acara
AUSTRALIAN CURRICULUM,
ASSESSMENT AND
REPORTING AUTHORITY

NAP Sample Assessment

Science Literacy

2018

Technical report

NAP
NATIONAL
ASSESSMENT
PROGRAM

The technical report was edited by Dr Eveline Gebhardt and Joanne Sim.

NAP–SL project staff: Michelle Robins, ACARA's Project Manager, Joanne Sim, ACARA's Online Assessment Specialist, Dr Rainer Mittelbach, ACARA's Science Curriculum Specialist and Matthew Stokoe, ACARA's NAP Sample Project Officer. Kate O'Malley from the Australian Council for Educational Research (ACER) was Project Manager for the Main Study Administration and Frances Eveleigh from ACER was the Project Manager for the Marking operation. The data analysis was carried out by Professor David Andrich, Dr Ida Marais and Sonia Sappl, psychometricians from The University of Western Australia (UWA) and Dr Eveline Gebhardt, ACARA's Psychometrics Senior Manager, Rassoul Sadeghi, ACARA's Lead Psychometrician, Nick Marosszeky and Ben Businovski, ACARA's Assisting Psychometricians.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 OVERVIEW

Joanne Sim - ACARA

The National Assessment Program (NAP) commenced as an initiative of ministers of education in Australia to monitor outcomes of schooling specified in the 1999 Adelaide Declaration on National Goals for Schooling in the 21st Century (Adelaide Declaration).

NAP was established to measure student achievement in relation to the national goals and to report this, using nationally comparable data in each of literacy, numeracy, science, information and communication technologies (ICT), and civics and citizenship.

In 2008, the Adelaide Declaration was superseded by the Melbourne Declaration on the Educational Goals for Young Australians (Melbourne Declaration). The work of NAP has continued and was refined, as necessary, to monitor and report on the goals specified in the Melbourne Declaration[1].

The National Assessment Program – Science Literacy (NAP–SL) is one of three national sample assessments developed and managed by the Australian Curriculum, Assessment and Reporting Authority (ACARA) under the auspices of the Education Council. The other assessments are civics and citizenship and information and communication technology (ICT) literacy.

The first science literacy assessment was conducted in 2003. The assessment has been repeated with a new sample of Year 6 students every three years to identify trends over time. In 2004 and 2005, similar national assessments were introduced for students in Years 6 and 10 in civics and citizenship, and information and communications technology (ICT) literacy. Each of these programs assesses a representative sample of Australian students and is repeated every three years.

In July 2016, the Education Council decided to extend the NAP–SL to Year 10 students from 2018. The purpose of this decision was to reinforce the need to assess the science literacy progress of Australian students using assessments that are closely aligned with the Australian Curriculum, in addition to using outcomes of the international assessments and surveys. Until now, the Programme for International Student Achievement (PISA) has been the primary national measure of performance for science literacy among secondary school students. Australian students also participate in the Trends in International Mathematics and Science Study (TIMSS) which includes assessment of Year 8 students' knowledge of both the mathematics and science curricula.

The 2018 assessment cycle was delivered to a representative sample of both Year 6 students and Year 10 students. This report documents the findings from NAP–SL 2018 and includes comparisons, as appropriate, with findings from previous assessment cycles.

[1] In December 2019, the Melbourne Declaration on the Educational Goals for Young Australians was superseded by the Alice Springs (Mparntwe) Education Declaration.

What is assessed in NAP–SL

The NAP Science Literacy assessment measures the ability of students:

> *to use scientific knowledge, understanding, and inquiry skills to identify questions, acquire new knowledge, explain science phenomena, solve problems and draw evidence-based conclusions in making sense of the world, and to recognise how understandings of the nature, development, use and influence of science help us make responsible decisions and shape our interpretations of information*
>
> (https://www.australiancurriculum.edu.au/f-10-curriculum/science/glossary/?letter=S).

The 2018 NAP–SL Assessment Framework content is organised according to the strands of the Australian Curriculum: Science. The strands are:

- Science Understanding
- Science as a Human Endeavour
- Science Inquiry Skills.

All strands were assessed for Years 6 and 10 in the 2018 NAP–SL assessment.

Further information about the 2018 NAP–SL Assessment Framework is provided in chapter 2 of this report and the 2018 NAP–SL Public Report.

NAP–SL and the Australian Curriculum: Science

NAP–SL 2018 was aligned to the Australian Curriculum: Science. The aims of the Australian Curriculum: Science are congruent with and reflected in the 2018 NAP–SL Assessment Framework. The specific aims of the Australian Curriculum: Science are:

- the understanding of important science concepts and processes, the practices used to develop scientific knowledge, science's contribution to society, and society's influence on science from a range of cultures
- the ability to think and act in a scientific way
- the ability to make informed decisions about local, national and global issues. https://www.australiancurriculum.edu.au/f-10-curriculum/science/rationale/

Every item used in the 2018 cycle was mapped against the Australian Curriculum: Science strands, sub-strands and the cognitive dimensions. Where applicable, items were also classified against the general capabilities, including the critical and creative thinking capability.

Assessment instrument

The 2018 NAP–SL test instrument included test items presented in units. Each unit comprised a set of items that were developed around a stimulus. The units were allocated to clusters which were allocated to test forms that were 'equivalent' in terms of framework coverage, item types, reading load and overall difficulty. Each test form contained three components: a set of objective test items, an inquiry task and a set of survey items. Each student was randomly allocated one of the possible test forms for their year.

Delivering the Assessments

The assessment instrument was administered online to samples of students in Year 6 and Year 10 in October and November 2018.

Students completed all parts of the assessment using internet-connected school computers. Given the secure nature of the tests, participating students undertook the tests via the locked down browser (LDB).

In preparation for the assessment, schools were contacted to assess their preparedness to use the online delivery mode. Schools were required to run an online Technical Readiness Test (TRT) on the computers designated for testing.

Reporting of the assessment results

The results of the assessment are reported in the 2018 National Assessment Program – Science Literacy Public Report.

The NAP–SL scale comprises proficiency levels that are used to describe the achievement of students both at Year 6 and Year 10. The scale was revised in 2006 to describe the performance of Year 6 students nationally and has a mean score of 400 with a standard deviation of 100 scale points. NAP–SL scale scores from the four previous assessment cycles have been reported using this same metric.

Following the 2017 pilot study, Year 10 students were included in the assessment sample in 2018. The introduction of Year 10 students necessitated a standard-setting process to determine the location on the measurement scale representing the proficient standard for Year 10.

The proficient standard is a point on the scale that represents a *challenging but reasonable* expectation of student achievement at that year level. The proportion of students who meet or exceed the proficient standard is the key performance measure for Science literacy at each year level.

As part of the inclusion of the new proficient standard for Year 10, a change was made to the width of the proficiency levels and the levels were re-labelled so that the proficient standard for Year 6 is now the boundary between levels 2 and 3 and the proficient standard for Year 10 is the boundary between levels 3 and 4. The proficient standard for Year 6 remained unchanged. Therefore, the percentage of Year 6 students attaining or exceeding the proficient standard can be compared with previous assessments. In 2018, 58 per cent of Year 6 students reached or exceeded the Year 6 proficient standard, whereas 50 per cent of Year 10 students were at or above the proficient standard for this year level.

Purposes of the Technical Report

This report describes the technical aspects of the NAP–SL 2018 sample assessment and summarises the main activities involved in the data collection, the data collection instruments and the analysis and reporting of the data and should be read in conjunction with the 2018 NAP–SL Public Report, which focuses on results and key findings (ACARA, 2019).

Chapter 2 summarises the development of the assessment framework and describes the process of item development and construction of the instruments.

Chapter 3 reviews the sample design and describes the sampling process. It also describes the weighting procedures that were implemented to derive population estimates and the calculation of participation rates.

Chapter 4 summarises the field administration of the assessment.

Chapter 5 deals with management procedures, including quality control and the cleaning and coding of the data.

Chapter 6 describes the scaling model and procedures, item calibration, the creation of plausible values and the standardisation of student scores. It discusses the procedures used for vertical (Year 6 to Year 10) and horizontal (2018 to 2015, 2012, 2009 and 2006) equating and the procedures for estimating equating errors.

Chapter 7 outlines the achievement levels and the procedures undertaken to determine the new proficient standard for Year 10 students.

Chapter 8 discusses the reporting of student results, including the procedures used to estimate sampling and measurement variance.

Chapter 2 TEST DEVELOPMENT AND TEST DESIGN

Joanne Sim - ACARA

The NAP – Science Literacy assessment measures science literacy as defined in the Australian Curriculum: Science, that is the ability:

> *to use scientific knowledge, understanding, and inquiry skills to identify questions, explain science phenomena, solve problems and draw evidence-based conclusions in making sense of the world, and to recognise how understandings of the nature, development, use and influence of science help us make responsible decisions and shape our interpretations of information*

(https://www.australiancurriculum.edu.au/f-10-curriculum/science/glossary/?letter=S).

The definition of science literacy in the NAP–SL is consistent with recent definitions of science literacy internationally. For example, PISA 2015 defined science literacy as

> *the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen* (OECD, 2016).[2]

PISA's definition includes being able to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically.

Science literacy assessment framework development

Historical description

In the previous NAP–SL cycles, the program was underpinned and guided by a science literacy progress map which was based on the construct of science literacy defined by the OECD-PISA assessment and on an analysis of the state and territory curriculum and assessment frameworks. The progress map described the development of science literacy across three strands of knowledge assessment framework that predated the Australian Curriculum. The three main areas of scientific literacy that were assessed were:

- Strand A: formulating or identifying investigable questions and hypotheses; planning investigations; and collecting evidence.
- Strand B: interpreting evidence and drawing conclusions from students' own or others' data; critiquing the trustworthiness of evidence and claims made by others; and communicating findings.
- Strand C: using science understandings for describing and explaining natural phenomena, and for interpreting reports about phenomena.

For a detailed description of previous assessment frameworks, see the 2015 NAP – Science Literacy Public and Technical reports.

[2] OECD (2016). PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy. Paris: OECD Publishing.

NAP–SL Assessment Framework 2018

The NAP – Science Literacy assessment framework was reviewed and extended during 2017 and is now published. The primary focus of this multi-step, collaborative review was to:

- align the science literacy assessment with the Australian Curriculum.
- identify and select content and contexts for scientific skills and conceptual knowledge that reflect the expectations set by the Year 10 Science Achievement Standard, that are considered to be 'essential' for Year 10 students, that is, essential to enable students to confidently engage in scientific issues relating to everyday life experiences, as well as successfully transition into senior secondary science; and
- elaborate the progression of knowledge and skills shared between primary and secondary year levels that could enable vertical linking of the Year 6 and Year 10 assessments.

The 2018 NAP – Science Literacy Assessment Framework underpinned the development of the 2018 assessments for both Years 6 and 10. The revised framework was guided by the Australian Curriculum: Science and provides guidance on the content to be assessed, the cognitive engagement that is expected of students and the types of assessment tasks and questions to be included in the assessment. The full assessment framework can be located on the NAP website.

The Australian Curriculum: Science requires students to develop an understanding of important science concepts and processes; the practices used to develop scientific knowledge; and science's contribution to our culture and society and its applications in our lives.

Accordingly, the Australian Curriculum: Science has three interrelated strands – *Science Understanding, Science as a Human Endeavour* and *Science Inquiry Skills* – which are designed to be taught in an integrated way. Table 2.1 lists the strands of the curriculum and the sub-strands within each strand.

Table 2.1 Strands and sub-strands in the Australian Curriculum: Science

Strand	Sub-strand
Science Understanding	Biological sciences
	Chemical sciences
	Earth and space sciences
	Physical sciences
Science as a Human Endeavour	Nature and development of science
	Use and influence of science
Science Inquiry Skills	Questioning and predicting
	Planning and conducting
	Processing and analysing data and information
	Evaluating
	Communicating

The seven general capabilities are a key dimension of the Australian Curriculum. They encompass knowledge, skills, behaviours and dispositions that, together with curriculum content in each learning area and the cross-curriculum priorities, can assist students to live and work successfully in the twenty-first century.

The capabilities identified as being most relevant and appropriate to the assessment of science, and hence reflected in NAP–SL, included the following:

- **Literacy:** aspects of the literacy capability are found within the reading comprehension demands of both the stimuli and the items of NAP–SL.
- **Numeracy:** aspects of the numeracy capability are found within NAP–SL, including the reading and construction of graphs and tables, calculations and measurement, as well as some elements of spatial reasoning.
- **Information and Communication Technology (ICT):** aspects of the ICT capability will arise from online delivery.
- **Critical and Creative Thinking:** aspects of the critical and creative thinking capability arise from important cognitive skills inherent in scientific inquiry.

Items and stimulus also drew on aspects of the personal and social capability, the ethical understanding capability, and the intercultural understanding capability when appropriate.

An important new feature of the 2018 NAP–SL Assessment Framework is the explicit definition of a cognitive dimension within the assessment of science literacy and across all three content domains. The addition of cognitive dimensions is consistent with many national and international frameworks, such as TIMSS, PISA and NAP – Civics and Citizenship.

The cognitive dimension seeks to make explicit the thinking skills and intellectual processes that will be engaged by the students to respond to the assessment tasks. The cognitive dimension includes three cognitive processes that underpin what students are required to do in a task. These are:

- Knowing and using procedures
- Reasoning, analysing and evaluating
- Synthesising and creating

See appendices 1 and 2 of the 2018 NAP – Science Literacy Public Report or the NAP website for full descriptions of the strands, sub-strands, general capabilities and cognitive dimensions included in the 2018 NAP–SL Assessment Framework.

Pilot study 2017

As part of the framework review, ACARA conducted a small-scale pilot study in 58 schools from 16 October to 3 November 2017 with 1658 Year 8 students from across Australia. The sample included students from major cities, inner regional, outer regional and remote areas of all states and territories. The schools also came from a range of socio-economic backgrounds with ICSEA scores ranging from 882 to 1218.

The purpose of the pilot was to provide empirical evidence for the new draft assessment framework about the progression of knowledge and skills, the development of age-appropriate assessments for Year 10 students and the vertical linking of Year 6 and 10 assessments. The tests administered to Year 8 students consisted predominantly of vertical link items with some unique Year 6 and unique Year 10 items to cover both ends of an expected Year 8 student ability range.

The analyses of the 2017 pilot data demonstrated robust psychometric properties of the items that would assist the trial and main study to achieve its aims of broadening the assessment to measure Year 10 science literacy, and provide more information on the development of age-appropriate assessments for Year 10 students and the vertical linking of Year 6 and 10 assessments. The results from this pilot also supported vertical equating of Year 6 and Year 10 tests and the capacity for the assessment to collect more information about the progression of science literacy between Year 8 and 10.

Mode effect study

A mode-effect study was designed to investigate the effect of a change in delivery from a paper-based to a computer- based assessment in the NAP–SL context. The outcome of this study was intended to inform 1) comparability of online results in 2018 and 2) the effort needed to place the results of the online NAP–SL 2018 onto the historical scale. Forty schools from Australian Capital Territory, New South Wales, Queensland, South Australia, Tasmania, Victoria and Western Australia were selected to participate in the study. In each school, approximately 20 - 25 students participated.

The mode effect test (36 historically linked items) contained two parts: Part A and Part B. Part A was the first half of the test; Part B was the second half of the test. Each part had a paper and an online version. Schools were randomly assigned into two groups. Group 1 (n=397) sat Part A on computer and Part B on paper while Group 2 (n=366) took Part A on paper and Part B on computer. The Rasch measurement model, using ACER ConQuest, was applied to calibrate items, perform DIF analysis and investigate the impact of mode effect at both test and item levels. The results of this analysis show that link items were easier when appeared on paper regardless. Consequently, a shift of 0.131 logits was added to the NAP–SL 2018 results to correct for this mode effect. More details about the mode effect study are included in Appendix 7.

Development of NAP–SL 2018 assessment

Audit of existing test item pool

All existing test and survey items were reviewed for linkage to the Australian Curriculum: Science and their psychometric viability. Consequently, content and skills areas were identified which required additional items to ensure coverage of all strands in the curriculum.

Item writing workshop

In consultation with jurisdictions, ACARA invited 19 primary and secondary teachers to attend a three-day workshop lead by ACARA’s Online Assessment Specialist. The workshop introduced the teachers to the NAP Sample program and the NAP – Science Literacy assessment. Participants then received training in writing diagnostic assessment items for an online environment. Using the 2018 NAP–SL Assessment Framework and content sequences as a guide, the teachers worked in teams to generate item sets which were then panelled by the whole group prior to the end of the workshop. Following the workshop, the new test items, along with existing pool items, were allocated to test forms for the field trial.

Field trial

The field trial was conducted in June 2018 with 1380 Year 6 students in 37 schools and 1107 Year 10 students in 35 schools. The field trial was conducted in ACT (1 school), New South Wales (31 schools), Queensland (17 schools), South Australia (1 school), Victoria (20 schools) and Western Australia (2 schools).

The major purpose of the field trial was to test field operations, the assessment platform and the psychometric properties of the items. Data collected from the field trial informed the implementation of the main study. The design and composition of the field trial test booklets are outlined early in this chapter.

Overall, the analysis of the collected data suggested that the field operations procedures, test instrument, scoring guides and scoring procedures had been successful and would form a solid foundation for the 2018 main study. As a result of findings from the field trial, there were a number of small changes made to different aspects of the instruments, guides and procedures, such as the addition of examples of student performance, some clarifications of wording in the scoring guides, and refinements of the test administration login system to make the data entry of student information by test administrators more efficient.

Review of test items

Following the field trial data analysis, the proposed final set of items for Main Study were reviewed by the members of the NAP–SL Working group and ACARA personnel. The reviewers judged the items against a range of criteria including:

Alignment with assessment framework: to ensure that items fit within the assessment framework and matched the specified strands, levels and concept areas.

Language demand: science stimulus may require some complex language, but it is important that the language is kept as simple as feasible.

Scientific accuracy: the science presented needed to be correct. In some cases, complex scientific ideas were explained in a simplified way suitable for the age of the audience.

Free from bias: items and stimulus were examined to ensure they were free from cultural or gender bias.

Metadata: the classifications of the items against multiple criteria were examined.

Item structure: the items were also examined in terms of how well they were likely to perform in a psychometrically validated test.

Main study

Main study was conducted from mid-October to mid-November 2018 and was to be attempted by 8621 students from Year 6 and Year 10 (5,578 from Year 6 and 3,043 from Year 10). These students were sampled randomly from 546 schools. The final participation numbers were 5,551 students from Year 6 and 3,032 students from Year 10. See chapter 3 for details relating to the student sample and participation numbers.

Following the closure of the test window, human marking of extended text responses occurred during November.

Data files for analysis were compiled between December and January 2019. Student background data were collected from schools and education systems during the main study by ACER. See chapter 4 for more details relating to student background data collection.

Analysis of final data set occurred during early 2019. See later chapters for more detailed information relating to the data analysis.

Standard – setting workshop

Twenty secondary science teachers from across Australia participated in a two-day workshop to recommendation a new proficient standard for Year 10. Following the workshop, extensive psychometric analysis occurred which resulted in a change to the width of the existing levels in the NAP–SL scale. See chapter 7 for a more detailed description of the standard – setting workshop.

Test design for 2018

Test specifications

Item and test development were based on the following specifications:

- develop/select approximately 140 items in total for the final test forms for both years (including historical link items for Year 6 and vertically linked items for Years 6 and 10)
- items to be presented as item sets with each group of contextually linked items associated with a stimulus, with a minimum of three items per set
- provide sufficient objective and inquiry assessment items for up to one hour of testing for each student in the national sample
- develop a diverse range of online test item types
- balance the core item types within the trial item pool to be approximately
    - 60 per cent multiple-choice and non-multiple-choice short response
    - 40 per cent extended text response

  the balance between process items (*Science Inquiry Skills*) and conceptual items (*Science Understanding and Science as a Human Endeavour*) would be approximately in the proportion half process and half conceptual items.

The assessment itself would be split into three parts:

- an objective test consisting of a mix of items gathered into thematically related item sets
- a set of inquiry tasks consisting of sets of items organised into a sequence that mimic the stages of a science investigation
- a set of survey questions to determine student attitudes to and interests in science and their science experiences in school.

Test booklet design

A rotational, incomplete block design was adopted for the NAP–SL assessment in order to provide a comprehensive coverage of content in science while minimising the task load on students taking part in the assessment. A rotational design minimises the effect of biased item parameters caused by varying item positions within the test booklets. Such a test design has been accepted as standard in large-scale assessments such as TIMSS and PISA and in other NAP tests.

It was established in NAP–SL 2006 that the 7-booklet rotational design provides a balance in terms of content coverage and test administration requirements. This design was used in the 2009, 2012 and 2015 cycles, thus the same design was proposed for the 2018 assessment.

Online item types

The online delivery of the NAP–SL assessment has broadened the types of test items that can be incorporated into the test. The item types used in 2018 included multiple choice, a range of interactive non multiple-choice short response items and constructed or extended text responses. Extended text responses required responses from a few words to a maximum of two paragraphs. See 2018 NAP–SL Public Report for a full description of the online item types.

Use of multimedia

The online test delivery platform can accommodate audio. The text of all stimuli was professionally recorded and was played to students when they landed on a screen containing a stimulus. The provision of the audio reduced the reading load for students.

Construction of field trial test booklets

In preparation for the introduction of a Year 10 test in 2018, a pilot study was undertaken in 2017. The items developed for the pilot study were audited and refined to form part of the pool of potential items for the 2018 assessment.

Existing Year 6 items were also audited to ascertain their suitability for 2018 from both a psychometric viewpoint and their link to the Australian Curriculum: Science. Items which were not able to be mapped to Year 6 were removed from the potential 2018 pool of items. These reviews allowed curriculum gaps to be identified which then required additional items to be developed to fill these gaps prior to assembling the trial test forms.

Available objective items were grouped to form eight clusters, C1–C8, for each year. Each test form contained two clusters of objective items and an inquiry task. The structure of the field trial test forms is shown in table 2.2.

Table 2.2 Structure of field trial test forms

Trial Forms	Cluster 1	Cluster 2	Inquiry task
Test form 1	C1	C2	Task 1
Test form 2	C2	C3	Task 2
Test form 3	C3	C4	Task 3
Test form 4	C4	C5	Task 4
Test form 5	C5	C6	Task 4
Test form 6	C6	C7	Task 3
Test form 7	C7	C8	Task 2
Test form 8	C8	C1	Task 1

Each Year 6 cluster contained approximately 20 items whilst Year 10 clusters contained approximately 25 items. There were four inquiry tasks for each year. Tasks 1 and 2 for each year were year specific whilst tasks 3 and 4 were allocated to both years.

