreting data and evidence, they are able to discriminate between relevant and irrelevant information and can draw on knowledge external to the normal school curriculum. They can distinguish between arguments that are based on scientific evidence and theory and those based on other considerations. Level 6 students can evaluate competing designs of complex experiments, field studies or simulations and justify their choices.
5	At Level 5, students can use abstract scientific ideas or concepts to explain unfamiliar and more complex phenomena, events and processes. They are able to apply more sophisticated epistemic knowledge to evaluate alternative experimental designs and justify their choices and use theoretical knowledge to interpret information or make predictions. Level 5 students can evaluate ways of exploring a given question scientifically and identify limitations in interpretations of data sets including sources and the effects of uncertainty in scientific data.
4	At Level 4, students can use more sophisticated content knowledge, which is either provided or recalled, to construct explanations of more complex or less familiar events and processes. They can conduct experiments involving two or more independent variables in a constrained context. They are able to justify an experimental design, drawing on elements of procedural and epistemic knowledge. Level 4 students can interpret data drawn from a moderately complex data set or less familiar contexts and draw appropriate conclusions that go beyond the data and provide justifications for their choices.
3	At Level 3, students can draw upon moderately complex content knowledge to identify or construct explanations of familiar phenomena. In less familiar or more complex situations, they can construct explanations with relevant cueing or support. They can draw on elements of procedural or epistemic knowledge to carry out a simple experiment in a constrained context. Level 3 students are able to distinguish between scientific and non-scientific issues and identify the evidence supporting a scientific claim.
2	At Level 2, students are able to draw on everyday content knowledge and basic procedural knowledge to identify an appropriate scientific explanation, interpret data, and identify the question being addressed in a simple experimental design. They can use everyday scientific knowledge to identify a valid conclusion from a simple data set. Level 2 students demonstrate basic epistemic knowledge by being able to identify questions that could be investigated scientifically.
1a	At Level 1a, students are able to use everyday content and procedural knowledge to recognise or identify explanations of simple scientific phenomenon. With support, they can undertake structured scientific enquiries with no more than two variables. They are able to identify simple causal or correlational relationships and interpret graphical and visual data that require a low level of cognitive demand. Level 1a students can select the best scientific explanation for given data in familiar personal, local and global contexts.
1b	At Level 1b, students can use everyday content knowledge to recognise aspects of simple scientific phenomenon. They are able to identify simple patterns in data, recognise basic scientific terms and follow explicit instructions to carry out a scientific procedure.

Figures 15.5, 15.6 and 15.7 provide the summary descriptions of knowledge and skills required to complete tasks located within the defined bands for the three competency subscales: *explaining phenomena scientifically*, *evaluating and designing scientific enquiry* and *interpreting data and evidence scientifically* respectively.

■ Figure 15.5 [Part 1/2] ■

Summary descriptions of the proficiency levels on the scientific literacy subscale *Explain phenomena scientifically*

Level	General proficiencies students should have at each level	Tasks a student should be able to do
6	At Level 6, students can draw on a range of inter-related scientific ideas and concepts from life, physical or Earth and space sciences to make predictions or to construct explanations of novel and unfamiliar phenomena, events and processes that may involve several steps. They can demonstrate the use of knowledge beyond standard science curricula and use procedural and epistemic knowledge appropriately.	<ul style="list-style-type: none"> ■ Construct acceptable scientific explanations, using a broad range of knowledge, ideas and concepts. ■ Recognise when data/information in the text does not answer the question. ■ Use given scientific knowledge and recall additional relevant scientific knowledge to explain an unfamiliar phenomenon. ■ Construct and run a mental model to offer an explanation or make a prediction in an unfamiliar situation. ■ Comment on the appropriate use of scientific models and their limitations.
5	Students at this level can use abstract scientific ideas or concepts to explain more complex phenomena, events and processes, which may be unfamiliar.	<ul style="list-style-type: none"> ■ Select an appropriate scientific explanation of an unfamiliar event, phenomenon or process. ■ Construct an appropriate explanation drawing upon abstract scientific ideas and constructs. ■ Apply theoretical scientific knowledge to interpret given information, develop an explanation or make a prediction.