Cluster 2 in Year 6 contained historical link items drawn from 2015. To allow vertical linking between Years 6 and 10, a significant number of items, which were successfully trialled, were included across the clusters in Year 6 except cluster 2. Sixteen item sets and two inquiry tasks were presented to both Years 6 and 10 in the field trial.

The allocation of the eight tests forms were randomly assigned to students as they logged into the online test platform.

Items selected for Field trial

The composition of Year 6 and 10 items selected for field trial, excluding the historical items, is presented in table 2.3.

Table 2.3 Characteristics of items selected for field trial

Australian Curriculum: Science strand	Objective items	Inquiry task items	Total
Science as a Human Endeavour	17	8	25
Science Inquiry Skills	87	82	169
Science Understanding	162	2	164
Cognitive dimensions			
Knowing and using skills	185	58	243
Reasoning, analysing and evaluating	74	32	106
Synthesising and creating	9	3	12
Test item types			
Extended text	28	24	52
Hotspot	7	2	9
Inline choices	26	10	36
Interactive gap match	21	5	26
Interactive graphic gap match	6	1	7
Interactive order	7		7
Multiple choice	128	39	167
Multiple choices	27	7	34
Position object	2		2
Select point	1	3	4
Text entry	1		1
Total	**254**	**91**	**345**

Construction of Main Study test booklets

The test design for Year 6 Main Study is presented in table 2.4 where C1 to C7 denotes seven different clusters of items, each containing approximately 15 - 20 minutes of testing material. Year 6 students were presented with one of two inquiry tasks. Task 1 (*Beaks*) was the Year 6 only task. Task 2 (*Bouncing Balls*) was given to both years.

Test number 8 for Year 6 contained only the historical link items including 19 items from the 2006, 2009 and 2012 tests and an additional 18 secure items from the 2015 test.

To allow vertical linking between Years 6 and 10, eight item sets were included across C1 to C7 in Year 6.

Each student was administered one online test. All eight tests were randomly allocated as each student logged into the test platform.

Table 2.4 Test design for Year 6 Main Study

Test	Block 1	Block 2	Block 3	Inquiry task
1	C1	C2	C4	Task 1
2	C2	C3	C5	Task 2
3	C3	C4	C6	Task 1
4	C4	C5	C7	Task 2
5	C5	C6	C1	Task 1
6	C6	C7	C2	Task 2
7	C7	C1	C3	Task 1
8	37 historical links items			Task 2

The test design for Year 10 Main Study is presented in table 2.5 where C1 to C8 denotes eight different clusters of items, each containing approximately 15 - 20 minutes of testing material. Year 10 students were each presented with one of two inquiry tasks. Task 1 (*Artificial Glaciers*) was the Year 10 only task. Task 2 (*Bouncing Balls*) was given to both years.

To allow vertical linking between Years 6 and 10, eight item sets were included across the eight clusters in Year 10.

Each student was administered one online test. All eight tests were randomly allocated as each student logged into the test platform.

Table 2.5 Test design for Year 10 Main Study

Test	Block 1	Block 2	Block 3	Inquiry task
1	C1	C2	C4	Task 1
2	C2	C3	C5	Task 2
3	C3	C4	C6	Task 1
4	C4	C5	C7	Task 2
5	C5	C6	C8	Task 1
6	C6	C7	C1	Task 2
7	C7	C8	C2	Task 1
8	C8	C1	C3	Task 2

The total item pool included 140 items for Year 6 students and 136 items for Year 10 students.

Items selected for Main Study

Table 2.6 shows the breakdown of the test items selected for Main Study, including the historical and vertical items.

Note that in classifying and mapping the test items into the strands and sub strands of the Australian Curriculum: Science, the cognitive demands of the item was used rather than the context portrayed in the stimulus for the item set. For example, an item may have stimulus relating to biology/living things, but the item may require the students to analyse data presented in a table which would see it mapped as a Science Inquiry Skills.

Table 2.6 Characteristics of items selected for Main Study

Australian Curriculum: Science strand	**Objective items**		**Inquiry task items**	
	Year 6	**Year 10**	**Year 6**	**Year 10**
Science as a Human Endeavour	10	9	2	2
Science Inquiry Skills	51	35	19	21
Science Understanding	58			1
Science Understanding sub strand				
Biological sciences	16	23		
Chemical sciences	18	16		
Earth and space sciences	10	12		
Physical sciences	14	17		1
Science as a Human Endeavour sub strand				
Nature and development of science	1	2		
Use and influence of science	9	7	2	2
Science Inquiry Skills sub strand				
Questioning and predicting	4		1	2
Planning and conducting	12	10	9	11
Processing and analysing data and information	34	25	6	7
Evaluating	1		3	1
Communicating				
Cognitive dimensions				
Knowing and using skills	82	78	13	16
Reasoning, analysing and evaluating	35	31	8	6
Synthesising and creating	2	3		2
Test item types				
Extended text	24	11	7	7
Hotspot	4	3		
Inline choices	8	13	4	1
Interactive gap match	7	9	2	2
Interactive graphic gap match	1	2	1	
Interactive order	2	5		1
Multiple choice	57	53	5	11
Multiple choices	15	16	2	2
Text entry	1			
Total	119	112	21	24

The 2018 NAP-SL Assessment Framework allowed for extended test items to be worth more score points than in previous cycles. Consequently, the extended items that were trialled in 2018 and then were included in Main Study were worth a range of score points from one to three. Some extended text items had two parts worth up to three score points for each part. Increasing the score value of these items allowed for an increased amount of diagnostic information being supplied to schools as the marking rubrics allowed for differentiation to occur with the quality of the students' responses.

Inquiry task design

All NAP–SL cycles have included an inquiry task component. The purpose of this component is to provide students with an opportunity to experience practical aspects of science within a formal assessment and assess the conventions of science literacy in more depth than was possible in the objective component.

Six new online inquiry tasks were developed for the 2018 trial. There were two tasks specifically aimed at Year 6, two tasks specifically aimed at Year 10 students and two tasks that were trialled in both Years 6 and 10. In the final 2018 assessment, three tasks which demonstrated the most robust measurement characteristics in trialling were administered to each year level. One inquiry task was common for both Years 6 and 10.

The three inquiry tasks delivered in the Main Study were *Beaks* (Year 6 only); *Artificial Glaciers (*Year 10 only) and *Bouncing Balls* (common task for Years 6 and 10). Each inquiry task contained between 10 and 12 items which followed a simulated investigation linked to a presented context. Each task commenced by introducing the students to the context and then stepped the students through the components of the scientific method for a linked investigation. Students were then required to apply the results of the simulated investigation to the original context. Each student was present with one inquiry task.

Student survey

As was the case for previous cycles of the NAP–SL assessment (2009–2015), there was a survey for students incorporated into the instrument.

The 2018 survey included some questions which were used in previous cycles for Year 6 along with some additional items for Year 10 students. The previously used questions allowed historical comparisons to occur for Year 6 students whilst some items were presented to both years to allow for vertical comparisons between Years 6 and 10.

The questions in the survey covered the following areas:

- Interest in Science
- Self-concept of science ability
- Value of Science
- Science teaching 1
- Time spent on Science
- Science teaching 2.

A copy of the student survey can be found in Appendix 1.

Chapter 3 SAMPLING and WEIGHTING PROCEDURES

Jorge Fallas – Australian Council for Educational Research

Martin Murphy – Australian Council for Educational Research

Kate O’Malley – Australian Council for Educational Research

This chapter describes the NAP–SL 2018 sample design, the achieved sample, and the procedures used to calculate the sampling weights. The sampling and weighting methods were used to ensure that the data provided accurate and efficient estimates of the achievement outcomes for the Australian Year 6 and Year 10 student populations.

Sampling

The target populations for the study were Year 6 and Year 10 students enrolled in educational institutions across Australia. In 2018, Year 10 students were also included in the target population for the first time.

A two-stage stratified cluster sample design was used in NAP–SL 2018, like that used in other Australian national sample assessments and in international assessments such as the Trends in International Mathematics and Science Study (TIMSS). The first stage consisted of a sample of schools, grouped in strata according to a combination of state and sector. Within each stratum, each school was sorted by performance on the 2017 NAPLAN test, geographic location and school size. The second stage consisted of a sample of 20 random students from the target year level in sampled schools. Samples were drawn separately for each year level.

The sampling frame

Schools were selected from the school sampling frame provided by ACARA, a comprehensive list of all schools in Australia, updated annually.

School exclusions

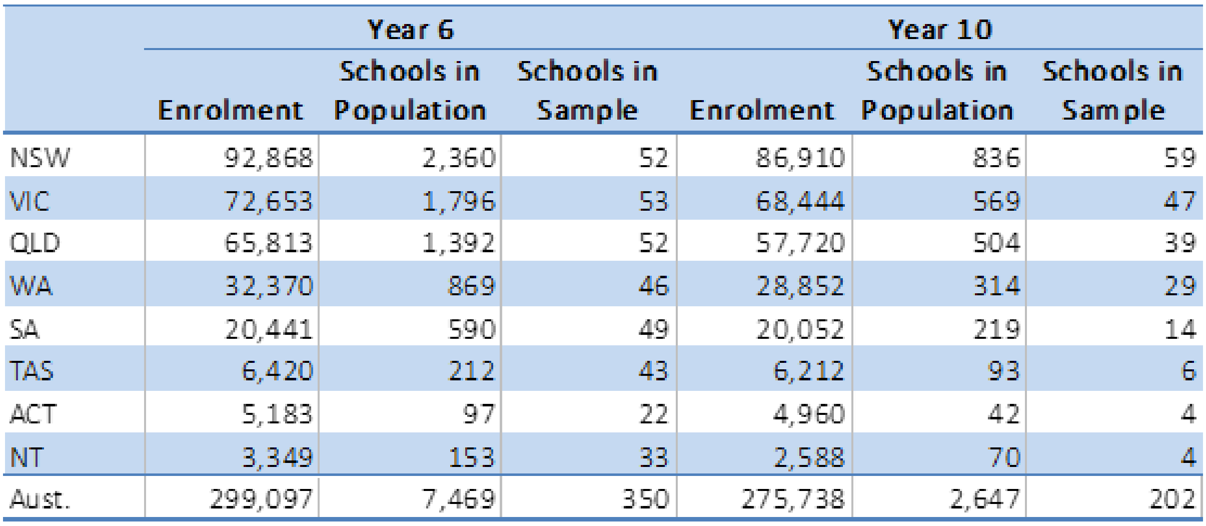
All schools that reported any student enrolment in Year 6 or Year 10 were considered part of the respective Year 6 and Year 10 target population. Schools excluded from the target population included: non-mainstream schools, such as schools for students with intellectual disabilities, hospital schools or distance education schools, among others. These exclusions accounted for 0.3 per cent of the Year 6 student population and 1.0 per cent of the Year 10 student population.

The designed sample

For both Year 6 and Year 10 samples, sample sizes were chosen to provide accurate estimates of achievement outcomes for all states and territories. As with previous studies at the Year 6 level, the expected 95 per cent confidence intervals were estimated in advance to be within approximately ±0.15 to ±0.2 of the population standard deviation for estimated means of the larger states. Confidence intervals of this magnitude require an effective sample size[3] of around 100-150 students in the larger states. This level of precision was considered an appropriate balance between the analytical demands of the study, the burden on individual schools and the overall costs of the study. The main requirement for achieving acceptable precision for a state or territory is to have a good-sized sample. Although a less important factor, sampling a larger proportion of the population will also improve precision. As the proportion of the total population surveyed becomes larger, the precision of the sample increases for a given sample size: This explains why the sample sizes for the smaller states and territories are smaller compared to the larger states and territories.

As 2018 was the first year of implementation of Science Literacy at the Year 10 level, and as 2018 was a year when both TIMSS and PISA were also in the field, there was concern about the burden of survey work across jurisdictions. Jurisdictions were consulted about whether they wished to have a Year 10 Science Literacy sample of a size to achieve similar precision as described above for year 6, or whether, for this first round of implementation, they wished to reduce the sample size to contribute to national estimates only. As can be observed in Table 3.1, the smaller jurisdictions took the latter option, and hence the overall sample size at the year 10 level was reduced.

Table 3.1 Year 6 and Year 10 target population and designed samples by state and territory



	Year 6			Year 10		
	Enrolment	Schools in Population	Schools in Sample	Enrolment	Schools in Population	Schools in Sample
NSW	92,868	2,360	52	86,910	836	59
VIC	72,653	1,796	53	68,444	569	47
QLD	65,813	1,392	52	57,720	504	39
WA	32,370	869	46	28,852	314	29
SA	20,441	590	49	20,052	219	14
TAS	6,420	212	43	6,212	93	6
ACT	5,183	97	22	4,960	42	4
NT	3,349	153	33	2,588	70	4
Aust.	299,097	7,469	350	275,738	2,647	202

Two sampling stages

Stratification by state and sector was explicit: separate samples were drawn for each sector within states and territories. Stratification by NAPLAN performance and Geographic Location was implicit: schools within each state and sector were ordered by size (according to the number of students in the target year level) within subgroups defined by a combination of NAPLAN performance quintile within each state and geographic location.

The selection of schools was carried out using a systematic probability-proportional-to-size (PPS) method. For large schools, the measure of size (MOS) was equal to the enrolment at

[3] The effective sample size is the sample size of a simple random sample that would produce the same precision as that achieved under a complex sample design.

the target year. The sum of the measures of size of schools within a stratum is calculated, and divided into *n* equal-sized intervals, where *n* is the number of schools to be sampled from the stratum. The school selection probability is equal to the measure of size of the school divided by the interval size:

$$\text{Pr (school selection)} = \text{MOS}_{\text{school}} / (\sum \text{MOS}_{\text{all schools in stratum}} / \text{n})$$

The number of students to be sampled from the school is known as the 'target cluster size' (TCS). Students are sampled from the school with equal probability and so the selection probability of a student from a larger school is:

$$\text{Pr (student selection within school)} = \text{TCS/ MOS}_{\text{school}}$$

The combined effect of this two-stage process is that most students are sampled with equal probability:

$$\text{Pr (student selection)} = \text{Pr(school selection)} * \text{Pr (student selection within school)}$$

$$= \text{MOS}_{\text{school}} / (\sum \text{MOS}_{\text{all schools in stratum}} / \text{n}) * \text{TCS/ MOS}_{\text{school}}$$

$$= \text{TCS/ } (\sum \text{MOS}_{\text{all schools in stratum}} / \text{n})$$

If a school is selected with target year enrolment less than the TCS (denoted as a 'small school'), all students from that school will be certain selections and the second term in the above expression becomes 1:

$$\text{Pr (student selection)} = \text{Pr(school selection)} * \text{Pr (student selection within school)}$$

$$= \text{MOS}_{\text{small school}} / (\sum \text{MOS}_{\text{all schools in stratum}} / \text{n}) * 1$$

$$= \text{MOS}_{\text{small school}} / (\sum \text{MOS}_{\text{all schools in stratum}} / \text{n})$$

In order to make the selection probability for these students the same as above, the starting point in the sample design is to set the measure of size for the smaller schools to TCS:

$$\text{MOS}_{\text{small school}} = \text{TCS}$$

$$\text{Pr (student selection)} = \text{TCS/ } (\sum \text{MOS}_{\text{all schools in stratum}} / \text{n})$$

For NAP–SL the TCS was set at 20 students. The starting point in the sample design is that all small schools with enrolments from 1 to 19, and all students from those schools, are sampled with equal probability.

This approach minimises variation in weights which is desirable. Large variations in weights can have a major impact on the precision of survey estimates.

The approach described above is used when small schools represent only a very small proportion of the total enrolment in the stratum. When the proportion of the total enrolment in small schools is larger, the number of schools to be sampled from the stratum is increased to cater for the fact that the yield from these smaller schools will be less than the target cluster size. In addition, the smallest of these smaller schools have their selection probabilities reduced, through a reduction in their measure of size, so that fewer of them are included in the sample, that is, they are under-sampled.

To under-sample small schools, all schools in the stratum are classified into one of the following groups based on their enrolment size:

- P1 ('extremely small'): enrolment of 2 or less
- P2 ('very small') enrolment between 3 and half the TCS
- Q ('moderately small'): enrolment from TCS/2 +1 to less than the TCS
- R ('large'): enrolment of TCS or larger

If the proportion of students in P1 and P2 schools in a stratum was 1% or more, or if the proportion of students in Q Schools was 4% or more, then the following adjustments were made:

1. The MOS for 'P1' schools was reduced to 0.25 TCS. In this case, with TCS = 20, the MOS for these extremely small schools is reduced to 5
2. The MOS for 'P2' schools was reduced to 0.5 TCS (i.e. MOS = 10)
3. The total number of schools to be sampled from the stratum is increased, to preserve the desired sample yield from the stratum to close to the product of the TCS and the number of schools to be sampled from the stratum (TCS * n).

The first two adjustments mean that the extremely small and very small schools are sampled at lower rates, to minimise the operational burden of having too many of these very small schools in the sample.

The net effect of these adjustments is that the desired yield from the sample is preserved, variation in weights is kept to a minimum, and the operational burden of having a large number of small schools included in the sample is reduced.

Due to the relatively high number of International Surveys carried out during 2018, as described above, it was decided to reduce the burden of schools that had been sampled for these surveys. For this reason, a school sample with minimum overlap control with TIMSS and PISA surveys was carried out. Schools who had already been selected for the PISA 2018 and the TIMMS Year 4 and Year 8 surveys had a lower probability of selection, adjusted using the methodology laid out in Chowdury et. al. (2000)[4]. These procedures make adjustments to the selection probability of schools based on the conditional probabilities of their selection in previous studies.

The standard process for the selection of schools with PPS is described as follows:

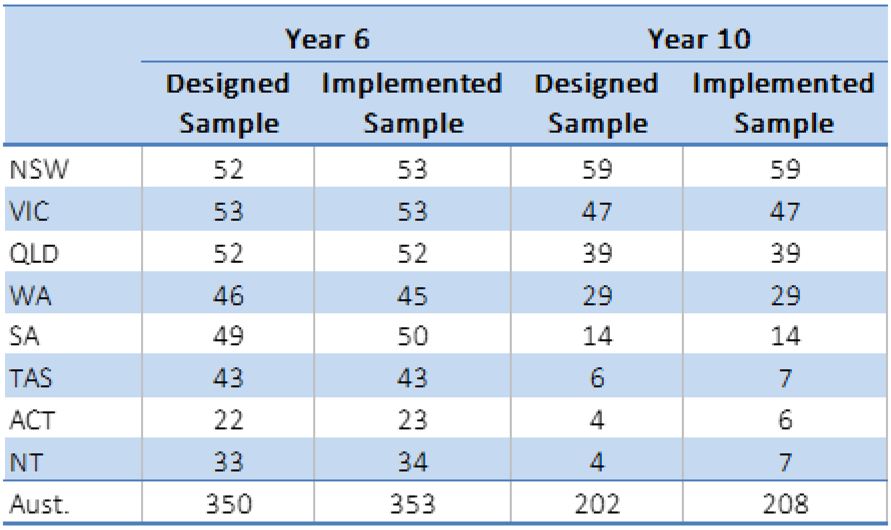
- The MOS was accumulated from school to school and the running total was listed next to each school. The total cumulative MOS was a measure of the size of the population of sampling elements. Dividing this figure by the number of schools to be sampled provided the sampling interval.
- The first school was sampled by choosing a random number between one and the sampling interval. The school who's cumulative MOS contained the random number was the first sampled school. By adding the sampling interval to the random number, a second school was identified. This process of consistently adding the sampling interval

[4] Chowdhury, S., Chu, A., & Kaufman, S. (2000). *Minimizing overlap in NCES surveys*. Proceedings of the Survey Methods Research Section. American Statistical Association, 174-179.

to the previous selection number resulted in a PPS sample of the required size.

On the basis of an analysis of small schools (schools with lower enrolments than the assumed cluster sample size of 20 students) undertaken prior to sampling, the school sample size in some strata was increased in order to ensure that the number of students sampled was close to expectations. As a result of both the small school analysis and overlap control, the actual number of schools sampled for Year 6 and Year 10 were 353 and 208, respectively. Both were slightly larger than the designed sample (see table 3.2). The actual sample drawn is referred to as the ‘implemented sample’.

Table 3.2 Year 6 and Year 10 designed and implemented samples by state and territory



	Year 6		Year 10	
	Designed Sample	Implemented Sample	Designed Sample	Implemented Sample
NSW	52	53	59	59
VIC	53	53	47	47
QLD	52	52	39	39
WA	46	45	29	29
SA	49	50	14	14
TAS	43	43	6	7
ACT	22	23	4	6
NT	33	34	4	7
Aust.	350	353	202	208

As each school was selected, the next school in the sampling frame was designated as a replacement school to be included in cases where the sampled school did not participate. The school before the sampled school was designated as the second replacement. It was used if neither the sampled nor the first replacement school participated. Due to the stratified sampling frame, the two replacement schools were generally similar (with respect to NAPLAN performance, geographic location and size) to the originally sampled school.

After the school sample had been drawn, several sampled schools were identified as meeting the criteria for exclusion. When this occurred, the sampled school and its replacements were removed from the sample and removed from the calculation of participation rates. Three schools were removed from the Year 6 sample and one school was removed from the Year 10 sample. These exclusions are included in the exclusion rates reported earlier.

Student exclusions

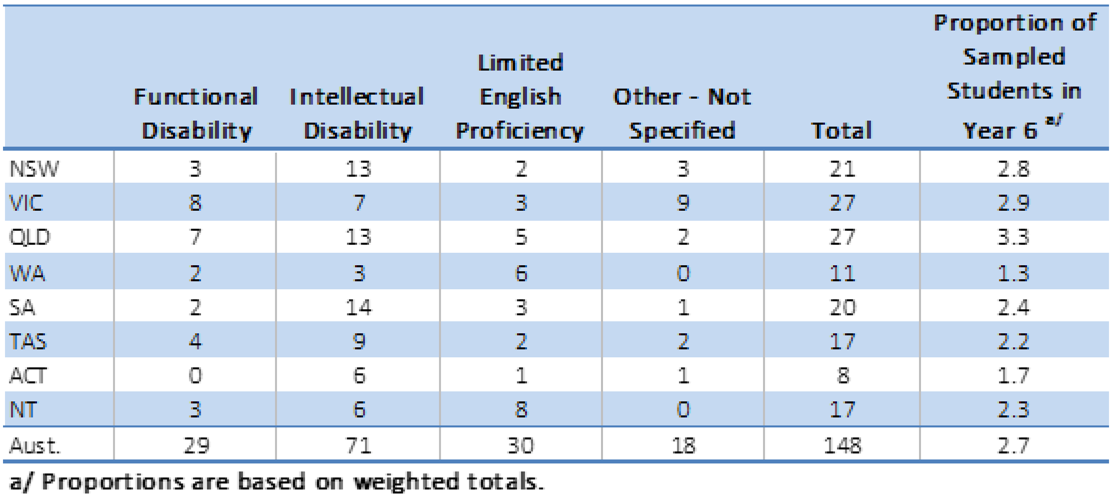
Within the group of sampled students, individual students were excluded from the assessment on the basis of the criteria listed below.

- *Functional disability*: Student has a moderate to severe permanent physical disability such that he/she cannot perform in the assessment situation.
- *Intellectual disability*: Student has a mental or emotional disability and is cognitively delayed such that he/she cannot perform in the assessment situation.

- *Limited assessment language proficiency*: The student is unable to read or speak the language of the assessment and would be unable to overcome the language barrier in the assessment situation. Typically, a student who has received less than one year of instruction in the language of the assessment would be excluded.

Tables 3.3 and 3.4 details the numbers and percentages of students excluded from the NAP-SL 2018 assessment, according to the reason given for their exclusion. The number of student-level exclusions was 148 at Year 6 and 136 at Year 10. This gives weighted exclusion rates of 2.7 per cent of the sampled Year 6 students and 4.1 per cent of sampled Year 10 students.

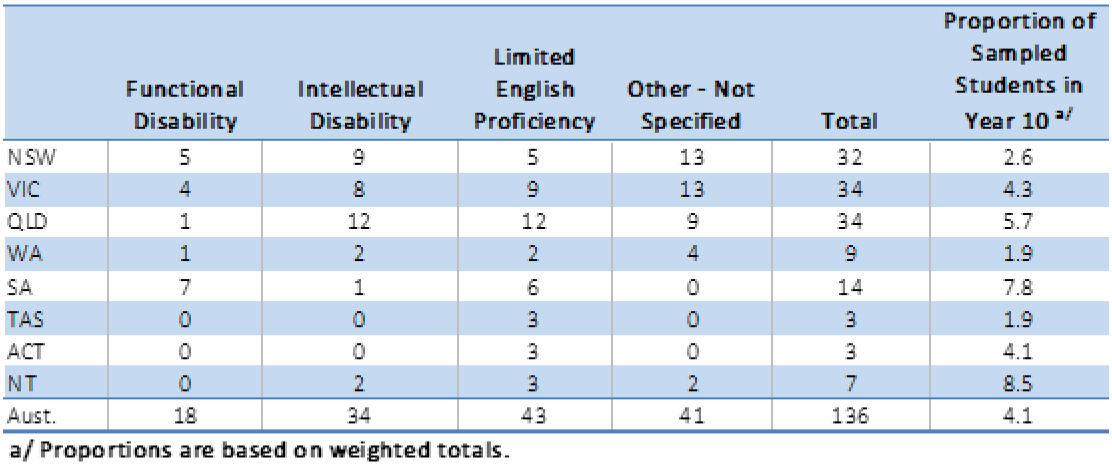
Table 3.3 Year 6 breakdown of student exclusions according to reason by state and territory



	Functional Disability	Intellectual Disability	Limited English Proficiency	Other - Not Specified	Total	Proportion of Sampled Students in Year 6 [a/]
NSW	3	13	2	3	21	2.8
VIC	8	7	3	9	27	2.9
QLD	7	13	5	2	27	3.3
WA	2	3	6	0	11	1.3
SA	2	14	3	1	20	2.4
TAS	4	9	2	2	17	2.2
ACT	0	6	1	1	8	1.7
NT	3	6	8	0	17	2.3
Aust.	29	71	30	18	148	2.7

a/ Proportions are based on weighted totals.

Table 3.4 Year 10 breakdown of student exclusions according to reason by state and territory



	Functional Disability	Intellectual Disability	Limited English Proficiency	Other - Not Specified	Total	Proportion of Sampled Students in Year 10 [a/]
NSW	5	9	5	13	32	2.6
VIC	4	8	9	13	34	4.3
QLD	1	12	12	9	34	5.7
WA	1	2	2	4	9	1.9
SA	7	1	6	0	14	7.8
TAS	0	0	3	0	3	1.9
ACT	0	0	3	0	3	4.1
NT	0	2	3	2	7	8.5
Aust.	18	34	43	41	136	4.1

a/ Proportions are based on weighted totals.

Weighting

While the multi-stage stratified cluster design provides a very economical and effective data collection process in a school environment, stratification, oversampling of sub-populations and non-response cause differential probabilities of selection for the ultimate sampling elements, the students. Consequently, one student in the assessment does not necessarily represent the same number of students in the population as another, as would be the case with a simple random sampling approach. To account for differential probabilities of selection due to the design and to ensure unbiased population estimates, a sampling weight was computed for each participating student. It was an essential characteristic of the sample design to allow the provision of proper sampling weights, since these were necessary for the computation of accurate population estimates.

The overall sampling weight is the product of weights calculated at the two stages of sampling:

- the selection of the school at the first stage
- the selection of students within the sampled schools at the second stage.

First stage weight

The first stage weight is the inverse of the probability of selection of the school, adjusted to account for school non-response.

The probability of selection of the school is equal to its measure of size (*MOS*)[5] divided by the sampling interval (*SINT*) or one, whichever is the lower. (A school with a *MOS* greater than the *SINT* is a certain selection and therefore has a probability of selection of one. Some very large schools were selected with certainty into the sample.)

The sampling interval is calculated at the time of sampling, and for each explicit stratum it is equal to the cumulative *MOS* of all schools in the stratum, divided by the number of schools to be sampled from that stratum.

This factor of the first stage weight, or the school base weight (BW_{sc}), was the inverse of this probability

$$BW_{sc} = \frac{SINT}{MOS}$$

Following data collection, counts of the following categories of schools were made for each explicit stratum:

[5] For larger schools, the measure of size is the number of students enrolled in Year 6 or Year 10. If under-sampling of small schools is required, the following adjustments were made to measures of size:

- For schools with an estimated enrolment of 2 or less students, the measure of size was set to 5.
- For schools with an estimated enrolment of more than 2 but less than 10, the measure of size was set to 10.
- For schools with an estimated enrolment between 11 and 19, the measure of size was set to 20.

- the number of schools that participated (n_p^{sc})
- the number of schools that were sampled but should have been excluded (n_x^{sc})
- the number of non-responding schools (n_n^{sc}).

Note that $n_p^{sc} + n_x^{sc} + n_n^{sc}$ equals the total number of sampled schools from the stratum.

Examples of the second class (n_x^{sc}) were:

- a sampled school that no longer existed
- a school that, following sampling, was discovered to have fitted one of the criteria for school-level exclusion (e.g. very remote, very small), but which had not been removed from the frame prior to sampling.

In the case of a non-responding school (n_n^{sc}), neither the originally sampled school nor the schools identified as possible substitutes participated.

Within each explicit stratum, an adjustment was made to account for school non-response. This non-response adjustment (*NRA*) for a stratum was equal to:

$$NRA_{strt} = \frac{\left(n_p^{sc} + n_n^{sc}\right)}{n_p^{sc}}$$

The first stage weight, or the final school weight, was the product of the inverse of the probability of selection of the school and the school non-response adjustment:

$$FW_{sc} = BW_{sc} * NRA_{strt}$$

Second stage weight

Following data collection, counts of the following categories of students were made for each sampled school:

- the total number of students in a school at relevant year level (n_{tot}^{st})
- the number of sampled students who participated (n_p^{st})
- the number of sampled students who were exclusions (n_x^{st})
- the number of non-responding, sampled students (n_n^{st}).

Note that $n_{samp}^{st} = n_p^{st} + n_x^{st} + n_n^{st}$equals the total number of sampled students from the sampled school.

The first factor in the second stage weight was the inverse of the probability of selection of the student from the sampled school.

$$BW_{st} = \frac{n_{tot}^{st}}{n_{samp}^{st}}$$

The student level non-response adjustment was calculated for each school as:

$$NRA_{sc} = \frac{n_p^{st} + n_n^{st}}{n_p^{st}}$$

The final student weight was:

$$FW_{st} = BW_{st} \times NRA_{sc}$$

Overall sampling weight

The full sampling weight (*FWGT*) was simply the product of the weights calculated at each of the two sampling stages:

$$FWGT = FW_{sc} \times FW_{st}$$

After computation of the overall sampling weights, the weights were checked for outliers, because outliers can have a large effect on the computation of the standard errors. A weight was regarded as an outlier if the value was more than four times the median weight within a subpopulation defined by year level, state or territory and sector (i.e. an explicit stratum). There were sixteen cases of outliers in the data for year 6 and 14 cases in Year 10, so these weights were trimmed to four times the median weight.

Post-Stratification Adjustment

A final adjustment to the weights was carried out, so the total sum of the weights would reflect the sum for the target population in each state and sector. For this purpose, the sample was divided by Year Level, State, Sector and Gender. For each combination of these categories a Post Stratification Adjustment (PSADJ) factor was estimated as follows:

$$PSADJ_{j,k,l} = \frac{popest_{j,k,l}}{\sum_{j,k,l} FWGT}$$

Where *popest*$_{j,k,l}$ is the total student population for the corresponding year level (Year 6 and Year 10) in state [j], sector [k], and gender [l] according to the latest estimates provided by the ABS[6]. The value in the denominator is the total sum of final weights for the corresponding combination of state, sector and gender within each year level. The final student weight used for analysis is therefore:

$$wt\_2018 = FWGT \times PSADJ$$

Table 3.5 shows the resulting post-stratification adjustments for NAP–SL 2018[7].

[6] From Table 42b of the Australian Bureau of Statistics (ABS) Schools Australia Report, available here http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4221.02017?OpenDocument

[7] Post-stratification adjustments are not to be applied as an automatic last step when creating student weights. In every study, the advantages and disadvantages of this adjustment need to be evaluated.

Table 3.5 Post-stratification adjustments for NAP–SL 2018

State	Sector	Year 6	Year 10
ACT	C	1.13	0.63
ACT	G	0.83	0.42
ACT	I	0.62	0.27
NSW	C	0.99	1.30
NSW	G	0.90	0.81
NSW	I	0.83	0.63
NT	C	0.89	0.53
NT	G	0.88	0.50
NT	I	1.33	0.26
QLD	C	0.90	0.76
QLD	G	0.93	0.86
QLD	I	1.01	0.73
SA	C	1.00	1.04
SA	G	0.88	0.66
SA	I	1.07	1.11
Tas.	C	0.91	0.46
Tas.	G	0.93	0.28
Tas.	I	1.34	1.63
Vic.	C	0.92	0.97
Vic.	G	0.87	0.84
Vic.	I	1.12	0.97
WA	C	1.01	1.49
WA	G	0.89	0.98
WA	I	1.15	1.30

Participation rates

Separate participation rates were computed: (1) with replacement schools included as participants, and (2) with replacement schools regarded as non-respondents. In addition, each of these rates was computed using unweighted and weighted counts. In any of these methods, a school and a student response rate were computed, and the overall response rate was the product of these two response rates. The differences in computing the four response rates are described below. These methods are consistent with the methodology used in TIMSS (Olson, Martin & Mullis, 2013).

Unweighted response rates including replacement schools

The unweighted school response rate, where replacement schools were counted as responding schools, was computed as follows:

$$RR_1^{sc} = \frac{n_s^{sc} + n_{r1}^{sc} + n_{r2}^{sc}}{n_s^{sc} + n_{r1}^{sc} + n_{r2}^{sc} + n_{nr}^{sc}}$$

where n_s^{sc} is the number of responding schools from the original sample, $n_{r1}^{sc} + n_{r2}^{sc}$ is the total number of responding replacement schools, and n_{nr}^{sc} is the number of non-responding schools that could not be replaced.

The student response rate was computed over all responding schools. Of these schools, the number of responding students was divided by the total number of eligible, sampled students.

$$RR_1^{st} = \frac{n_r^{st}}{n_r^{st} + n_{nr}^{st}}$$

where n_r^{st} is the total number of responding students in all responding schools and n_{nr}^{st} is the total number of eligible, non-responding, sampled students in all responding schools.

The overall response rate is the product of the school and the student response rates.

$$RR_1 = RR_1^{sc} \times RR_1^{st}$$

Unweighted response rates excluding replacement schools

The difference of the second method with the first is that the replacement schools were counted as non-responding schools.

$$RR_2^{sc} = \frac{n_s^{sc}}{n_s^{sc} + n_{r1}^{sc} + n_{r2}^{sc} + n_{nr}^{sc}}$$

This difference had an indirect effect on the student response rate because fewer schools were included as responding schools and student response rates were only computed for the responding schools.

$$RR_2^{st} = \frac{n_r^{st}}{n_r^{st} + n_{nr}^{st}}$$

The overall response rate was again the product of the two response rates.

$$RR_2 = RR_2^{sc} \times RR_2^{st}$$

Weighted response rates including replacement schools

For the weighted response rates, sums of weights were used instead of counts of schools and students. School and student base weights (*BW*) are the weight values before correcting for non-response, so they generate estimates of the population being represented by the responding schools and students. The full weights (*FW*) at the school and student levels are the base weights corrected for non-response.

School response rates are computed as follows:

$$RR_3^{sc} = \frac{\sum_i^{s+r1+r2}\left(BW_i \times \sum_j^{r_i}(FW_{ij})\right)}{\sum_i^{s+r1+r2}\left(FW_i \times \sum_j^{r_i}(FW_{ij})\right)}$$

where i indicates a school, $s + r1 + r2$ all responding schools, j a student, and r_i the responding students in school *i*. First, the sum of the student final weights FW_{ij} for the responding students from each school was computed. Second, this sum was multiplied by the school's *BW* (numerator) or the school's *FW* (denominator). Third, these products were summed over the responding schools (including replacement schools). Finally, the ratio of these values was the response rate.

As in the previous methods, the numerator of the school response rate is the denominator of the student response rate:

$$RR_3^{st} = \frac{\sum_i^{s+r1+r2}\left(BW_i \times \sum_j^{r_i}(BW_{ij})\right)}{\sum_i^{s+r1+r2}\left(BW_i \times \sum_j^{r_i}(FW_{ij})\right)}$$

The overall response rate is the product of the school and student response rates:

$$RR_3 = RR_3^{sc} \times RR_3^{st}$$

Weighted response rates excluding replacement schools

Practically, replacement schools were excluded by setting their school *BW* to zero and applying the same computations as above. More formally, the parts of the response rates are computed as follows:

$$RR_4^{sc} = \frac{\sum_i^{s}\left(BW_i \times \sum_j^{r_i}(FW_{ij})\right)}{\sum_i^{s+r1+r2}\left(FW_i \times \sum_j^{r_i}(FW_{ij})\right)}$$

$$RR_4^{st} = \frac{\sum_i^{s}\left(BW_i \times \sum_j^{r_i}(BW_{ij})\right)}{\sum_i^{s}\left(BW_i \times \sum_j^{r_i}(FW_{ij})\right)}$$

$$RR_4 = RR_4^{sc} \times RR_4^{st}$$

Reported participation rates

The Australian school participation rate in Year 6 was 97 per cent when including replacement schools and 94 per cent when excluding replacement schools. In Year 10, the respective percentages were 96 per cent and 90 per cent. These are the unweighted response rates and are very similar to the weighted response rates.

Overall unweighted participation rates including replacement schools were 87 per cent for Year 6 and 79 per cent for Year 10.

Tables 3.5 and 3.6 detail the Years 6 and Year 10 participation rates according to the four methods described above.

Table 3.6 Overall school and student participation rates in Year 6

	Unweighted, including replacement schools			Unweighted, sampled schools only			Weighted, including replacement schools			Weighted, sampled schools only		
	Overall	School	Student	Overall	School	Student	Overall	School	Student	Overall	School	Student
NSW	0.92	1.00	0.92	0.90	0.98	0.92	0.91	1.00	0.91	0.89	0.98	0.91
VIC	0.89	0.98	0.90	0.87	0.96	0.90	0.88	0.98	0.89	0.86	0.96	0.89
QLD	0.80	0.92	0.87	0.77	0.88	0.87	0.82	0.94	0.87	0.78	0.89	0.87
WA	0.89	0.98	0.91	0.89	0.98	0.91	0.89	0.98	0.90	0.89	0.98	0.90
SA	0.88	0.98	0.90	0.85	0.94	0.90	0.87	0.98	0.89	0.83	0.93	0.89
TAS	0.92	1.00	0.92	0.92	1.00	0.92	0.91	1.00	0.91	0.91	1.00	0.91
ACT	0.89	1.00	0.89	0.89	1.00	0.89	0.88	1.00	0.88	0.88	1.00	0.88
NT	0.79	0.91	0.87	0.70	0.79	0.89	0.84	1.00	0.84	0.76	0.88	0.87
Aust.	0.87	0.97	0.90	0.85	0.94	0.90	0.88	0.98	0.89	0.85	0.95	0.89

Table 3.7 Overall school and student participation rates in Year 10

	Unweighted, including replacement schools			Unweighted, sampled schools only			Weighted, including replacement schools			Weighted, sampled schools only		
	Overall	School	Student	Overall	School	Student	Overall	School	Student	Overall	School	Student
NSW	0.81	0.97	0.83	0.75	0.90	0.83	0.82	1.00	0.82	0.76	0.93	0.82
VIC	0.76	0.94	0.82	0.71	0.87	0.81	0.73	0.96	0.76	0.69	0.91	0.76
QLD	0.79	0.95	0.83	0.79	0.95	0.83	0.78	0.98	0.80	0.78	0.98	0.80
WA	0.80	1.00	0.80	0.80	1.00	0.80	0.78	1.00	0.78	0.78	1.00	0.78
SA	0.74	0.93	0.79	0.52	0.64	0.81	0.68	0.95	0.71	0.47	0.60	0.78
TAS	0.85	1.00	0.85	0.85	1.00	0.85	0.81	1.00	0.81	0.81	1.00	0.81
ACT	0.84	1.00	0.84	0.84	1.00	0.84	0.84	1.00	0.84	0.84	1.00	0.84
NT	0.64	0.86	0.75	0.53	0.71	0.74	0.49	0.72	0.68	0.46	0.68	0.67
Aust.	0.79	0.96	0.82	0.74	0.90	0.82	0.77	0.98	0.79	0.72	0.91	0.79

Chapter 4 TEST ADMINISTRATION PROCEDURES AND DATA PREPARATION

Kate O'Malley – Australian Council of Educational Research

Frances Eveleigh – Australian Council of Educational Research

The validity and rigour of any assessment are determined by the quality of its data inputs. It is therefore essential that the collection of school- and student-level data is underpinned by well-crafted data collection procedures and quality control processes. Over the course of many years, ACER has continued to refine these processes in order to ensure that test administration is intuitive, well-designed and uniform across all participating schools and that the data collected from jurisdictions, schools and students is of high quality. This chapter outlines the data collection and test administration procedures implemented for the 2018 Main Study.

Pre-assessment preparation

As the 2018 assessment took place within schools, the contribution of both educational authorities and school staff in the organisation of, and preparation for the assessment was an essential part of the field administration. This section outlines the different stages and key roles of the NAP–SL pre-assessment preparation phase.

Contact with schools

In the lead up to the administration of the assessment, several stages of school liaison were necessary to collect school and student level information that would ensure the smooth-running of the assessment on the scheduled date. An overview of the school liaison process is outlined in table 4.1.

At each of the stages that required information to be sent from participating schools, a timeframe was provided to the relevant individual (i.e. Principal, School Contact, STSO, Test Administrator) for the provision of this information. If the school did not respond within the designated timeframe, follow-up contact was made via email and telephone. In some instances, assistance from the educational authorities was needed to ensure the information was received in a timely manner.

Table 4.1 School liaison overview

Stage	Jurisdictional Activity	ACER Project Team Activity	School Activity
1.	Educational authorities inform sampled schools of their selection in the assessment. If a sampled school is unable to participate (as confirmed by the jurisdiction) the relevant replacement school is contacted	ACER contacts principals of sampled schools to request the nomination of a school contact person and school technical support officer	Principals of contacted schools supply requested contact information via secure online form
2.		ACER contacts nominated School Contacts and requests preferred assessment dates and student lists for target year level (either Year 6 or Year 10 cohort)	School Contacts submit preferred assessment dates and student list via School Administration Website
3.		ACER contacts nominated School Technical Support Officers (STSOs) and provides Technical Readiness Test (TRT) instructions. ACER provides technical support and troubleshooting advice to STSOs via the Helpdesk	STSOs undertake the TRT to ensure the school’s computer resources are test-ready
4.		ACER notifies School Contacts of finalised assessment date and selected students via the School Administration Website	School Contact makes relevant school-level test day arrangements (including room bookings and informing sampled students of their selection)
5.	Educational authorities provide SBD for students in schools for which this information is held centrally	ACER requests Student Background Data (SBD) from School Contacts for all sampled students (where SBD cannot be provided by the jurisdiction)	School Contacts provide SBD for all sampled students via the School Administration Website
6.		ACER provides detailed test administration manual and test login credentials to all nominated Test Administrators. ACER continues to provide support to schools via the Helpdesk	Test Administrators familiarise themselves with the processes and procedures outlined in the test administration manual and consult with ACER Helpdesk staff to confirm understanding of protocol and circumvent any perceived issues prior to the scheduled assessment date.

The School Contact

Each school participating in the assessment was asked to appoint a School Contact person to liaise with the Project Team at ACER and oversee the administration of the assessment at the school. School Contacts were supplied with an instructional manual which outlined their duties and provided an overview of the assessment program. Their duties included:

- providing the Project Team at ACER with information about the school's preferred assessment dates, student cohort list at the target year level, and provision of Student Background Data for the sampled students (if this information could not be provided centrally by the jurisdiction)
- scheduling the assessment and booking the relevant rooms and/or devices for the assessment session
- notifying teachers, students and parents about the assessment, in line with their school's policies. Informational brochures for teachers and parents/carers were provided to schools for this purpose
- nominating the Test Administrator, who would be tasked with conducting the assessment session with the sampled students on the scheduled test day. In most schools, School Contacts would nominate themselves for this task.

The School Technical Support Officer

A School Technical Support Officer (STSO) was also nominated by the principal at each participating school. STSOs were issued with a short instructional handbook which provided a step-by-step guide to ensuring the school's devices were test ready. These individuals were responsible for:

- determining, in consultation with the School Contact, which devices were to be used for the assessment
- ensuring that all devices met the minimum requirements needed to access the online assessment platform by performing the Technical Readiness Test (TRT)
- ensuring all devices were switched on, logged in and fully charged (if connection to a power source was not possible) on the day of assessment.