■ Figure 15.5 [Part 2/2] ■

Summary descriptions of the proficiency levels on the scientific literacy subscale
Explain phenomena scientifically

Level	General proficiencies students should have at each level	Tasks a student should be able to do
4	At Level 4, students can recall or use given scientific ideas to construct explanations of relatively complex or less familiar events and processes, or to make simple predictions.	<ul style="list-style-type: none"> Identify or construct an appropriate causal explanation for a more complex or less familiar phenomenon, event or process. Identify the relationship between simple physical quantities and use this to explain a phenomenon. Predict how one quantity will change when other quantities change. Use scientific knowledge to evaluate a claim or to interpret an unfamiliar phenomenon. Recognise relationships between physical quantities.
3	Students at this level can draw upon moderately complex scientific facts and ideas to identify or construct appropriate simple explanations of familiar phenomena. In less familiar or more complex situations, they can construct an explanation with relevant cueing or support.	<ul style="list-style-type: none"> Construct simple explanations of familiar phenomena drawing on knowledge from life, physical or Earth and space sciences. Identify a conclusion consistent with given information in an unfamiliar context. Select from multiple components and place them in a logical order to construct simple explanations. Identify causal factors which explain a phenomenon.
2	At Level 2, students can recall and apply simple scientific facts and ideas, or select a simple scientific explanation, given relevant cues and support.	<ul style="list-style-type: none"> Use familiar and simple scientific knowledge to draw an appropriate conclusion. Select the correct explanation of a relatively familiar scientific situation. Choose appropriate alternatives to complete an explanation. Use simple scientific knowledge to identify causal relationships. Reconstruct a temporal sequence for a familiar scientific phenomenon.
1a	Students at this level can select an appropriate example of a given simple scientific concept or identify an appropriate scientific explanation for a familiar event or process that is consistent with given information.	<ul style="list-style-type: none"> Use familiar content and procedural knowledge to recognise or identify explanations. Select the best scientific explanation from a list for given data in familiar contexts.
1b	Students at this level can recognise scientific terms and use single scientific facts close to their personal experience to recognise very simple cause and effect relationships.	<ul style="list-style-type: none"> Recognise simple scientific language or scientific conventions used in everyday life situations. Use familiar content knowledge to recognise scientific aspects of simple phenomena in tasks that require a low level of cognitive demand.

■ Figure 15.6 [Part 1/2] ■

Summary descriptions of the seven proficiency levels on the scientific literacy subscale
Evaluate and design scientific enquiry

Level	General proficiencies students should have at each level	Tasks a student should be able to do
6	At Level 6, students can evaluate competing designs of complex experiments, field studies or simulations and justify their choices.	<ul style="list-style-type: none"> Evaluate an investigation involving multiple variables requiring the identification of the independent or dependent variable. Justify choices and the range of data to be collected drawing on relevant epistemic and/or procedural knowledge. Evaluate and comment on the model inherent to experimental designs.
5	Students at this level can evaluate alternative experimental designs or data interpretations and justify their choices. They can identify limitations of the interpretations of data sets.	<ul style="list-style-type: none"> Evaluate whether an empirical question can be answered scientifically or not. Justify a more detailed feature of an experimental design. Provide a procedural justification for the inadequacy of a set of data. Choose between two experimental designs and justify the choice drawing on procedural, epistemic or content knowledge. Justify a data collection procedure in a context involving several independent variables.
4	At Level 4, students can conduct experiments involving two or more independent variables in a constrained context and justify aspects of their experimental design, drawing on procedural and epistemic knowledge. They can interpret data drawn from more complex or less familiar contexts and draw appropriate conclusions that go beyond the data. They can use data from less familiar contexts to identify trends and make predictions.	<ul style="list-style-type: none"> Carry out and interpret a simple experiment involving the manipulation of more than one independent variable. Follow instructions to identify the outcome of several variable choices. Manipulate variables to answer a scientific question, identify a trend, interpolate between, or extrapolate beyond, the data. Justify the conclusions of an experimental design drawing on procedural or epistemic knowledge. Identify the question of an investigation of a more complex or less familiar experimental design.