The Test Administrator

The School Contact at each participating school was asked to nominate a Test Administrator (TA) who was tasked with administering the test to the sampled students on assessment day. In most cases, School Contacts nominated themselves for this task. TAs were supplied with a handbook which outlined their role and provided instructions for leading students through an assessment session. Specifically, the TA was required to:

- familiarise themselves with all test administration materials, including the test instructions and TA 'script'
- download the TA and student test logins and distribute to the appropriate individuals
- administer the test session as per the TA instructions and invigilate the assessment
- record student attendance on the School Administration Website.

Provision of Student Background Data

As per NAP protocol, student background data were collected for all participating students and matched to students' assessment and survey results for analysis and reporting purposes. The data variables collected for participating students are set out in the Data Standards Manual (ACARA, 2017) and included:

- gender
- date of birth
- Indigenous status
- geolocation of the students' school
- parents' school education
- parents' non-school education
- parents' occupation group
- students' and parents' home language.

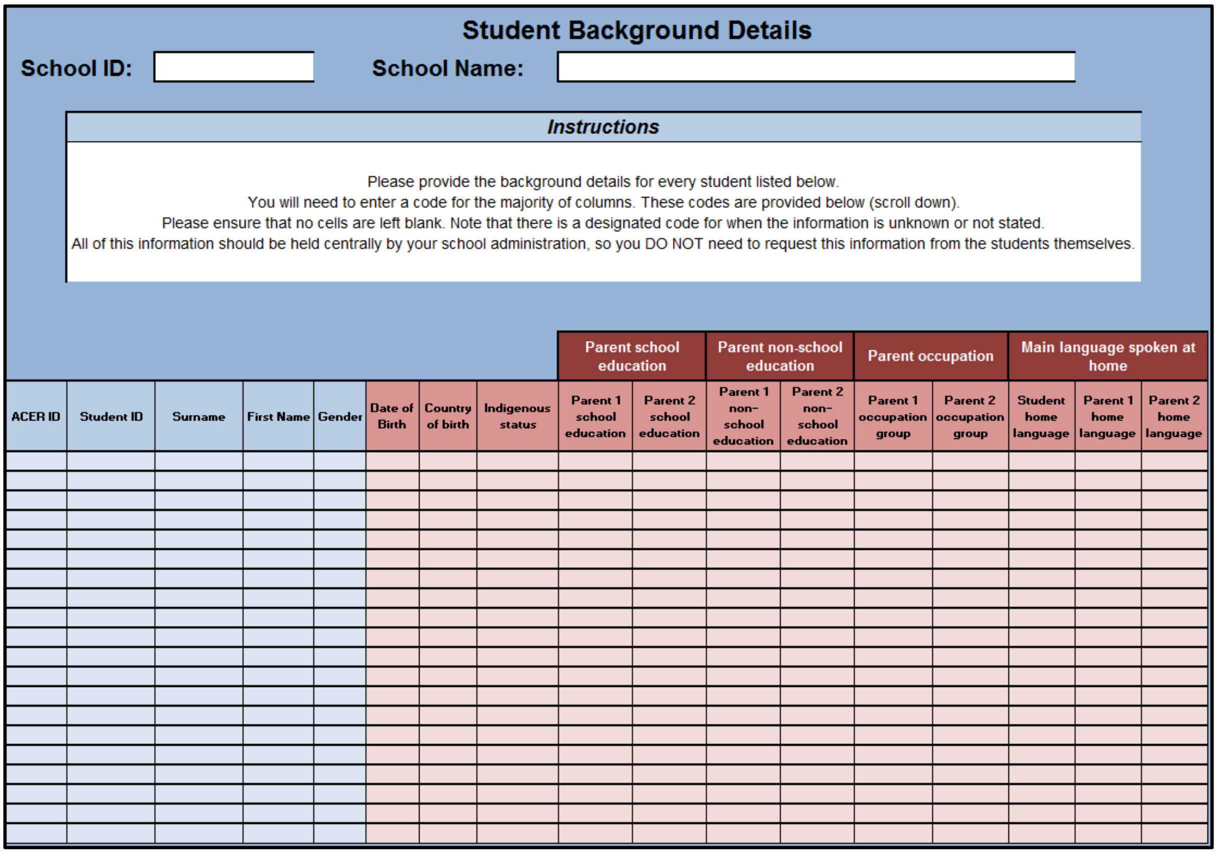
Schools are required to collect this information from the time of student enrolment and the data are often held centrally by a school's educational authority. Where data were held centrally, ACER sought the student background data from the relevant educational authority so that schools were not unnecessarily burdened with this administrative task. This occurred in half (12 out of 24) of the jurisdictions across the country. The source of student background data for the 2018 NAP–SL Main Study is outlined in table 4.2.

Table 4.2 Student Background Data Provision

Jurisdiction	Sector	Source	Jurisdiction	Sector	Source
ACT	Government	ACT DET	**SA**	Government	SA DECD
	Catholic	ACT DET		Catholic	SA CEO
	Independent	ACT DET		Independent	School
NSW	Government	NSW DET	**Tas**	Government	Tas DoE
	Catholic	School		Catholic	Tas CEO
	Independent	School		Independent	School
NT	Government	NT DET	**Vic**	Government	VIC DET
	Catholic	School		Catholic	School
	Independent	School		Independent	School
Qld	Government	QLD DETE	**WA**	Government	WA DET
	Catholic	School		Catholic	School
	Independent	School		Independent	School

Where central data collection was not possible, ACER collected this information from the schools themselves. To do this, the ACER Project Team created a spreadsheet template into which schools could enter the coded background details for each sampled student. This template was then uploaded by each school onto the secure NAP–SL School Administration Website. An example of the Student Background Data template (figure 4.1), and the accompanying code list (figure 4.2) is presented below.

Figure 4.1 Student Background Data template



Student Background Details

School ID: ______ **School Name:** ______

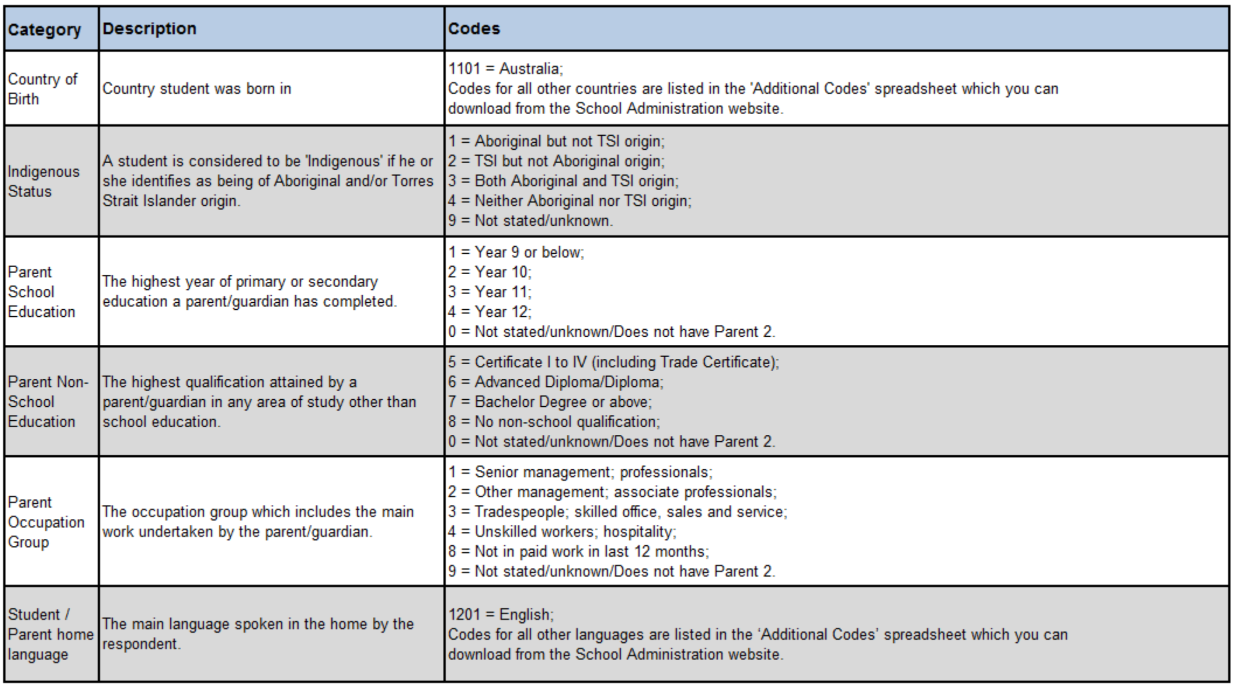
Instructions

Please provide the background details for every student listed below.
You will need to enter a code for the majority of columns. These codes are provided below (scroll down).
Please ensure that no cells are left blank. Note that there is a designated code for when the information is unknown or not stated.
All of this information should be held centrally by your school administration, so you DO NOT need to request this information from the students themselves.

								Parent school education		Parent non-school education		Parent occupation		Main language spoken at home		
ACER ID	Student ID	Surname	First Name	Gender	Date of Birth	Country of birth	Indigenous status	Parent 1 school education	Parent 2 school education	Parent 1 non-school education	Parent 2 non-school education	Parent 1 occupation group	Parent 2 occupation group	Student home language	Parent 1 home language	Parent 2 home language

The ability of the ACER Project Team to collect student background data to the level required for data analysis purposes depends on how complete the records are kept at participating schools and central authorities. Where data variables were labelled as unknown or left blank by the school or jurisdiction, and the absence of data was confirmed upon follow up from the project team, these values were coded as missing. The percentage of missing values for the derived background data variables, along with the percentages for all valid codes, are presented in the national report.

Figure 4.2 Student Background Data codes



Category	Description	Codes
Country of Birth	Country student was born in	1101 = Australia; Codes for all other countries are listed in the 'Additional Codes' spreadsheet which you can download from the School Administration website.

Indigenous Status	A student is considered to be 'Indigenous' if he or she identifies as being of Aboriginal and/or Torres Strait Islander origin.	1 = Aboriginal but not TSI origin;
		2 = TSI but not Aboriginal origin;
		3 = Both Aboriginal and TSI origin;
		4 = Neither Aboriginal nor TSI origin;
		9 = Not stated/unknown.

Parent School Education	The highest year of primary or secondary education a parent/guardian has completed.	1 = Year 9 or below;
		2 = Year 10;
		3 = Year 11;
		4 = Year 12;
		0 = Not stated/unknown/Does not have Parent 2.

Parent Non-School Education	The highest qualification attained by a parent/guardian in any area of study other than school education.	5 = Certificate I to IV (including Trade Certificate);
		6 = Advanced Diploma/Diploma;
		7 = Bachelor Degree or above;
		8 = No non-school qualification;
		0 = Not stated/unknown/Does not have Parent 2.

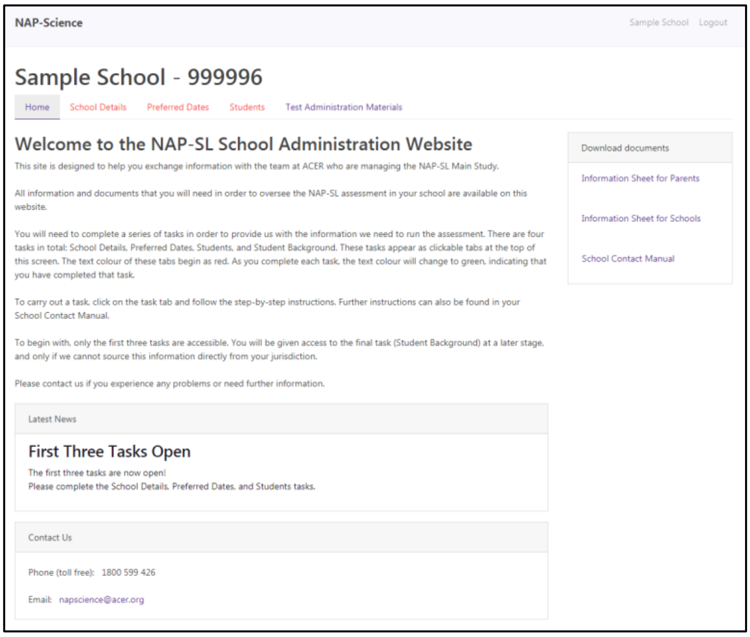
Parent Occupation Group	The occupation group which includes the main work undertaken by the parent/guardian.	1 = Senior management; professionals;
		2 = Other management; associate professionals;
		3 = Tradespeople; skilled office, sales and service;
		4 = Unskilled workers; hospitality;
		8 = Not in paid work in last 12 months;
		9 = Not stated/unknown/Does not have Parent 2.

Student / Parent home language	The main language spoken in the home by the respondent.	1201 = English; Codes for all other languages are listed in the 'Additional Codes' spreadsheet which you can download from the School Administration website.

School administration website

The NAP-SL School Administration Website was created to facilitate the exchange of information between participating schools and the ACER project team. It aimed to ease the administrative burden on School Contacts by providing a convenient, intuitive and secure repository for all school data relating to the assessment. To access the website, School Contacts needed to create a secure password and activate their school-specific account. Once their account was activated, they were able to download all relevant administrative materials from this site, as well as using it to provide information to ACER regarding school contact details, assessment date preferences, and student-related information as required. Figure 4.3 shows a screenshot from the homepage of the website.

Figure 4.3 Figure 4.3 School Administration Website homepage



Sample School - 999996

Welcome to the NAP-SL School Administration Website

This site is designed to help you exchange information with the team at ACER who are managing the NAP-SL Main Study.

All information and documents that you will need in order to oversee the NAP-SL assessment in your school are available on this website.

You will need to complete a series of tasks in order to provide us with the information we need to run the assessment. There are four tasks in total: School Details, Preferred Dates, Students, and Student Background. These tasks appear as clickable tabs at the top of this screen. The text colour of these tabs begin as red. As you complete each task, the text colour will change to green, indicating that you have completed that task.

To carry out a task, click on the task tab and follow the step-by-step instructions. Further instructions can also be found in your School Contact Manual.

To begin with, only the first three tasks are accessible. You will be given access to the final task (Student Background) at a later stage, and only if we cannot source this information directly from your jurisdiction.

Please contact us if you experience any problems or need further information.

Latest News

First Three Tasks Open

The first three tasks are now open!
Please complete the School Details, Preferred Dates, and Students tasks.

Contact Us

Phone (toll free): 1800 599 426

Email: napscience@acer.org

Increased data security provisions

Given the sensitive nature of much of the data uploaded by schools to the School Administration Website, heavy investments were made to security upgrades to this data repository. Improvements to the website's user password protocol and file upload mechanisms were introduced to ensure the secure handling of these information assets. These upgrades, together with rigorous, company-wide data handling policies and procedures helped to ensure compliance with nationally and internationally recognised standards, including:

- ISO 27002:2015 Information technology - Security techniques - Code of practice for information security controls
- The Australian Government Information Security Manual (ISM) produced by the Australian Signals Directorate, and
- The Australian Government Protective Security Policy Framework

Technical Readiness Test (TRT)

The Technical Readiness Test (TRT) was a series of tasks which aimed to ensure the compatibility of every participating school's IT resources with the NAP-Science Literacy assessment platform. To promote the smooth running of the assessment at a participating school, the TRT was performed by the nominated School Technical Support Officer in the weeks leading up to the scheduled assessment. The TRT comprised the following tasks:

1. **Identify appropriate devices to be used for the assessment**. STSOs were asked to clearly identify the 20 (or up to 20) devices that would be used to undertake the assessment on the scheduled day. These could be students' own devices (if the school operated a BYOD policy) or could be a bank of devices supplied by the school (in a computer lab setup, for instance). STSOs were asked to ensure the School Contact and all staff and students involved were aware of which devices were to be used on test day.

2. **Run a bandwidth test on the identified devices.** A minimum of 2Mbps download speed and 100 Kbps upload speed was required to run the assessment without issue. Any schools not reaching this minimum threshold were flagged for further troubleshooting or special test day arrangements (e.g. staggering the test administration in order to minimise bandwidth load).

3. **Download and install the Locked Down Browser (LDB) on the identified devices.** The LDB was designed to prevent students from accessing other applications and websites whilst undertaking the assessment. It also disabled features such as the camera, spell check and operating system commands, thereby ensuring all students had as uniform and standardised testing experience as possible.

4. **Perform a device check on all identified devices.** STSOs were asked to perform a short device check on each of the devices scheduled for use on the day of the assessment. The device check was to be performed via the LDB which ensured that the LDB was configured correctly, that there were no firewall or filtering issues at play and that the device met all required minimum specifications for a student's optimal test experience.

5. **Confirm device test readiness by completing a short IT questionnaire.** STSOs were asked to complete a short online questionnaire in order to confirm that they had completed

the TRT and to add any special arrangements relating to the IT setup at the school (that the School Contact would be providing headphones for students at a later time, for instance).

Any schools that did not complete the TRT within the allocated timeframe were contacted by the ACER Project Team and technical assistance was provided to them, if required.

Helpdesk provision and online support

An 1800 helpdesk support number and a dedicated email address were made available to schools for the entire Main Study administration phase (June – December 2018). Using these channels, the ACER Project Team supported schools through all administrative, technical and operational tasks related to the administration of the NAP-Science Literacy assessment. Project staff were also on hand to provide any urgent assistance required during, or immediately preceding, the assessment session itself. The helpdesk hours of operation during the assessment window were 8am-6pm AEST so that school hours across Australia's various time zones could be accommodated.

For complex technical matters concerning the assessment platform, issues were escalated to ACARA's Technology Partner for prompt troubleshooting assistance.

Assessment administration

The NAP–SL 2018 assessment was conducted within a two-week window at the beginning of Term 4 at each of the participating schools. The window commenced from Week 2 of Term 4 in each state and territory, as below:

QLD, Vic & WA:	Monday 15 October – Friday 26 October
ACT, NSW, NT, SA & TAS:	Monday 22 October – Friday 2 November

Schools generally undertook the test session on one day within the testing window, though a small number nominated to run the test with smaller groups of students over several days for logistical or technical reasons.

Furthermore, if attendance on the scheduled day of assessment fell below 80 per cent, schools were asked to schedule a follow-up session later within the testing window with as many of the absent students as possible. To maximise participation for the follow-up sessions, an additional testing week was added to the original assessment window for schools in all states and territories.

Data capture

The 2018 cycle of the NAP–SL assessment was delivered exclusively via the Online National Assessment Platform which has been developed to deliver NAPLAN and other NAP assessment events. The platform is managed by Educational Services Australia.

All student cognitive and survey data were captured via this online method and students used school or student supplied devices which were connected to the internet. Given the widespread compatibility of schools' IT systems with the online platform, offline delivery methods such as USB or school-server solutions were not used to administer the assessment in 2018.

All student survey and achievement data for NAP–SL 2018 were collected electronically which meant that no scanning or manual data entry of student responses was required.

Student test experience

The NAP–SL assessment comprised a single test session of 60 minutes for Year 6 students, and 75 minutes for Year 10 students. The entire assessment administration time was no more than two hours in total. This two-hour period included time for settling the students into the test room, logging students into the assessment platform, reading the instructions to the students, conducting a short student survey and administering the test itself.

Follow-up test sessions

In schools where a significant number (i.e. more than 20 percent) of students was absent on the scheduled assessment day, Test Administrators were asked to administer the scheduled session as normal with the students in attendance, and then conduct a follow-up test session at another time with as many of the absent students as possible. This ensured a participation rate of at least 80 percent in most schools administering the NAP-SL assessment.

Quality monitor visits

In order to document the quality and uniformity of the administrative procedures undertaken, a random selection of five per cent of schools across all sectors and jurisdictions were visited by National Quality Monitors on the scheduled day of the assessment. Selected schools were notified of the Quality Monitor’s visit before the scheduled assessment day so that appropriate permissions could be obtained for the Quality Monitor’s admission to the school.

National Quality Monitors were trained by the ACER project team in all aspects of test administration procedures and NAP-protocol prior to their deployment in schools. Their responsibility was to observe and record whether tasks in the procedural manuals were followed during the assessment session and to report their findings to the ACER project team via the completion of a structured online Quality Monitor Report.

In total, 28 schools from both year levels and a range of jurisdictions across Australia were visited by Quality Monitors. The Quality Monitor report template is provided in appendix 2.

Post-assessment procedures

To facilitate the requisite data analysis for, firstly, the production of school reports and then the final national report, all student responses to assessment items had to be scored appropriately. Student responses were scored either automatically by the assessment system or, where extended text responses were elicited, by groups of trained markers in a central marking location.

The following sections detail the various marking processes and quality control measures implemented during the marking operation.

Automated marking

Items that did not elicit open-ended responses from participating students were automatically scored as correct or incorrect by the assessment platform. These item types were either multiple choice, or what was termed 'non-multiple choice', where items involved drag and drop, dropdown menu, sequencing and hotspot functionalities.

As a quality control measure, students' raw responses for these items were also extracted from the system and compared to the item key in the codebook to ensure there were no anomalies with the automated scoring algorithm. Analysis of raw responses for these items was also undertaken for the later stages of data analysis.

Marking of extended text responses

The marking of extended text responses for this assessment took place at the ACER marking centre in Sydney directly after the test administration period in November 2018.

ACER employed several markers to score the NAP–SL student responses. Markers were organised into groups, with each group overseen by an experienced Group Leader who reported to ACARA's Chief Marker. Groups of markers were trained by the Chief Marker on one item at a time and then scored all student responses for this question before being trained in the next item. This train-mark, train-mark model meant that a given group of markers was focused on a single item at any one time, making it easier to recall scoring criteria and enabling markers to rapidly score a large set of data.

Regarding quality control processes, control scripts were set for each of the marked items. These control scripts were pre-selected and given a 'true score' by the Chief Marker. Markers provided scores for these scripts which were then in turn compared with the true score. If a marker gave a score that was inconsistent with the score given by the Chief Marker, the scoring criteria were clarified.

In addition to the use of control scripts, spot checking was instituted as a quality control measure throughout the marking operation. For each marked item, approximately 10 per cent of responses were spot checked (i.e. marked again) by the designated Group Leaders or the Chief Marker. The spot-checking process provided an opportunity to identify when items were being marked inconsistently, either by the whole group or an individual marker. If inconsistent marking was identified, the markers were retrained on the specific item and the responses were re-marked. This in turn improved the quality of the data used in school and public reports.

To ensure the consistent application of marking rubrics between the 2015 and 2018 NAP–SL cycles, a reliability check was undertaken on the items common to both assessments. A sample of the paper scripts from 2015 were provided to ACER and from these a random sample was taken. The 2018 markers 'remarked' these scripts and the 2015 scores and 2018 scores were subsequently data-entered. The scores were compared, and, in all cases, discrepancies were checked by the Chief Marker, with the overall discrepancy below 4 per cent.

Data processing for school reporting

Once all student responses were marked, the following data processing steps were implemented in preparation to produce school summary reports:

- Collation of all marked student data and creation of a single data file for each year level.
- Cleaning of the student response data file, including removal of introductory practice items for each student and separation of student survey data (which was not included in the

analysis for school summary reports).

- Inspection of student response data file and comparison with codebook to ensure no major data anomalies.
- Recoding of student response data to ensure missing responses were differentiated from incorrect responses[8].
- Identification of embedded missing and not reached missing responses. Not reached missing responses were excluded from item per cent correct analysis.
- Calculation and application of preliminary student weights for item per cent correct analysis.
- Calculation of item per cent correct for each NAP-SL item in standard NAP-Sample format (e.g. 75,23 where 0,1,2 item becomes 75 (facility of 1 and 2), 23 (facility of 2 only)).
- Formatting of data file to required specifications for import into the ACER Online Assessment and Reporting System (OARS).

School summary reports

After all student test data underwent the data processing steps, the final data set was imported into ACER OARS to create and distribute the online summary reports to participating schools.

The NAP–SL 2018 School Summary Reports provided schools with information about the specific items each student was administered, the level of credit each student received for every item they were administered, and the weighted proportion of students who received different levels of credit for each item. The reports were interactive in that users could filter and sort data to view information grouped by categories of interest, such as by student gender or item type. Furthermore, the reports were password-protected so that only the designated School Contact person could access them on the OARS platform and could then disseminate to other staff and/or students in line with their school's specific policy in this regard.

Whilst preliminary student weights were applied for the per cent correct analysis, scaled scores were not provided in the school reports. This was because there was not enough time to complete the equating and scaling analysis between the end of the marking process (mid-November) and the end of the school year (early mid-December). Provision of weighted, unscaled scores to schools is in line with school reporting protocol for other NAP–Sample assessments (NAP–CC and NAP–ICTL) due to the rapid turnaround of reports for participating schools.

Appendix 3 provides the instructional guide that was sent to School Contacts at participating schools which outlines how to access and read the NAP-SL school summary reports.

[8] Note: differentiation between missing and deleted responses for extended text response items was not possible due to the absence of a deleted text identifier.

Chapter 5 DATA MANAGEMENT

Kate O’Malley – Australian Council for Education Research

A robust and thorough data management strategy is integral to the accuracy and integrity of the data derived from a large-scale assessment such as NAP-SL. This chapter outlines the various data management practices and processes utilised for the conduct of the NAP-Science Literacy Main Study in 2018, including systems of identification and tracking, data capture and data verification and cleaning.

Data Security

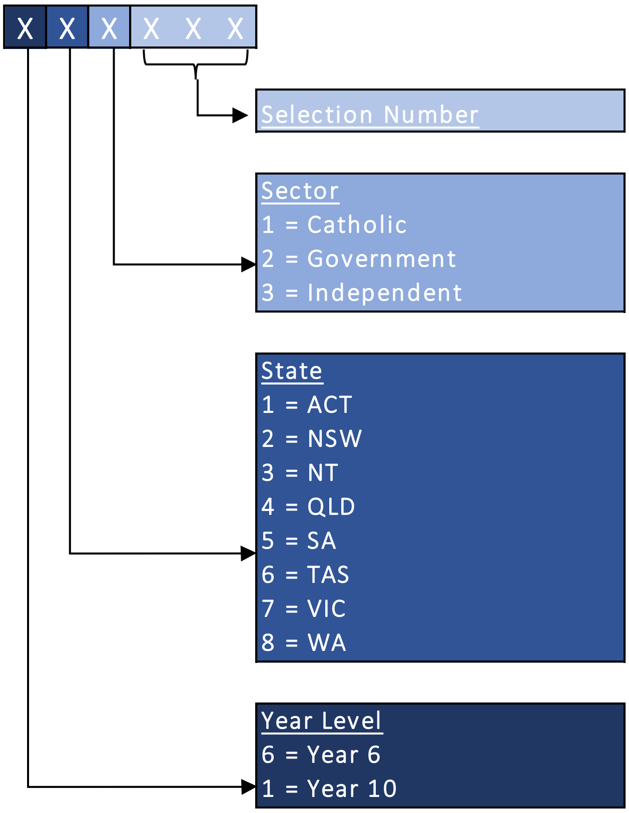
In the context of collecting, transferring, storing and disposing of school- and student-level data, it is important to ensure that all systems, staff and processes are handling those information assets securely for the life of the project. Considering this, the team at ACER ensured that all policies and procedures implemented in the conduct of NAP–Science Literacy 2018 complied with the following three standards:

- ISO 27002:2015 Information technology - Security techniques - Code of practice for information security controls
- The Australian Government Information Security Manual (ISM) produced by the Australian Signals Directorate, and
- The Australian Government Protective Security Policy Framework

Data Identification

A system of identification (ID) codes was used to track and monitor data throughout the life of the project. At the school level, a unique ID was created for each school at the time the sample was drawn. This school ID consisted of a six-digit concatenation of codes relating to year level, state, sector as well as a unique selection number sourced from the school sample file. The specific codes used for each variable are outlined in Figure 1 below.

Figure 5.1: School ID creation



For the purposes of student identification and tracking, a student ID was created that comprised the 6-digit school ID followed by a two-digit student number (01–20) that was unique to each sampled student within the school. This student ID was included in the student cognitive, contextual and student background data files so that data could be accurately matched and tracked throughout the data capture, cleaning and analysis stages.

Sampling data

The sampling data file was produced by the sampling contractor and comprised a list of all sampled schools together with their respective substitute schools. Information provided about each school included the ACARA ASL ID, state, sector, geo-location, and the Socio-Economic Indexes for Areas [SEIFA]), NAPLAN Performances Quintile, as well as the expected enrolment numbers for the grade level being assessed (either Year 6 or Year 10).

The participation status of each school was updated as needed by ACER during the conduct of the project. Post-assessment, this information was required for computing the school sample weights needed to provide accurate population estimates.

School and student data

School-level data were derived from both the sampling data file and the details provided directly to ACER by each of the participating schools. These data included contact details for the school contact person, principal and STSO, as well as other information obtained from the school via the NAP–Science Literacy School Administration Website. This information included data about the school’s computer resources, preferred assessment dates and the list of sampled students from each school.

Regarding the data ultimately used for the analysis undertaken for public reporting, data was sourced from:

- the cognitive assessment data and student survey data
- the student background data provided by the education authorities in each jurisdiction (directly, where possible) or the schools themselves
- student participation data obtained from the attendance form on the School Administration Website
- school-level variables obtained from the sample database.

In addition to these variables, student weights and replicate weights were computed for the purposes of analysis.

Data capture

Student cognitive and survey data were captured via the Online National Assessment Platform program using the Locked Down Browser installed on school or student computers. As outlined in the Data Collection Procedures chapter, the widespread compatibility of schools' IT systems with the online platform meant that offline delivery methods such as USB or school-server solutions were not used to administer the assessment in 2018.

As all the student survey and achievement data were collected electronically, scanning and manual data entry of student responses were not required.

Regarding the collection of student background data, this information was collected electronically, either from the jurisdiction via ACARA's secure ftp site, or from individual schools via the School Administration Website. Table 1 below provides the definition of each of the variables collected in this dataset.

Table 5.1: Variable definitions for student background data

Category	Description	Codes
Sex	Sex of student	1 = female 2 = male
Date of birth	Date of birth of student	Free response dd/mm/yyyy
Country of birth	Country student was born in	1101 = Australia (Codes for all other countries as per Standard Australian Classification of Countries [SACC] Coding Index 2nd edn)
Indigenous status	A student is Indigenous if he or she identifies as being of Aboriginal and/or Torres Strait Islander origin.	1 = Aboriginal but not TSI origin 2 = TSI but not Aboriginal origin 3 = Both Aboriginal and TSI origin 4 = Neither Aboriginal nor TSI origin 9 = Not stated/unknown
Parent school education	The highest year of primary or secondary education each parent/guardian has completed	1 = Year 9 or equivalent or below 2 = Year 10 3 = Year 11 4 = Year 12 0 = Not stated/unknown/Does not have Parent

Category	Description	Codes
		1 or 2
Parent non-school education	The highest qualification attained by each parent/guardian in any area of study other than school education	5 = Certificate I to IV (including Trade Certificate) 6 = Advanced Diploma/Diploma 7 = Bachelor’s Degree or above 8 = No non-school qualification 0 = Not stated/unknown/Does not have Parent 1 or 2
Parent occupation group	The occupation group, which includes the main work undertaken by each parent/guardian	1 = Senior management; professionals 2 = Other management; associate professionals 3 = Tradespeople; skilled office, sales and service 4 = Unskilled workers; hospitality 8 = Not in paid work in last 12 months 9 = Not stated/unknown/Does not have Parent 1 or 2
Student/Parent home language	The main language spoken in the home by the respondent	1201 = English (Codes for all other languages as per the Australian Standard Classification of Languages [ASCL] Coding Index 2nd edn)

Data cleaning and verification

Data cleaning and verification relate to processes of ensuring that all data received from various sources are free from error. For NAP-SL, a series of data cleaning steps was undertaken on all data collected from jurisdictions, schools and students. With respect to student background data, the following steps were performed:

1. Student names (for the purposes of school reporting) were corrected where there was obvious first name/surname reversal, or where foreign characters (e.g. ?, !, %) were included. Some instances of correction had to be confirmed with the school directly.
2. Missing sex of the student was attributed where it could be inferred from the school type (e.g. where single-sex) or name of the student. Some instances of correction had to be confirmed with the school directly.
3. All dates of birth were converted to the standard dd/mm/yyyy format, and any auto-formatting executed by the spreadsheet template that rendered dates of birth illegible was reversed and corrected.
4. Any free text or abbreviated text was coded as per the variable coding schema above.
5. Any out of range, implausible or missing values were double-checked with the school or jurisdiction that provided the data. Where possible, the correct values were inputted. Where no further information was provided or available, the data were recoded to missing.

With respect to the student cognitive and survey data, the following preliminary data cleaning steps were performed:

1. Instances of invalid IDs were investigated and, after liaison with the test administration team, corrected where possible or else removed from the dataset.
2. Instances of spare IDs were matched with valid Student IDs and recoded accordingly. This often necessitated confirmation and cross-checking with the attendance roll data and notes from the test administration team.
3. Patterns of missing values were explored and, where appropriate, recoded to ‘9’ for embedded missing, ‘R’ for unreached missing or ‘N’ for not administered.

Chapter 6 SCALING PROCEDURES and EQUATING

Prof David Andrich – University of Western Australia

Ida Marais – University of Western Australia

Sonia Sappl – University of Western Australia

Eveline Gebhardt - ACARA

Both cognitive and survey items were scaled using item response theory (IRT) scaling methodology. The cognitive items were used to derive a one-dimensional NAP–Science Literacy achievement scale, while several scales were constructed based on different sets of survey items.

The scaling model

Test items were scaled with the one-parameter model (Rasch, 1960). In the case of dichotomous items, the model predicts the probability of selecting a correct response (value of one) instead of an incorrect response (value of zero), and is modelled as:

$$P_i(\theta_n) = \frac{exp(\theta_n - \delta_i)}{1 + exp(\theta_n - \delta_i)}$$

where $P_i(\theta_n)$ is the probability of person *n* scoring 1 on item *i*, θ_n is the estimated ability of person *n,* and δ_i is the estimated location of item *i* on this dimension. For each item, item responses are modelled as a function of the latent trait θ_n.

For items with more than two (*k*) categories (as for example with Likert-type items) the polytomous Rasch model (partial credit parameterization) was applied which takes the form of:

$$P_{x_i}(\theta_n) = \frac{exp \sum_{k=0}^{x}(\theta_n - \delta_i + \tau_{ik})}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h}(\theta_n - \delta_i + \tau_{ik})} \qquad x_i = 0,1, \dots, m_i$$

where $P_{xi}(\theta_n)$ denotes the probability of person *n* scoring *x* on item *i*, θ_n denotes the person's ability, the item parameter δ_i gives the location of the item on the latent continuum, and τ_{ij} denotes the threshold *k* between adjacent categories.

The analysis of item characteristics and the estimation of model parameters were carried out with the ACER ConQuest (Adams, Wu & Wilson, 2015) and RUMM2030 (Andrich, Sheridan & Luo, 2018) software packages.

SCALING COGNITIVE ITEMS

This section outlines the procedures for analysing and scaling the cognitive test items measuring science literacy. The procedures are somewhat different from scaling the survey items, which will be discussed in a subsequent section.

Calibration sample

In NAP Sample, jurisdictions contribute equally to the estimation of item difficulties. Choosing a sample with equal numbers in each jurisdiction and each year level was problematic because of the low number of Year 10 students in some jurisdictions. Therefore, and because the main comparison is that of Year 6 with previous years and it is important to maintain the Year 6 scale, a calibration sample of only Year 6 students was chosen. An alternative method would be to use all available data and weigh the cases so that the sum of the weights is equal across jurisdictions (senate weights). However, this is not possible in all software packages.

A random subset of the Year 6 responses was chosen to ensure that each jurisdiction had an equal representation in the sample. Since the ACT had the smallest number of responses, all 399 responses were included in the calibration sample. For each of the other jurisdictions, a random sample of 399 responses was selected. Consequently, the calibration sample consisted of 3192 (399×8) responses.

The full set of Year 10 responses were added to the calibration sample, anchored on the calibration sample item estimates, to estimate the parameters for the unique Year 10 items. This sample, with approximately equal numbers of Year 6 and 10 responses, was used for analysing year level DIF.

Item statistics

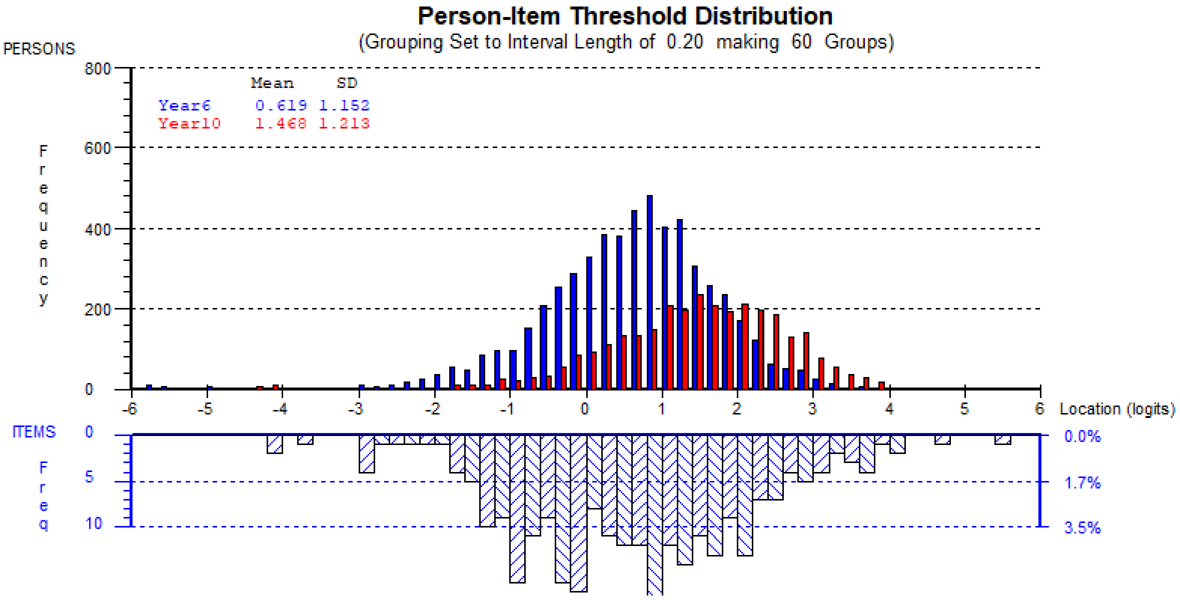
A complete list of items with their relevant statistics can be found in appendix 7. The table in the appendix shows that one item was deleted due to poor fit across the trait, showing less discrimination relative to the other items. For each item the following information is shown:

- Item code
- Item name
- Maximum score
- Whether the item was a vertical link
- Whether the item was a horizontal (historical) link
- Estimated difficulty from the 2018 free calibration
- Estimated difficulty on the historical scale (including the mode effect adjustment). This is shown for Response Probabilities (RP) of 0.5 and of 0.62 and on the SL scale (after standardisation). Response Probability refers to the probability for a student to respond correctly to an item of the same difficulty as the student's ability. The default probability in Rasch models is 0.5.
- Percentage of students with correct responses
- Weighted fit (MNSQ) statistic

Item-person maps

The responses from Years 6 and 10 were analysed both separately and simultaneously. Figure 6.1 shows an item-person map from the analysis of Years 6 and 10 simultaneously.

Figure 6.1 Item-person map of 2018 main study items



The horizontal scale in figure 6.1 shows the location of the students and items, with the student proficiency distributions shown above the x-axis and the item difficulty distribution shown below the x-axis. The items are placed in item difficulty order, where items to the right are most difficult. These are initial estimates before the horizontal equating. Likewise, the students are placed in order of proficiency, where students to the right are most proficient on the test. These are initial estimates before the addition of sampling weights and the calculation of plausible values.

Figure 6.1 shows that the items cover a wide range of difficulty levels. The match between item difficulties and person proficiencies is quite good overall for both year levels. On this scale, the average item difficulty for the Year 6 test is zero logits while the average proficiency is 0.619 logits. The average item difficulty for the Year 10 test is 0.837 logits and the average proficiency is 1.468 logits.

Figure 6.2 shows the same information as figure 6.1 but with the items labelled.

Figure 6.2 Item-person maps of 2018 main study items with the items labelled

```
                         YEAR 6                                                             YEAR 10
                            |                                                                  |
                            |                                                                  | 239.3
                            |                                                                  |
 5.0                        |                                                                  |
                            |                                                                  |
                            |                                                                  | 241.3
                            |                                                                  |
                            |                                                                  |
 4.0                        | 92.2                                                             | 216.6
                            |                                                                × | 193
                            | 138.2 136                                                     ×× | 194.3 234
                            | 79.4                                                         ××× | 230 194.4
                          × |                                                            ××××× | 209 138.2
 3.0                      × | 104                                                       ×××××× | 208 240.2 238.3
                         ×× |                                                     ×××××××××××× | 128 199 217 136 222
                         ×× | 94                                                   ××××××××××× | 216.5 221 214
                        ××× | 85 105.2 134                                    ×××××××××××××××× | 219 211 239.2 240.1
                      ××××× | 63 128                                          ×××××××××××××××× | 236.3 241.2 218 194.2 223
 2.0                ××××××× | 140.2 91 86.2 71                              ×××××××××××××××××× | 227 216.3 220 134 216.4 140.2 237.2 215 194.1 188
                  ××××××××× | 79.5 81 127 105.1                               ×××××××××××××××× | 235.2 179 127 143 236.2 196
                 ×××××××××× | 126.2 51 92.1 124 132 131                      ××××××××××××××××× | 177 168 216.2 126.2 212 237.1 124 207.2 131 200
               ×××××××××××× | 135.2 88.3 69 133 120 11 90 138.1           ×××××××××××××××××××× | 135.2 241.1 137 197 133 120
          ××××××××××××××××× | 101 113 88.2 100 79.3                      ××××××××××××××××× | 173 203 195 233 113 239.1 162 148 181 238.2
 1.0       ×××××××××××××××× | 53.2 140.1 103 139 54                      ××××××××××××××××× | 202 185 192 140.1 164 155 139 138.1 178
        ××××××××××××××××××× | 89.2 96 118 10 68 137 75 126.1 102                ×××××××××××× | 235.1 118 189 167 125 180 236.1 132 183 207.1 225 126.1 144 238.1
         ×××××××××××××××××× | 108 83.2 34 39 135.1 29 28                         ××××××××××× | 159 108 158 135.1 147 210 171
            ××××××××××××××× | 98 97 121 125 16 110 37 21 116                     ××××××××××× | 121 116 161 226 151
            ××××××××××××××× | 84 95 43 52 23 53.1                                  ××××××××× | 152 191 216.1 224 176
 0.0          ××××××××××××× | 17 86.1 89.1 66                                       ×××××××× | 169 154 170 201
                ××××××××××× | 4 36 13 129 46 48 130 117 83.1 45 50                   ××××××× | 175 172 106 23 110 129 229 130 117
                 ×××××××××× | 31 22 47 12 93 79.1 74 35 112                              ×××× | 146 198 160 165 184 153 112 182
                   ×××××××× | 59 8 7 76 57 40 19                                          ××× | 150 141
                     ×××××× | 1 56 42 65 107 2                                             ×× | 204 142 213 107 174 157
-1.0                   ×××× | 38 27 111 123 30 49 109 5 106 33 15 14                       ×× | 190 231 186 111 123 109 228
                       ×××× | 32 3 79.2 122 77 80                                          ×× | 156 122 166 149
                        ××× | 73 55 67 20 62 60 64 70 99                                    × | 206
                         ×× | 72 58 9                                                       × | 145 205
                         ×× | 41 24 119 114                                                 × | 119 114
-2.0                      × | 115                                                          | 115
                          × | 25                                                           |
                          × | 44                                                           |
                            | 26                                                           |
                            |                                                                  | 187
-3.0                        | 87 82 88.1 78                                                |
                            |                                                                  |
                            |                                                                  |
                            |                                                                  |
                            | 18                                                           |
-4.0                        |                                                                  |
                            | 61 6                                                           × |
                            |                                                                  |
                      × = 25 Persons                                                     × = 12 Persons
```

Test reliability

Person separation reliability for the 2018 NAP–SL tests is 0.89 for the Year 6 calibration sample and 0.92 for the Year 10 sample. In comparison, the reported reliability for PISA 2003 mathematics is 0.85, and 0.89 for TIMSS 2003 Grade 8 mathematics.