■ Figure 15.6 [Part 2/2] ■

**Summary descriptions of the seven proficiency levels on the scientific literacy subscale
*Evaluate and design scientific enquiry***

Level	General proficiencies students should have at each level	Tasks a student should be able to do
3	Students at this level can draw on procedural or epistemic knowledge to design, and justify aspects of the design of, a simple experiment in a constrained context. They can distinguish between scientific, technological and non-scientific issues.	<ul style="list-style-type: none"> Identify which variable to control in a two variable experiment. Drawing on epistemic or procedural knowledge, provide a justification for aspects of a simple experimental design. Identify the role of simulations in scientific enquiry. Discriminate between issue that can be solved by science or other means. Within a constrained context, identify a set of data that could answer a specified question about a phenomenon.
2	Students at Level 2 are able to draw on procedural and basic content knowledge to identify the question being addressed in a simple experimental design. They can collect and interpret data to answer questions that require only simple or everyday content knowledge. They can distinguish between a non-scientific and scientific question.	<ul style="list-style-type: none"> Given a simple experimental design, identify the question being addressed. Distinguish between simple scientific and simple non-scientific questions. Interpret simple data sets and draw an appropriate conclusion using everyday knowledge. Carry out a straightforward procedure to collect a data set to answer a simple question. Identify the aspects of a simple model and the external features they represent.
1a	Students at this level can, with support, carry out a simple experiment involving one independent and one dependent variable to generate data to answer a question.	<ul style="list-style-type: none"> Identify the independent variable in a given situation. Follow instructions to carry out a simple experiment to investigate how an outcome changes when one independent variable is changed.
1b	At this level, students typically are able to follow simple instructions to carry out a scientific procedure.	<ul style="list-style-type: none"> Run a simulation to extract a single data point.

■ Figure 15.7 ■

**Summary descriptions of the seven proficiency levels on the scientific literacy subscale
*Interpret data and evidence scientifically***

Level	General proficiencies students should have at each level	Tasks a student should be able to do
6	At Level 6, students can evaluate the strength of support provided by data for competing hypotheses and construct and justify a conclusion using abstract science concepts. They can also discriminate between relevant and irrelevant information, and draw on outside knowledge to construct an explanation.	<ul style="list-style-type: none"> Evaluate a complex set of data to determine whether each piece of data supports one, both or neither of two or more competing hypotheses. Provide a reason for their choice using abstract science concepts and applying procedural or epistemic knowledge.
5	Students at this level can interpret a moderately complex data set to construct and justify a conclusion using abstract science concepts. They can also identify sources and effects of uncertainty in scientific data.	<ul style="list-style-type: none"> Analyse complex data to identify which of several inferences is correct. Generate a set of data from a simulation by manipulating a single variable to identify the correct outcome from a number of possibilities.
4	At Level 4, students can interpret and manipulate a moderately complex data set expressed in a number of formats to select or justify appropriate conclusions. They can also distinguish between scientific and social or personal issues when interpreting data.	<ul style="list-style-type: none"> Analyse moderately complex data to identify which of several inferences is correct. Analyse more complex data to identify the appropriate conclusion of an experiment using controls and provide a reason that justifies their choice.
3	Students at this level can interpret and transform data to support a claim or conclusion. They can identify the evidence supporting a scientific claim.	<ul style="list-style-type: none"> Analyse a data table to identify which of several inferences is correct. Use data to identify the appropriate conclusion from an experiment using controls or a set of data and provide a reason that justifies their choice.
2	Students at this level can identify data that support a claim or conclusion and interpret data to select relevant explanations.	<ul style="list-style-type: none"> Analyse tabular or graphic data to identify which of several hypotheses or claims are supported by the data. Identify the pattern in a data set such as a graph or table.
1a	Students at this level can identify whether simple data support a claim or conclusion. They can make straightforward interpretations of simple data sets presented in different formats.	<ul style="list-style-type: none"> Identify the trend in simple data set. Transform simple data representations between pictorial, graphical, tabular and text. Use a simple data set to identify data that support a conclusion.
1b	Students at this level can identify simple patterns in data.	<ul style="list-style-type: none"> In response to a specific question showing a simple pictorial representation of objects, make comparisons and judgments about the differences observed.



Notes

1. A typical item is one that has an item slope parameter of 1.0 in this example. In the more general statistical model used for 2015, items are allowed to vary in their slope parameter, which quantifies the strength of the relationship between proficiency and item response. This slope parameter was introduced to allow tasks to be more appropriately described if not fitted well by assuming a common slope for all items. This leads to a reporting model that described the observed student responses much more appropriately but, as a consequence, it requires talking about typical items in terms of RP62 and proficiency levels.
2. For a detailed description of the scaling procedures used in PISA 2015, see Chapter 9 of this report.

References

- OECD (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264281820-en>.
- OECD (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance (Volume VI)*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264112995-en>.
- OECD (2010), *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264091450-en>.
- OECD (2002), *Reading for Change: Performance and Engagement across Countries: Results from PISA 2000*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264099289-en>.