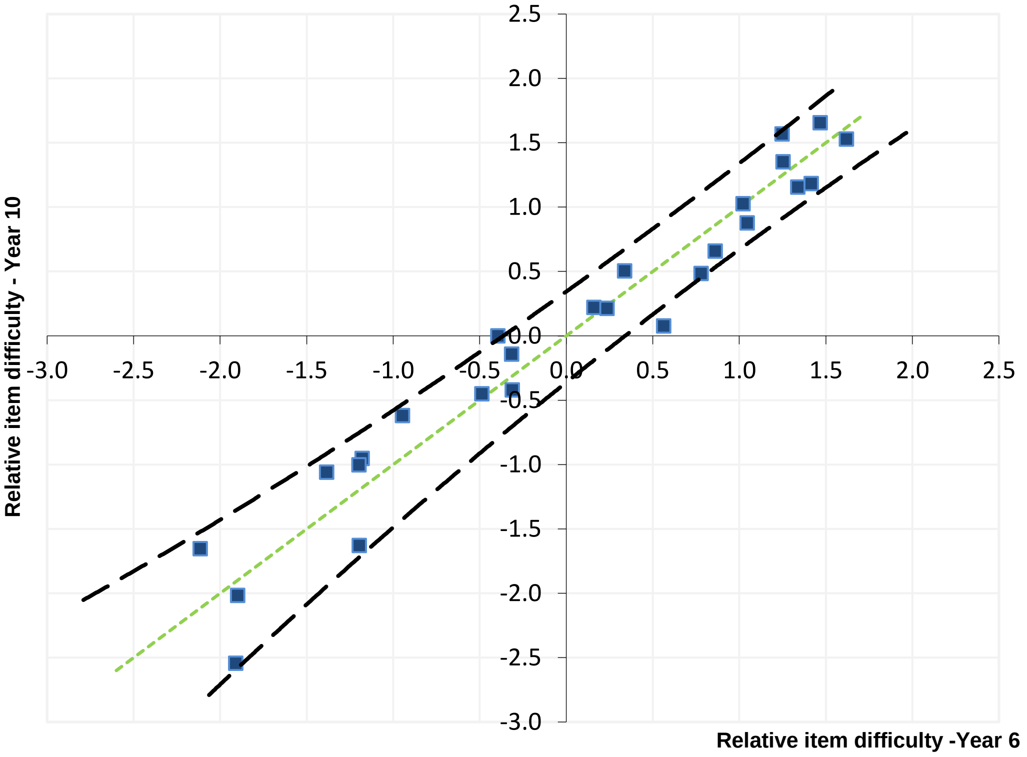
Differential Item Functioning

Item response models assumes that the probability of responding correctly to an item is only dependent on a student's ability, not on any group membership. Violation of this assumption is called differential item functioning (DIF).

DIF for grade was examined in separate analyses of the grade levels. In order to conduct comparisons between Year 6 and Year 10 vertical link item locations in the separate analyses, the locations were adjusted to be mean deviated. Items that were clear outliers were removed with a final set of 26 link items (out of 35) retained. A plot of the mean deviated Year 6 and the Year 10 item difficulties for the final set of vertical link items, including graphical representation of the 95 per cent confidence interval for the statistical difference between item locations, is given in figure 6.3.

The table of item statistics in appendix 7 indicates the nine vertical link items which were broken as links into Year 6 and Year 10 unique items. These items have an 'a' (Year 6) or 'b' (Year 10) in their labels.

Figure 6.3 Scatter plot of relative item difficulties for Year 6 and Year 10 vertical link items



Eighteen items showed DIF for gender (7 items favouring females and 11 favouring males). In order to test whether this DIF was of concern, the 18 items were deleted and the impact on the relative means was studied. The means for each gender remained very similar after deleting the 18 items, so it was concluded that this DIF cancels out in the means of boys and girls.

Test form effect

‘Test form effect’ refers to the differences in test form difficulties after equating of the forms has been carried out. That is, students may be advantaged or disadvantaged by taking a particular test form, even after forms have been equated. Table 6.2 shows the test form difficulty estimates.

Table 6.1 Test form difficulty parameters in Years 6 and 10

Form	Year 6		Year 10	
	Mean (logit)	**SD**	**Mean (logit)**	**SD**
1	0.679	1.084	1.545	1.190
2	0.645	1.089	1.334	1.260
3	0.751	1.088	1.395	1.289
4	0.574	1.081	1.497	1.204
5	0.633	1.196	1.549	1.136
6	0.721	1.079	1.451	1.174
7	0.638	1.049	1.504	1.270
8	0.500	1.071	1.465	1.165

The test form parameters shown in Table 6.2 are very similar in magnitude, indicating that test form effect was not a serious issue for the assessments. It is noted that the Year 6 test form 8 appears to be somewhat easier than the other test forms, however this form contained only historical link items necessary for horizontal equating.

In the Year 6 calibration sample, the ANOVA for test form was not significant with a Bonferroni adjustment (F(7,3184)=2.128, p=0.038). In the Year 10 sample, anchored on item estimates from the calibration sample, the ANOVA for test form was also not significant (F(7,3047)=1.422, p=0.192).

Plausible values

Plausible values methodology was used to generate estimates of students’ science literacy. Using item parameters anchored at their estimated values from the calibration process, plausible values were randomly drawn from the marginal posterior of the latent distribution (Mislevy, 1991; Mislevy & Sheehan, 1987; von Davier, Gonzalez, & Mislevy, 2009). Here, ‘not reached’ items were included as incorrect responses, just like other (embedded) missing responses. Estimations are based on the conditional item response model and the population model, which includes the regression on background variables used for conditioning (see a detailed description in Adams & Wu, 2002). The ACER ConQuest Version 4.0 software was used for drawing plausible values.

This cycle, no analyses were planned on the relationship between survey responses and science literacy. Therefore, survey responses were not included in the conditional model. It is recommended to include these in future cycles of NAP–SL so that relationships between survey responses and achievement can be reported on.

The conditioning variables included are:

- school mean proficiency (average of students' weighted likelihood estimates for each school adjusted for student's own performance)
- stratum (jurisdiction and sector)
- gender
- Indigenous status
- geographic location
- parental occupation
- parental education
- language background.

Horizontal equating

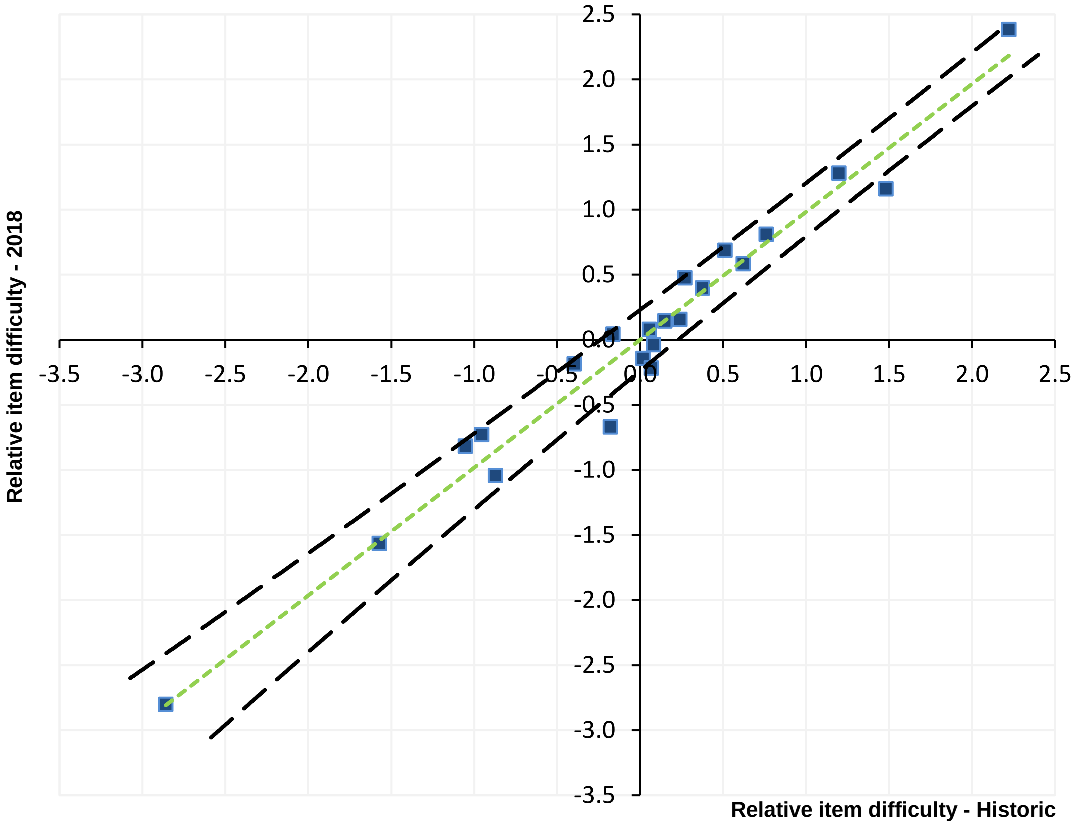
The 2018 results were equated to 2006 results. To carry out the equating, horizontal (historical) link items between the 2018, 2015, 2012, 2009 and 2006 tests were used.

Horizontal link items

In order to equate the 2018 results to the science literacy scale, a total of 37 historical items were included in the 2018 assessment. This included five link items from the 2006 assessment, two link items from the 2009 assessment, 12 items from the 2012 assessment and 18 items from the 2015 assessment. Care was taken to find items that performed well psychometrically and also covered the range of science literacy strands A, B and C and the concept areas.

The list of link items was refined based on the comparison of item locations in 2018 and the location of the items on the historical (2006) scale. First, the 2018 location of link items was independently estimated. In order to conduct comparisons of item locations between the historical scale and the 2018 scale, the locations were adjusted to be mean deviated. Items that were clear outliers were removed with a final set of 22 link items retained, which had the same mean and standard deviation in 2006 and 2018. A plot of the mean deviated historic and 2018 item difficulties for the final set of link items, including graphical representation of the 95 per cent confidence interval for the statistical difference between item locations, is given in figure 6.4.

Figure 6.4 Calibrated item difficulties in 2006 and 2018 for the final horizontal link item set



Equating procedures

The mean difference between the final set of link item parameters for 2018 compared to historic estimates was used to obtain the 'shift' required to place 2018 results onto the historic scale. This approach to equating was used for NAP–SL in 2015 because it is a simple and common approach to equating. The methodology used for equating NAP–SL to historical cycles in previous cycles provided similar results to the 'shift' method and hence, the simplest equating model was retained. In addition, the adjustment for students' interaction with online historic items was included in the transformation formula. Based on the results of the mode effect (paper or online) study, the 2018 results were shifted up by 0.131 of a logit to correct for the effect that the switch from paper-based testing to computer-based testing had on student performance.

The result of the equating process was the derivation of a transformation formula for the 2018 results to be placed on the 2006 scale:

$$\text{logit}_{2006} = \text{logit}_{2018} - 0.507 + 0.131$$

For item parameter estimates, an additional shift of 0.49 was applied to change the response probability from 0.5 to 0.62.

To place the results in logits onto the reporting scale, the following additional standardisation parameters were applied

$$\text{Reporting scale} = 100 * (\text{logit}_{2006} - \text{mean}_{2006}) / \text{SD}_{2006} + 400,$$

where the Year 6 mean of 2006 was equal to 0.201 and the standard deviation to 0.955.

Uncertainty in the link

The shift that equates the 2018 data with the 2006 data depends upon the change in difficulty of each of the individual link items. Consequently, the sample of link items that have been chosen will influence the estimated shift. This means that the resulting shift could be slightly different if an alternative set of link items had been selected. Consequently, there is an uncertainty associated with the equating, which is due to the choice of link items, like the uncertainty associated with the sampling of schools and students.

The uncertainty which results from the selection of a subset of link items is referred to as *linking or equating error*. This error should be considered when making comparisons between the results from different data collections across time. Just as with the error that is introduced through the process of sampling students, the exact magnitude of this linking error cannot be determined. We can, however, estimate the likely range of magnitudes for this error and take this error into account when interpreting results. As with sampling errors, the likely range of magnitude for the combined errors is represented as a standard error of each reported statistic.

The following approach has been used to estimate the equating error. Suppose we have a total of *L* score points in the link items in *K* modules. Use *i* to index items in a unit and *j* to index units so that $\hat{\delta}_{ij}^{y}$is the estimated difficulty of item *i* in unit *j* for year *y*, and let:

$$c_{ij} = \hat{\delta}_{ij}^{2018} - \hat{\delta}_{ij}^{2006}$$

The size (number of score points) of unit *j* is m_j so that:

$$\sum_{j=1}^{K} m_j = L \text{ and } \bar{m} = \frac{1}{K}\sum_{j=1}^{K} m_j$$

Further let:

$$c_{\bullet j} = \frac{1}{m_j}\sum_{i=1}^{m_j} c_{ij}, \text{ and } \bar{c} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{m_j} c_{ij}$$

and then the link error, taking into account the clustering is as follows:

$$LinkError_{2018,2006} = \sqrt{\frac{\sum_{j=1}^{K} m_j^2 (c_{\bullet j} - \bar{c})^2}{K(K-1)\bar{m}^2}} = \frac{\sum_{j=1}^{K} m_j^2 (c_{\bullet j} - \bar{c})^2}{L^2}\frac{K}{K-1}$$

Link errors between 2018 and each of the previous cycle are presented in Table 6.3.

Table 6.2 Example of link error application in calculating standard error of difference

Link errors between 2018 and	In logits	On reporting scale
2015	0.042	4.39
2012	0.064	6.68
2009	0.071	7.42
2018	0.079	8.28

The link error is used only when comparisons across 2018, 2015, 2012, 2009 and 2006 results are made. For example, to test whether the mean achievement in 2018 differs from the mean achievement in 2006, the link error is added to the standard error of the difference, as illustrated in table 6.3. Additional information about the use of link errors can be found in chapter 8.

Table 6.3 Example of link error application in calculating standard error of difference

Year 6	2018 mean on 2006 scale & S.E.	2006 mean & S.E.	2018 mean – 2006 mean	Standard error of difference	Standardised difference
SA	400 (7.89)	392 (5.10)	8	SQRT $(7.89^2+5.10^2+8.28^2)$	0.67 = 8/12.53 (not significant)

Scaling survey items

Besides the item-by-item analysis of cognitive responses, survey responses were also analysed using the Rasch model (RUMM2030 software) to provide a single measure for each student on a survey scale. All responses were reverse scored, so that a high score indicates a positive response to the items. Students with high levels of agreement on the statements in the items are then placed high on the scale.

Fourteen of the survey items were presented to both year groups. These items were used as ‘link items’ to place Year 6 and 10 students on the same scale.

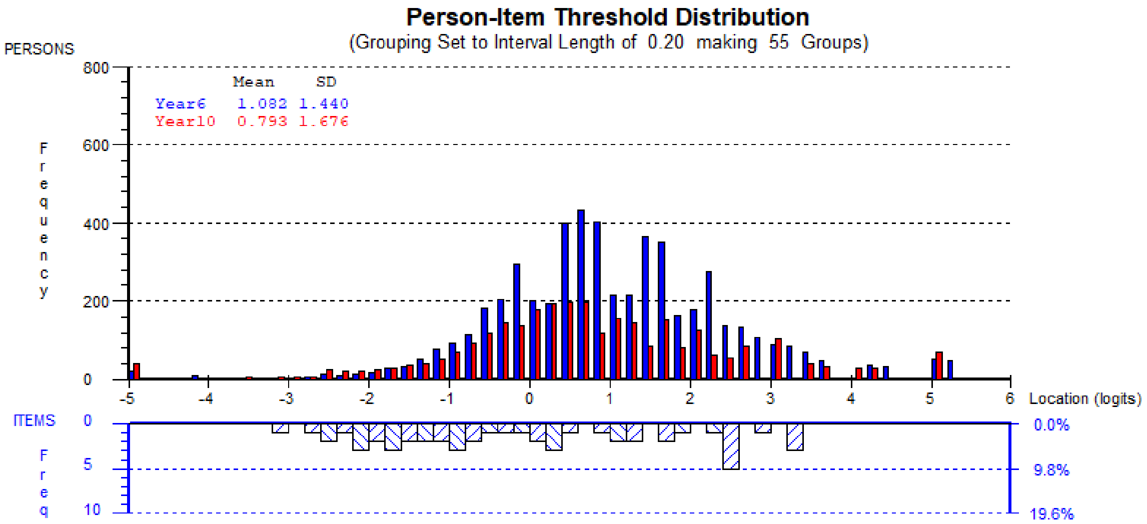
The fit statistics showed that there wasn’t an underlying unidimensional construct. It would not be valid to provide one measure for each student on *all* survey scale items. Items from *Teaching and learning science* (Groups 4 to 6) did not fit with items from *Student engagement with science* (Groups 1 and 2) and *Science as a human endeavour* (Group 3). This is not surprising as the items from Groups 4 to 6 are items about *how often and in what way science is taught* at a student’s school, whereas items from Groups 1 to 3 are about a student’s *interest, self-efficacy and perception of the value of science.*

Because of this, items from Groups 4 to 6 were deleted from the scale (12 items) leaving only items from Groups 1 to 3 (15 items). The remaining items did not show DIF for jurisdiction and gender and all had ordered thresholds, showing that the response categories functioned as intended.

Because the link items used to place Years 6 and 10 on the same scale should function invariantly for the two Grade groups, DIF for Grade was examined. Two items showed DIF for Grade. For the same level of overall attitude to science, Year 10 students scored higher on item 14 (*Science is part of my everyday life*). For the same level of overall attitude to science, Year 6 students scored higher on item 13 (*It is important that all students learn science*).

Figure 6.5 shows the person/item threshold estimates of the final scale, after deletion of items in Groups 4 to 6 and resolving items 13 and 14 into separate Grade items. The graph shows that the items are fairly well spread out in terms of ‘difficulty to endorse’ and that the items align well with the locations of the persons. Year 6 students had a higher mean (1.082), indicating a more positive attitude to science than Year 10 students (0.793). The difference between the Year 6 mean (1.082) and the Year 10 mean (0.793) was 0.289. The Year 10 students have a higher standard deviation (1.676) compared to Year 6 students (1.440), indicating a bigger variation of attitudes to science.

Figure 6.5 Histograms of the Rasch person and item threshold estimates of the final scale for Year 6 and Year 10



References:

Andrich, D., Sheridan, B., and Luo, G. (2018). RUMM2030: Rasch unidimensional models for measurement. Perth UWA, Australia: RUMM Laboratory

Chapter 7 ACHIEVEMENT LEVELS AND THE PROFICIENT STANDARD

Julian Fraillon – Australian Council for Education Research

Eveline Gebhardt - ACARA

One of the main objectives of NAP–SL is to monitor and report trends in science literacy performance. One convenient and informative way of doing so is to reference students' results to the NAP–SL proficiency levels. Typically, students whose results are located within a proficiency level can demonstrate the understandings and skills associated with that level and possess the understandings and skills of lower proficiency levels.

Prior to the 2018 cycle, the NAP–SL scale contained only one proficient standard for Year 6 students. The proficient standard is a point on the scale that represents a 'challenging but reasonable' expectation of student achievement at that year level. With the addition of Year 10 students, a standard-setting process was conducted to establish a new Year 10 proficient standard.

Setting the Year 10 proficient standard

The NAP–SL assessment provides the basis on which national key performance measures (KPMs) can be reported and a mechanism for monitoring progress towards the Melbourne Declaration on Educational Goals for Young Australians (Melbourne Declaration). The KPM for Year 6 science was established as part of the first cycle of the NAP–SL as the proportion of students performing at or above Level 3.2 on the Science Literacy scale (MCEETYA, 2004, ACARA, 2015).

Before 2018, the Programme for International Student Achievement (PISA) was the primary national measure of performance for science literacy among secondary school students. The KPM for science in Australia was the proportion of 15-year-old students performing at or above Level 3 on the OECD PISA combined scientific literacy scale (ACARA, 2015).

In July 2016, the Education Council decided to extend the NAP–SL to Year 10 students from 2018. It was therefore necessary to establish a KPM for Year 10 Science Literacy on the NAP Science Literacy scale.

This section of the report describes the process by which the standard was established.

Background to standard-setting process

There are many different standard-setting procedures, but they can be classified according to three main approaches:

a) *Holistic* methods require experts to make judgements about predicted candidate achievement on a whole test.

b) *Content-based* methods require experts to make judgements about predicted candidate achievement on the individual items within a test.

c) *Performance-based* methods use candidate performance data as the basis for establishing the cut-scores.

(Kellow and Wilson, 2008)

What all these approaches share is that they ‘invoke the judgement of experts in the content area of interest’ and that ‘these individuals possess substantive content knowledge as well as intimate familiarity with the target population’ (Kellow & Wilson, 2008).

ACARA recruited a panel of 20 science educators from across Australia to form the panel. The majority of the educators were actively working as secondary-school science teachers (and science curriculum leaders) and many had also participated in item development workshops that contributed new item material to NAP–SL 2018.

Hambleton (1998) described a generic set of steps in standard-setting exercises that are common to the process of setting performance standards, regardless of the specific approach or procedure adopted. An adapted set of these steps is presented in Table 7.2 (adapted from Cizek, 2006).

Table 7.1 Generic steps in setting performance standards

Step	Description
1	Select a large and representative panel of experts.
2	Choose a standard-setting procedure: prepare training materials and standard-setting meeting agenda.
3	Prepare descriptions of the referent candidate or group.
4	Train participants to use the standard-setting method.
5	Compile item judgements/ratings from experts and summarise outcomes to provide feedback on ratings.
6	Facilitate a discussion amongst the experts based on the feedback from the rating exercise.
7	Provide experts the chance to revise their ratings on the basis of the feedback and discussion; this may include repeating Steps 5 and 6.
8	Ask experts to review/reconfirm their ratings to determine a ‘final’ recommended standard.
9	Conduct an evaluation of the process with the participants to confirm that they are satisfied with and have confidence in the process.
10	Assemble documentation of the process and other evidence that may have a bearing on the validity of the resultant standards.

The procedure used to establish the *proficient standard* for Year 10 science reflected the generic steps presented in Table 7.2. The standard was established using a content-based approach informed by presentation of candidate performance data to the participating experts following their initial ratings of the test items. Broadly the presentation of candidate achievement data to judges took place as part of the first iteration of Step 6 in Table 7.2.

This blend of a content-based approach informed by performance data was used because the previous Year 10 KPM for science was established against the OECD PISA Scientific Literacy scale and not with reference to NAP–SL achievement. This blended approach is supported under such circumstances. For example, Linn (2003) states:

> *Assuring that judges on standard setting panels understand the context in which the standards will be used is a minimal requirement for obtaining reasonable performance standards. Normative information needs to be made part of the process for judges to anchor their absolute judgments with an understanding of current levels of performance of students and likely consequences. As Zieky (2001) has noted, considering both absolute and normative information "in setting a cut-score can help avoid the establishment of unreasonably high or low values" (p. 38).*

As the pre-existing KPM for Year 10 in NAP–SL had been the boundary between Level 2 and Level 3 on the OECD PISA Scientific Literacy Scale, a final validation check (as part of what is listed as step 10 in table 7.2) was included as part of the process. This involved judges first comparing the nature of the items in the range they had recommended with level descriptors and release items from OECD PISA. Furthermore, the judges compared both the NAP–SL outcomes and the OECD PISA Science Literacy scale levels to the Australian Curriculum: Science achievement standards.

Standard-setting procedures

There remains no agreed best method for setting standards (see, for example Zieky, 2001 or Linn, 2003). There is however agreement that, wherever possible, it is wise to consider both the outcomes of more than one method and any additional relevant statistical information when establishing benchmark cut scores (Linn, 2003, Jaeger, 1989).

The standard-setting procedure for NAP–SL comprised firstly a modified Angoff procedure, the *Yes/No* method (see, for example Cizek, 2006 or Kellow and Wilson, 2008). After this, the judges were presented with the actual candidate facility data (the percentage of candidates correctly answering an item in NAP–SL 2018) for each item before finally completing an adapted form of the Bookmark method procedure (see, for example, Mitzel et al, 2001, or Kellow & Wilson, 2008). This combination of a modified Angoff and adapted Bookmark methods informed by candidate achievement data is similar to that used in establishing the *Proficient Standards* in the Australian National Sample assessments of ICT Literacy and Civics and Citizenship (Wernert et. al., 2006, Ainley et. al, 2008).

Steps in the standard-setting procedure

Following is a description of the steps used to establish the recommended *proficient standard* (or acceptable range for the standard).

Step 1: Selecting the panel of expert judges

As described previously ACARA recruited an panel of 20 experienced secondary school science teachers (and curriculum leaders) from In addition to the professional experiences and attributes that led to their inclusion on the panel, the experience of participating in panel meetings further qualified the members for the standard-setting work The panel members were well acquainted with: the genesis and full context of the work, the candidature; the Content Parameters – Advice for Test Developers document that underpinned the test development work, and the test items themselves.

Step 2: Convening the standard-setting meeting

The standard-setting meeting took place over two consecutive days, March 27 and 28, 2019 at the ACARA Sydney office. All judges on the panel were required to attend both full days of the meeting.

Hambleton (2001) listed nine characteristics of effective standard-setting panellist training:

1. explaining and modelling the steps to follow in setting standards (e.g. estimating the performance of borderline candidates or sorting examinee papers into ordered categories)
2. showing the scoring keys and/or scoring rubrics and ensuring they are understood
3. completing easy to use rating forms
4. providing practice in providing ratings
5. explaining any normative data that will be used in the process, and so on
6. familiarising panellists with assessment content (e.g., the assessment tasks)
7. developing borderline descriptions (if used)
8. taking the test under standard or near standard conditions
9. reviewing the item pool on which the performance items will be set

(Hambleton, 2001)

The selection of the judges meant that they were already extremely familiar with the NAP–SL test and the role of NAP–SL in Australia. The standard-setting meeting began with an overview of the purpose of the meeting (to establish the Year 10 proficient standard on the NAP–SL scale to be used as the Key Performance Measure for Year 10 Science). The judges were then introduced to the mechanisms of the different standard-setting procedures to be applied and to conceptualisations of the *proficient standard*. The judges were provided opportunities to share and discuss their ratings as part of the exercise. Details of these processes are provided in the descriptions of the remaining steps in the standard-setting procedure.

Step 3: Developing a shared understanding of the proficient standard

Central to the success of setting standards is that panellists develop a consistent conceptualisation of the key referents about which they make their judgements (Cizek and Bunch, 2007. p48). In the case of this standard-setting exercise, the essential conceptualisations required by the judges related to the notion of the *proficient standard* as used across all the NAP sample assessments. This standard is described as *challenging but reasonable* or ‘achievement at a year level with students needing to demonstrate more than elementary skills expected at that year level’ (ACARA, 2015, p. 5).

The meeting included a detailed discussion with judges of the meaning of the standard in practice including that it reflected more than minimal/basic proficiency but is still realistically achievable by a student who has had ‘typical’ exposure to schooling.

The judges were then asked to consider and discuss the concept of the hypothetical borderline candidate (see, for example Cicek, 2006 p. 248) with respect to the *proficient standard*. The discussion continued until there was a consensus amongst the judges that they had individually and collectively developed sound and consistent conceptualisations of the standard and the context in which it was to be applied.

Step 4: Completing the modified Angoff (Yes/No) standard-setting method

Judges were provided a booklet containing 100 items, each with its scoring guide. The items were presented in random order in the booklet. Items for which students could receive credit of more than one-score point (known as partial-credit items) were presented once, together with the scoring guide for that item.

Under the modified Angoff (Yes/No) method, judges were required to make a judgement for each individual item (or different score category for partial credit items) about whether a hypothetical borderline candidate would answer the item sufficiently well to receive credit (Yes) or not sufficiently well to receive credit (No). Judges entered their ratings by writing '1' (Yes), '0' (No) for each item dichotomous item (each item with a maximum score of 1) on a record sheet. For partial credit-items the judges wrote the score of the highest category for which they believed the hypothetical borderline candidate would receive credit (0, 1, 2 or 3). The judges completed this exercise alone and without further consultation, except to ask for points of clarification of process with the group facilitator.

Step 5: Small group and then whole group discussion of the judges' ratings

At this point the judges were asked to discuss their item ratings in groups of four. This allowed the judges to see the variability in their judgements with others and to discuss differences. The purpose was not for the judges to achieve consensus within their groups and judges were encouraged not to change their ratings unless they felt they had made a clear error. The main purpose of this exercise was for the judges to continue to clarify the notion of the *proficient standard* and the *marginally proficient student*.

After the small group discussions, a whole group discussion was convened to discuss differences in ratings and, more importantly, to revisit the definition of the *proficient standard* and the *marginally proficient student*.

Once at the end of the first day, the judge's ratings were entered into a spreadsheet so they could be collated and viewed.

Step 6: Whole group discussion of the judges' ratings with respect to student data

The second day of the meeting began with a brief reflection on the first day and the opportunity for judges to raise any points for clarification.

Following this, the judges were presented with a visual display of their responses to all 100 items. For this purpose, the partial credit items were each separated into one row for each score category with a '1' indicating that a judge believed that the score category would be achieved and a '0' indicating that a score category would not be achieved. In total there were 118 item categories presented to the judges.

For each item the judges were shown: their judgement (1 or 0); the percentage of judges who gave each item a 1 (i.e. indicated that a *marginally proficient student* could complete the item); the weighted percentage correct (or achieved) for that item by Year 10 students in NAP–SL 2018; and the difference between the actual percentage achieved by students and the percentage of judges who believed a *marginally proficient student* would correctly complete the item. The items were ordered from the largest to the smallest difference between the estimates of the judges and the actual performance of the students (Table 7.3) with differences greater than 20 percentage points highlighted in red (where the percentage achievement estimated by the judges was less than the percentage correct achieved by students) and green (where the percentage achievement estimated by the judges was greater than the percentage achieved by the students). The data shown in Table 7.3 formed the basis for a discussion with judges. The discussion first addressed the 39 items for which the difference between the prediction of the judges and the achievement of the students differed by more than 20 percentage points.

This discussion was aimed at having the judges consider and understand why their judgements did not accord clearly with the candidate achievement. This included further discussion of the concept of the *challenging but reasonable* standard, *marginally proficient student* and the demands of each item that extended beyond knowledge and understanding of the necessary scientific content.

In some cases, this discussion resulted in judges reconsidering their judgements regarding given items. This can happen when, through the discussion the judges perceive a previously unnoticed property of an item that could account for the discrepancies in their ratings. The judges were neither required nor requested to change their ratings.

The judges were then invited to suggest for discussion any other items for which they felt student achievement was difficult to estimate.

The purpose of this extended discussion of the items and the judgements was to further support the judges to develop a common understanding of the *challenging but reasonable* standard and to raise their awareness of the properties of test items that may influence their difficulty (and judgements about their difficulty) that go beyond the scientific content knowledge and understanding in required to answer the item correctly.

Table 7.2 Summary outcomes of the modified Angoff (Yes/No) procedure

Item	**Judge**																				**Judges' (J) %**	**Actual (A)%**	**Diff (A-J) (%)**
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	36	-64
2	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	20	82	62
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	58
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	47	-53
5	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	85	38	-47
6	1	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	70	24	-46
7	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	80	35	-45
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	43
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	58	-42
10	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	90	49	-41
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	61	-39

Item	Judge																				Judges' (J) %	Actual (A)%	Diff (A-J) (%)
12	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	1	1	1	0	1	70	34	-36
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	64	-36
14	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	51	36
15	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	95	60	-35
16	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	25	60	35
17	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	95	61	-34
18	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	90	56	-34
19	1	0	0	1	0	1	1	1	1	0	0	0	1	1	1	1	1	0	0	0	55	22	-33
20	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	90	57	-33
21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	67	-33
22	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	0	1	1	0	1	55	23	-32
23	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	0	0	1	0	0	65	35	-30
24	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	40	70	30
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	90	60	-30
26	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	35	65	30
27	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	25	54	29
28	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	73	-27
29	1	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	25	52	27
30	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	0	75	49	-26
31	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	85	60	-25
32	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	40	64	24
33	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	15	39	24
34	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	95	71	-24
35	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	85	61	-24
36	1	0	0	0	0	1	0	0	1	0	1	0	0	1	1	0	0	0	1	0	35	13	-22
37	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	90	68	-22
38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	95	74	-21
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	95	74	-21
40	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	60	40	-20

Item	Judge																				Judges' (J) %	Actual (A)%	Diff (A-J) (%)
41	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	80	-20
42	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	80	60	-20
43	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	81	-19
44	1	0	1	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	30	49	19
45	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	80	61	-19
46	1	1	0	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	75	57	-18
47	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	10	28	18
48	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	82	-18
49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	82	-18
50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	82	-18
51	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	90	73	-17
52	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	95	78	-17
53	1	0	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	1	1	1	55	39	-16
54	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	10	26	16
55	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	90	74	-16
56	1	0	1	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	40	24	-16
57	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	5	20	15
58	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	30	45	15
59	1	0	0	1	0	1	1	1	0	1	1	0	1	1	1	0	1	1	0	0	60	46	-14
60	1	1	0	1	1	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1	65	51	-14
61	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	70	56	-14
62	1	0	1	1	1	0	1	1	0	1	1	1	1	1	0	0	0	1	0	1	65	79	14
63	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	95	82	-13
64	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	95	82	-13
65	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	85	73	-12
66	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	88	-12
67	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	80	68	-12
68	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	0	0	70	58	-12
69	0	0	0	0	0	0	1	0	1	0	1	1	1	1	1	1	0	1	0	1	50	39	-11

Item	Judge																				Judges' (J) %	Actual (A)%	Diff (A-J) (%)
70	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	89	-11
71	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	25	36	11
72	1	0	0	1	1	0	1	1	0	1	0	0	1	0	1	0	0	0	1	0	45	56	11
73	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	85	74	-11
74	1	1	1	1	0	1	1	0	1	0	0	0	1	1	1	0	0	0	0	0	50	60	10
75	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	90	80	-10
76	1	0	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	0	0	70	80	10
77	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	20	30	10
78	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	90	-10
79	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	95	85	-10
80	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	1	0	1	75	85	10
81	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	85	76	-9
82	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	91	-9
83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	9
84	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	95	86	-9
85	1	0	1	0	1	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	35	44	9
86	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	95	87	-8
87	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	92	-8
88	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	30	22	-8
89	0	1	1	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	30	37	7
90	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	93	-7
91	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	95	88	-7
92	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	15	22	7
93	1	0	0	1	1	0	1	1	1	1	1	1	1	0	0	1	1	0	0	0	60	54	-6
94	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	1	80	74	-6
95	1	0	0	1	1	1	1	0	0	1	0	0	1	1	1	1	0	0	0	0	50	56	6
96	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	95	90	-5
97	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	90	85	-5
98	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	90	86	-4

Item	Judge																				Judges' (J) %	Actual (A)%	Diff (A-J) (%)
99	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100	96	-4
100	1	1	0	1	1	1	1	1	1	0	0	1	0	1	0	0	1	1	0	1	65	69	4
101	1	0	1	1	0	1	0	0	0	1	1	0	0	1	0	0	0	1	1	1	50	54	4
102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
103	0	1	0	0	0	1	0	0	1	1	1	1	0	0	0	1	0	0	0	0	35	38	3
104	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	75	78	3
105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
106	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	20	17	-3
107	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	95	92	-3
108	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	25	27	2
109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
110	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	7	2
111	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	90	89	-1
112	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	80	81	1
113	1	0	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	65	66	1
114	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
115	0	1	0	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	0	1	70	69	-1
116	0	0	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1	1	45	44	-1
117	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10	10	0
118	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	10	10	0

Following this discussion, the judges were presented with the data shown in Table 7.3 (above) ordered by student percentage correct (from lowest to highest percentage correct (i.e. with the more difficult items at the top of the table). By ordering the items according to their difficulty (noting that percentage correct was regarded as an enough proxy for scaled item difficulty), it was possible to see the region of the table where the judges' judgements were shifting from all estimating correct through to all estimating incorrect. The group discussed the nature of the items in this 'middle' region (the region in which the judges' ratings were changing from all 'Yes' to all 'No' in which, according to the modified Angoff procedure it was likely the standard would lie.

Step 7: Applying the adapted Bookmark method of establishing the recommended proficient standard

The judges were then presented with a fresh booklet containing all the test items in order from least to most difficult (based on the NAP–SL scale). Judges were instructed to identify the items at the top and bottom of a range of achievement within which they were confident (on the basis of all the information with which they had been presented and discussed previously) that the relevant benchmark cut-score was located. According to this definition, the items above the top items in the range were deemed by a judge to be too difficult to reasonably expect a *marginally proficient student* to answer correctly, and the items below the bottom items in the range were deemed by the judge to be sufficiently easy for a *marginally proficient student* to be reasonably expected to answer correctly. At this point, each judge had identified an 'acceptable range' (marked by an upper and lower limit) in which they felt the standard could be set.

Step 8: Discussing the judges' ranges for the proficient standard and establishing a 'whole group' recommendation.

The judges then provided their individual recommendations for the upper and lower limits of the *proficient standard* for discussion. The recommendations were collated and displayed to the group (Figure 7.1).

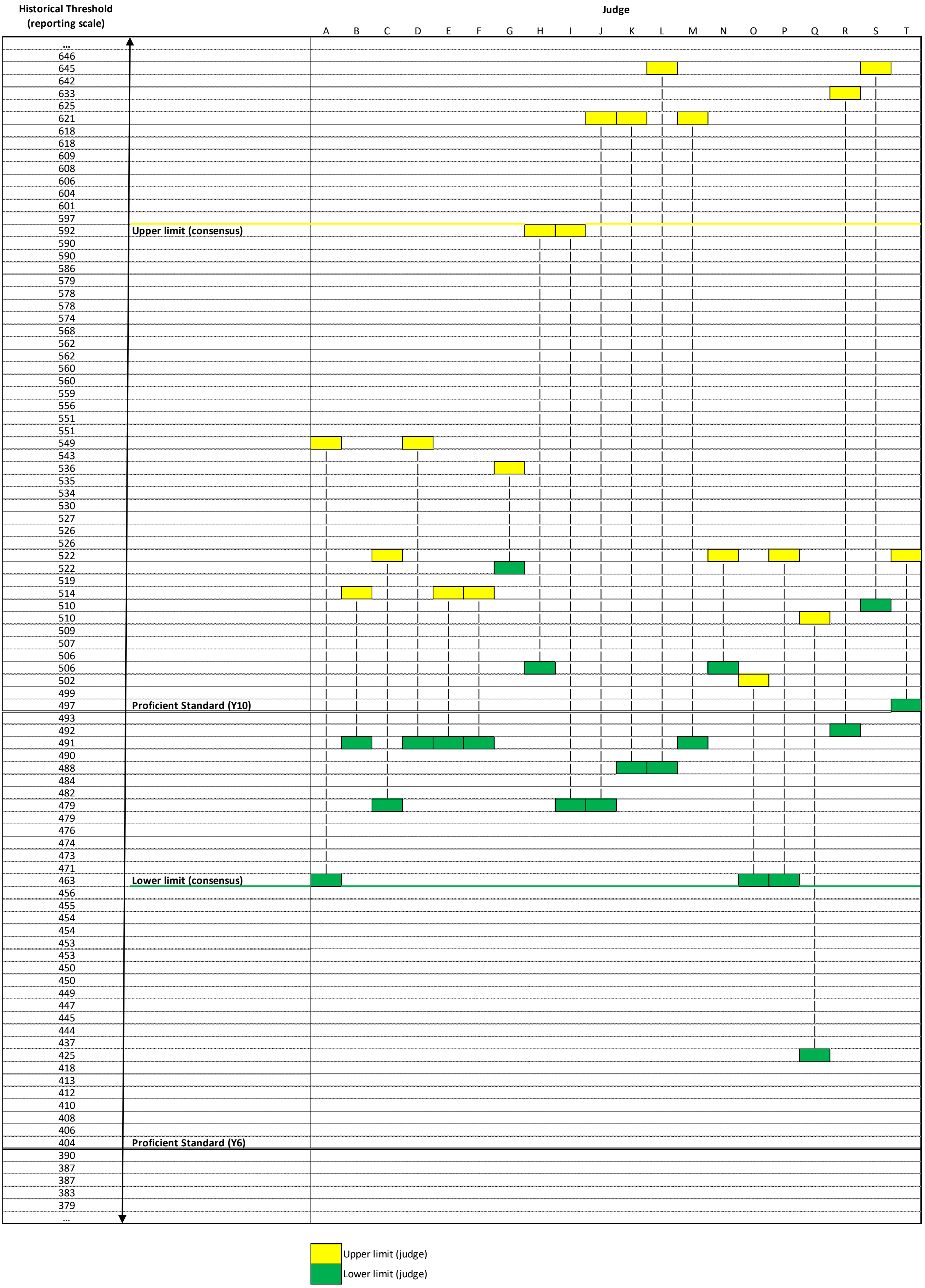
The aim of the discussion of the judges' recommended upper and lower limits was to achieve consensus among the judges for an acceptable upper and lower limit (i.e. a consensus range) within which the *proficient standard* could be set. Figure 7.1 shows that the range of the judges' lower limit judgements (shown in green) was smaller than that of their upper limit estimates (shown in yellow).

The discussion began with consideration of the lower limits proposed by the judges. The discussion focussed on the content of the items at the lower limits with further reference to expectations of a *marginally proficient student* at the standard. Ultimately, the judges agreed on a lower limit corresponding to an item located at 463 scale points on the NAP–SL scale. This was at the low-end of the judges' original recommendations.

A similar discussion was had regarding the upper limit. In this case, there were two groups of items, separated by roughly 100 NAP–SL scale points, around which judges had tended to make their recommendations for the upper limit (Figure 7.1). In this case, following discussion, the group agreed on an upper limit that was closer to the top of this range corresponding to an item located at 592 scale points on the NAP–SL scale.

As a result of the complete standard-setting procedure (modified Angoff and adapted Bookmark) the group agreed that an acceptable range for the *proficient standard* was one bounded by items at 463 (the lower limit) and 592 scale points (the upper limit) on the NAP–SL scale (Figure 7.1).

Figure 7.1 Summary outcomes of the adapted Bookmark procedure



Step 9: Validating the range of the proficient standard against the OECD PISA Science Literacy Scale and the Australian Curriculum: Science achievement standards.

As a final exercise, the judges were provided both the OECD PISA Science Literacy Scale and the Australian Curriculum: Science achievement standards for Years 6 to 10.

The judges then compared the contents of each of the OECD PISA Scientific Literacy Scale and the Australian Curriculum: Science achievement standards to the items within the range of the NAP–SL *proficient standard* recommended by the judges. The focus of their discussions were on the OECD PISA scale at Levels 2, 3 and 4 (as the existing KPM for 15-year-old Australian students was level 3 and above on the OECD PISA Scientific Literacy scale) and at Years 9 and 10 for the Australian Curriculum (given that Year 10 students complete NAP–SL in the final term of the school year).

Following the small group discussions, a whole group discussion was held to consider whether the judges were confident that the items within the proposed range for the *proficient standard* reflected comparable and reasonable levels of achievement in comparison to the PISA and Australian Curriculum standards. The ACARA Science Curriculum expert participated in this discussion with the group.

The group agreed that the proposed range for the *proficient standard* (between 463 and 592 NAP–SL scale points inclusive) was appropriate and consequently confirmed the range as their recommendation.

Confirming the proficient standard from within the acceptable range

Following the benchmarking activity, ACARA staff were provided with the full set of recommendations from the standard-setting meeting together with additional data on student achievement in NAP–SL 2018.

The most likely location for the Year 10 proficient standard was one proficiency level above the Year 6 proficient standard. While this location was within the proposed range for the Year 10 standard (scaled score of 522.9), the percentage of Year 10 students reaching the proficient standard (40%) was much lower than the percentage of Year 6 students reaching their proficient standard (58%).

The solution for this problem was to change the width or the proficiency levels. The width of the levels in previous cycles was 130 scaled score points (or 1.25 logits). Generally, the width of proficiency levels is chosen so that a student achieving at the very bottom of a level is likely to respond correctly to 50 per cent of the items in that level. All other students within the same level are likely to respond correctly to more than 50 per cent of the items in a level. Therefore, students within a level can be regarded as mastering the skills that are required to answer the items within the same level correctly.

This is achieved by changing the default response probability (RP) of 0.5. The corresponding RP for a width of 1.25 logits is 0.65. Changing the RP to 0.62 results in a level width of 1 logit (or 105 scaled score point). Making this change to the proficiency levels decreased the location of the Year 10 proficient standard to 497.3 and consequently increased the percentage of Year 10 students reaching their proficient standard to 50. The location of the Year 6 proficient standard was kept at the same location (in logits), keeping the percentage of Year 6 students reaching their proficient standard at 58.

Given these changes to the proficiency levels of the NAP–SL measurement scale, it was decided to relabel the levels. Table XX shows the new and original labels for the levels and the new cut points between the levels. A consequence of the change in proficiency levels is that percentages of students within levels could not be compared to published percentages of previous assessment cycles. The only percentage that is comparable with previous cycles is the percentage of Year 6 students reaching their proficient standard.

Table 7.3 Cut points for proficient standards and between proficiency levels

		Lower boundary		**Percentage of students**	
Original label	**New label**	**Logits**	**Scaled score**	**Year 6**	**Year 10**
Level 4 or above	Level 5 or above	2.13	602.0	2	16
Level 3.3	Level 4	1.13	497.3	17	**33**
Level 3.2	Level 3	0.13	392.6	**39**	31
Level 3.1	Level 2	-0.87	287.9	30	15
Level 2 or below	Level 1 or below			12	5

Equating errors on percentages

When comparing outcomes between cycles, uncertainty caused by the choice of link items needs to be taken into account. The previous chapter outlined the estimation of equating errors on differences between mean performance. When comparing percentages at or above the proficient standard an equating error on percentages needs to be included in a similar way.

Equating errors on percentages have been estimated using a replication method. In every replication, the proficient standard is slightly changed within the size of the equating error on means (see previous chapter) and the percentage reaching the proficient standard is recalculated. The variation in the recalculated percentages is an indication of the uncertainty caused by the current equating method. The equating error on percentages varies across subgroups in the student population and needs to be calculated for each subgroup. Table 7.4 lists the equating errors between 2018 and each of the previous cycles for all Year 6 students and by subgroup within this population.

Table 7.4 Equating errors on percentages at or above the proficient standard 9 (Year 6)

	Equating between 2018 and			
	2015	**2012**	**2009**	**2006**
All students	1.89	2.79	3.08	3.41
Non-ATSI	1.94	2.85	3.14	3.47
ATSI	1.23	1.93	2.18	2.47
Male	1.83	2.75	3.04	3.36
Female	1.96	2.85	3.14	3.47
Metropolitan	1.97	2.87	3.15	3.47
Regional	1.81	2.74	3.04	3.40
Remote	1.06	1.79	2.01	2.24
Non-LBOTE	1.85	2.76	3.05	3.40
LBOTE	2.06	2.97	3.25	3.55

ACT	1.19	1.87	2.11	2.39
NSW	1.78	2.64	2.92	3.25
NT	1.22	1.91	2.15	2.43
QLD	1.93	2.91	3.21	3.56
SA	1.60	2.33	2.59	2.90
TAS	1.94	2.99	3.32	3.68
VIC	2.18	3.08	3.37	3.71
WA	2.07	3.03	3.32	3.63

References

ACARA (2015) Measurement Framework for Schooling in Australia. Sydney: ACARA

(Accessed 12 June 2012).

Cizek, G.J. (2006). Standard Setting. In Downing, S. & Haladyna, T. (Eds.). *Handbook of Test Development (pp225-258).* Mahwah: Lawrence Erlbaum Associates.

Cizek, G. J., and Bunch, M. B., (2007). SECTION I Fundamentals of Standard Setting. Standard Setting. Thousand Oaks: SAGE Publications, Inc.

Hambleton, R. M. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In Hansche, L. N. (Ed.). *Handbook for the development of performance standards* (pp. 87-114). Washington, DC: Council of Chief State School Officers.

Hambleton, R. M. (2001): Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In Cizek (Ed.) *Setting Performance Standards Concepts, Methods and Perspectives.* (pp 89-116) Mahwah: Lawrence Erlbaum Associates Inc.

Kellow, J & Wilson, L. (2008) *Setting Standards and Establishing Cut Scores on Criterion-Referenced Assessments Some Technical and Practical Considerations*. In Osborne, J (Ed.) Best Practices in Quantitative Methods (pp 14-28). Thousand Oaks: SAGE Publications, Inc.

Linn, R. (2003). *Performance standards: Utility for different uses of assessments*. Education Policy Analysis Archives, 2003, 11 (31).

MCEETYA (2004). National Year 6 Science Assessment Report. Melbourne: Curriculum Corporation

https://www.nap.edu.au/_resources/2003_NAP_SL_Public_report.pdf

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). *The bookmark procedure: Psychological perspectives*. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives, (pp. 249-281). Majwah, NJ: Lawrence Erlbaum Associates, Inc.

Zieky, M. J. (2001). So much has changed: How the setting of cut-scores has evolved since the 1980s. In G. J Cizek (Ed.) *Setting performance standards: Concepts, methods and perspectives* (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum.

Chapter 8 REPORTING OF RESULTS

Eveline Gebhardt - ACARA

The students assessed in NAP–SL 2018 were selected using a two-stage stratified sampling procedure. At the first stage, schools were sampled from a sampling frame with a probability proportional to their size as measured by student enrolments in the relevant year level. In the second stage, several students at each year level were randomly sampled within schools. Applying cluster sampling techniques is an efficient and economical way of selecting students in educational research. However, as these samples were not obtained through (one-stage) simple random sampling, standard formulae to obtain sampling errors of population estimates are not appropriate. In addition, NAP–SL estimates were obtained using plausible value methodology (see chapter 6 on scaling procedures), which allows for estimating and combining the measurement error of achievement scores with their sampling error.

This chapter describes the method applied for estimating sampling as well as measurement error. In addition, it contains a description of the types of statistical analyses and significance tests that were carried out for reporting of results in the *NAP–SL 2018 Public Report.*

Transformation of logits to a scale with mean 400 and standard deviation 100

To facilitate the interpretation of the results, it is a common practice to transform logit scores. It was decided that, for NAP–SL assessments, the proficiency scale should have a national mean of 400 and a standard deviation of 100. This scale was chosen to avoid having negative values on the scale representing student proficiency. Further, a standard deviation of 100 provides easy interpretation of performance levels in terms of how far away a score is from the mean.

As part of the equating process the 2018 logit scores are first translated to the 2006 scale (refer to Chapter 6 for details), then transformed to the 400/100 scale.

Note that the mean of 400 is the national mean, computed using student sampling weights to reflect the average achievement of all Year 6 students in Australia. It is not the average of jurisdiction means, as that average does not consider the number of students in each jurisdiction. In summary, house weights are used to set the average score of 400, not senate weights.

Computation of sampling and measurement variance

Unbiased standard errors from studies should include both sampling variance and measurement variance. One way of estimating sampling variance on population estimates from cluster samples is by utilising the application of replication techniques. The sampling variances of population means, differences, percentages and correlation coefficients in NAP–SL studies were estimated using the jackknife repeated replication technique (JRR). The other component of the standard error of achievement test scores, the measurement variance, can be derived from the variance among the five plausible values for NAP–SL. In addition, for comparing achievement test scores with those from previous cycles (2006, 2009, 2012 and 2015), an equating error was added as a third component of the standard error.

Replicate weights

When applying the JRR method for stratified samples, primary sampling units (PSUs) – in this case schools – are paired into pseudo-strata, also called sampling zones. The assignment of schools to these sampling zones needs to be consistent with the sampling frame from which they were sampled (to obtain pairs of schools that were adjacent in the sampling frame) and zones are always constructed within explicit strata of the sampling frame. This procedure ensures that schools within each zone are as similar to each other as possible.

Within each sampling zone, one school was randomly assigned a value of two whereas the other one received a value of zero. To create replicate weights for each of these sampling zones, the jackknife indicator variable was multiplied by the original sampling weights of students within the corresponding zone so that one of the paired schools had a contribution of zero and the other school had a double contribution, whereas schools from all other sampling zones remained unmodified.

Standard errors

In order to compute the sampling variance for a statistic *t*, *t* is estimated once for the original sample *S* and then for each of the jackknife replicates J_h. The JRR variance is computed using the formula:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where *H* is the number of replicate weights, *t(S)* the statistic *t* estimated for the population using the final sampling weights, and $t(J_h)$ the same statistic estimated using the weights for the h_{th} jackknife replicate. For all statistics that are based on variables other than student test scores (plausible values), the standard error of *t* is equal to:

$$\sigma(t) = \sqrt{Var_{jrr}(t)}$$

The computation of JRR variance can be obtained for any statistic. However, many standard statistical software packages like SPSS® do not generally include any procedures for replication techniques. Therefore, specialist software, the SPSS® replicates add-in, was used to run tailored SPSS® macros to estimate JRR variance for means and percentages.

Population statistics for NAP–SL scores were always estimated using all five plausible values with standard errors reflecting both sampling and measurement error. If *t* is any computed statistic and t_i is the statistic of interest computed on one plausible value, then:

$$t = \frac{1}{M} \sum_{i=1}^{M} t_i$$

with *M* being the number of plausible values.

The sampling variance *U* is calculated as the average of the sampling variance for each plausible value U_i :

$$U = \frac{1}{M}\sum_{i=1}^{M} U_i$$

Using five plausible values for data analysis allows the estimation of the error associated with the measurement of NAP–SL due to the lack of precision of the test instrument. The measurement variance or imputation variance B_m was computed as:

$$B_m = \frac{1}{M-1}\sum_{i=1}^{M}(t_i - t)^2$$

To obtain the final standard error of NAP–SL statistics, the sampling variance and measurement variance were combined as:

$$SE = \sqrt{U + \left(1 + \frac{1}{M}\right) B_m}$$

with *U* being the sampling variance.

The 95 per cent confidence interval, as presented in the *NAP–SL 2018 Public Report*, was computed as 1.96 times the standard error. The actual 95 per cent confidence interval of a statistic is between the value of the statistic *minus* 1.96 times the standard error and the value of the statistic *plus* 1.96 times the standard error.

Reporting of mean differences

The NAP–SL 2018 Public Report included comparisons of achievement test results across states and territories; that is, means of scales and percentages were compared in graphs and tables. Each population estimate was accompanied by its 95 per cent confidence interval. In addition, tests of significance for the difference between estimates were provided, in order to flag results that were significant at the 5 per cent level ($p < 0.05$) which indicate a 95 per cent probability that these differences are **not** a result of sampling and measurement error.

The following types of significance tests for achievement mean differences in population estimates were reported:

- between states and territories
- between student sub-groups
- between this assessment cycle and previous ones in 2015, 2012, 2009 and 2006.

Mean differences between states and territories and year levels

Pairwise comparison charts allow the comparison of population estimates between one state or territory and another or between Year 6 and Year 10. Differences in means were considered significant when the test statistic *t* was outside the critical values ±1.96 (α = 0.05). The *t* value is calculated by dividing the difference in means by its standard error, which is given by the formula:

$$SE_{dif\_ij} = \sqrt{SE_i^2 + SE_j^2}$$

where SE_{dif_ij} is the standard error of the difference and SE_i and SE_j are the standard errors of the two means *i* and *j*. This computation of the standard error was only applied for comparisons between two samples that had been drawn independently from each other (for example, jurisdictions or year levels).

Mean differences between dependent sub-groups

The formula for calculating the standard error described in the previous section is not appropriate for sub-groups from the same sample. Here, the covariance between the two standard errors for sub-group estimates needs to be considered and JRR should be used to estimate correct sampling errors of mean differences. Standard errors of differences between statistics for sub-groups from the same sample (for example, groups classified according to student background characteristics) were derived using the SPSS® replicates add-in. Differences between sub-groups were considered significant when the test statistic *t* was outside the critical values ±1.96 (α = 0.05). The value *t* was calculated by dividing the mean difference by its standard error.

Mean differences between assessment cycles (2006, 2009, 2012, 2015 and 2018)

The NAP–SL 2018 Public Report also included comparisons of achievement results across assessment cycles. The process of equating tests across different achievement cycles introduced a new form of error when comparing population estimates over time: the equating or linking error. When computing the standard error, equating error as well as sampling and measurement error were considered.

The value of the equating error between 2018 and the previous assessment in 2015 was 4.39 score points on the NAP–SL scale for both year levels (see Table 6.3). When testing the difference of a statistic between these two assessment cycles, the standard error of the difference was computed as follows:

$$SE(t_{18} - t_{15}) = \sqrt{SE_{18}^2 + SE_{15}^2 + EqErr_{18\_15}^2}$$

Where t can be any statistic in units on the NAP–SL scale (mean, percentile, gender difference, but **not** percentages), SE_{18}^2 is the respective standard error of this statistic in 2018, SE_{15}^2 the corresponding standard error in 2015 and $EqErr_{18\_15}^2$ the equating error for comparing 2018 with the 2015 results.

When comparing population estimates between 2018 and the third assessment in 2012, two equating errors (between 2018 and 2015 and between 2015 and 2012) had to be taken into account. This was achieved by applying the following formula for the calculation of the standard error for differences between statistics from 2018 and 2012:

$$SE(\mu_{18} - \mu_{12}) = \sqrt{SE_{18}^2 + SE_{12}^2 + EqErr_{18\_12}^2}$$

where $EqErr_{18\_12}^2$ reflects the uncertainty associated with the equating between the assessment cycles of 2018 and 2015 (4.39 score points) as well as between 2015 and 2012 (5.03 score points). This combined equating error was equal to 6.68 score points and was calculated as:

$$EqErr_{18\_12} = \sqrt{EqErr_{18_{15}}^2 + EqErr_{15_{12}}^2}$$

Similarly, for comparisons between 2018 and the first NAP–SL assessment in 2006, the equating errors between each adjacent pair of assessments had to be considered and standard errors for differences were computed as:

$$SE(\mu_{18} - \mu_{06}) = \sqrt{SE_{18}^2 + SE_{06}^2 + EqErr_{18\_06}^2}$$

$EqErr_{18\_06}^2$ reflects the uncertainty associated with the equating between the assessment cycles of 2018 and 2015 (4.39 score points), between 2015 and 2012 (5.03 score points), between 2012 and 2009 (3.24 score points) and between 2009 and 2006 (3.68 score points). The combined equating error was equal to 8.28 score points, and was calculated as

$$EqErr_{18\_06} = \sqrt{EqErr_{18_{15}}^2 + EqErr_{15_{12}}^2 + EqErr_{12_{09}}^2 + EqErr_{09_{06}}^2}$$

To report the significance of differences between percentages at or above proficient standard s, the corresponding equating error had to be estimated using a different approach. To obtain an estimate, the following replication method was applied to estimate the equating error for percentages at the proficient standard s.

For the cut-point that defines the corresponding proficient standard at each year level (393 for Year 6 and 497 for Year 10), a number of *n* replicate cut-points were generated by adding a random error component with a mean of 0 and a standard deviation equal to the estimated equating error of 4.39 score points for comparisons between 2018 and 2015, 6.68 score points for comparisons between 2018 and 2012, 7.42 score points for comparisons between 2018 and 2009, and 8.28 score points for comparisons between 2018 and 2006. Percentages of students at or above each replicate cut-point *(ρ_n)* were computed and the equating error was estimated as:

$$EquErr(\rho) = \sqrt{\frac{(\rho_n - \rho_o)^2}{n}}$$

where ρ_o is the percentage of students at or above the (reported) proficient standard. The standard errors of the differences in percentages at or above proficient standards between 2018 and 2015 were calculated as:

$$SE(\rho_{18} - \rho_{15}) = \sqrt{SE(\rho_{18})^2 + SE(\rho_{15})^2 + EqErr(\rho_{18\_15})^2}$$

where ρ_{18} is the percentages at or above the proficient standard in 2018 and ρ_{15} in 2015,

$SE(\rho_{18})$ and $SE(\rho_{15})$ their respective standard errors, and $EqErr(\rho_{18\_15})$ the equating error for comparisons. For estimating the standard error of the corresponding differences in percentages at or above proficient standards between 2018 and 2012, the following formula was used:

$$SE(\rho_{18} - \rho_{12}) = \sqrt{SE(\rho_{18})^2 + SE(\rho_{12})^2 + EqErr(\rho_{18\_12})^2}$$

Likewise, for estimating the standard error of the corresponding differences in percentages at or above proficient standards between 2018 and 2009 and between 2018 and 2006, the following formulae were used:

$$SE(\rho_{18} - \rho_{09}) = \sqrt{SE(\rho_{18})^2 + SE(\rho_{09})^2 + EqErr(\rho_{18\_09})^2}$$

$$SE(\rho_{18} - \rho_{06}) = \sqrt{SE(\rho_{18})^2 + SE(\rho_{06})^2 + EqErr(\rho_{18\_06})^2}$$

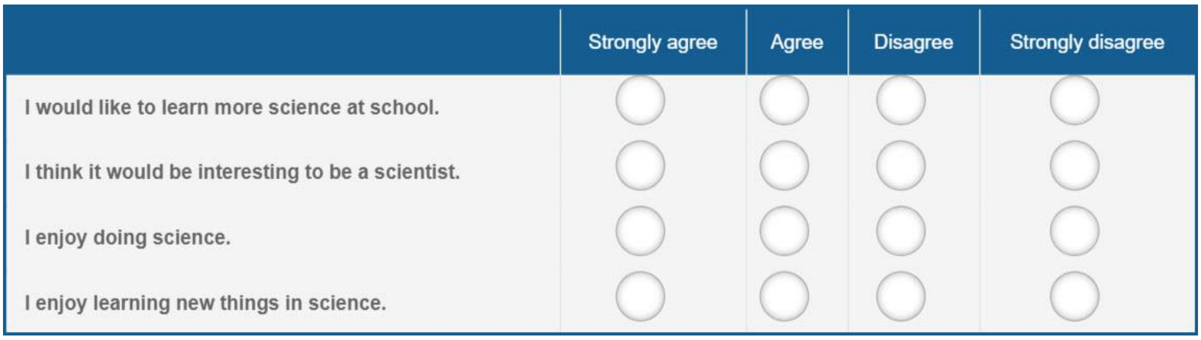
For NAP–SL 2018, equating errors on percentages were estimated for each sample or subsample of interest. Table 8.1 shows the values of these equating errors of Year 6.

Table 8.1 Year 6 equating errors for comparisons between percentages

Group	2018/2015	2018/2012	2018/2009	2018/2006
Aust.	**1.89**	**2.79**	**3.08**	**3.41**
NSW	1.78	2.64	2.92	3.25
Vic.	2.18	3.08	3.37	3.71
Qld	1.93	2.91	3.21	3.56
WA	2.07	3.03	3.32	3.63
SA	1.60	2.33	2.59	2.90
Tas.	1.94	2.99	3.32	3.68
ACT	1.19	1.87	2.11	2.39
NT	1.22	1.91	2.15	2.43
Females	1.96	2.85	3.14	3.47
Males	1.83	2.75	3.04	3.36
Non-Indigenous	1.94	2.85	3.14	3.47
Indigenous	1.23	1.93	2.18	2.47
Not LBOTE	1.85	2.76	3.05	3.40
LOBTE	2.06	2.97	3.25	3.55

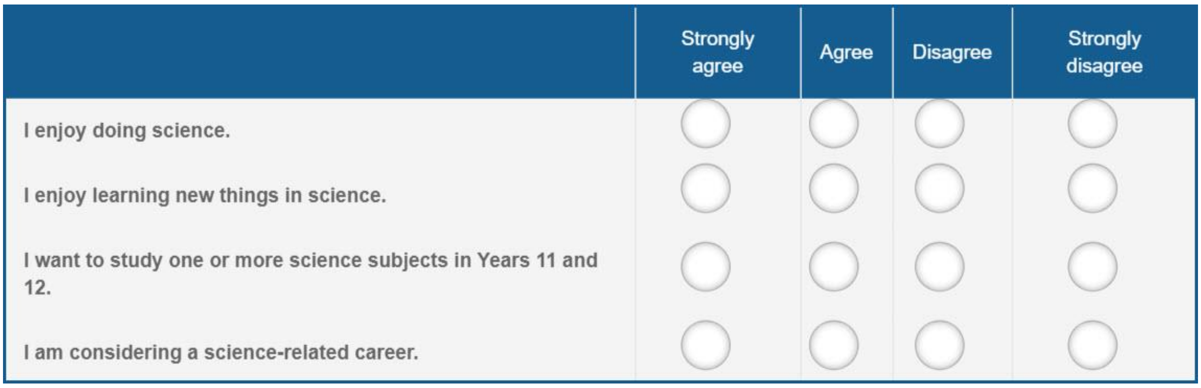
APPENDIX 1 STUDENT SURVEY

Interest in Science – Year 6



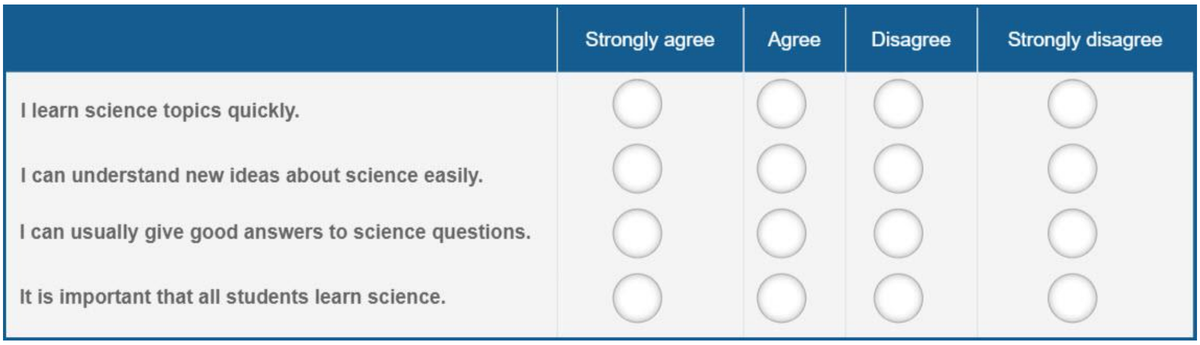
	Strongly agree	Agree	Disagree	Strongly disagree
I would like to learn more science at school.	○	○	○	○
I think it would be interesting to be a scientist.	○	○	○	○
I enjoy doing science.	○	○	○	○
I enjoy learning new things in science.	○	○	○	○

Interest in Science – Year 10



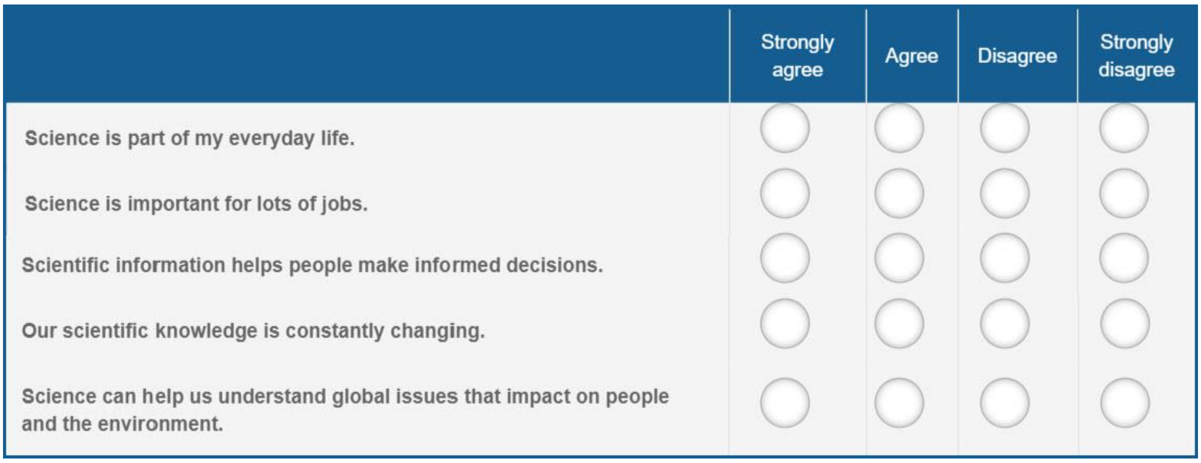
	Strongly agree	Agree	Disagree	Strongly disagree
I enjoy doing science.	○	○	○	○
I enjoy learning new things in science.	○	○	○	○
I want to study one or more science subjects in Years 11 and 12.	○	○	○	○
I am considering a science-related career.	○	○	○	○

Self-concept of Science Ability – Years 6 and 10



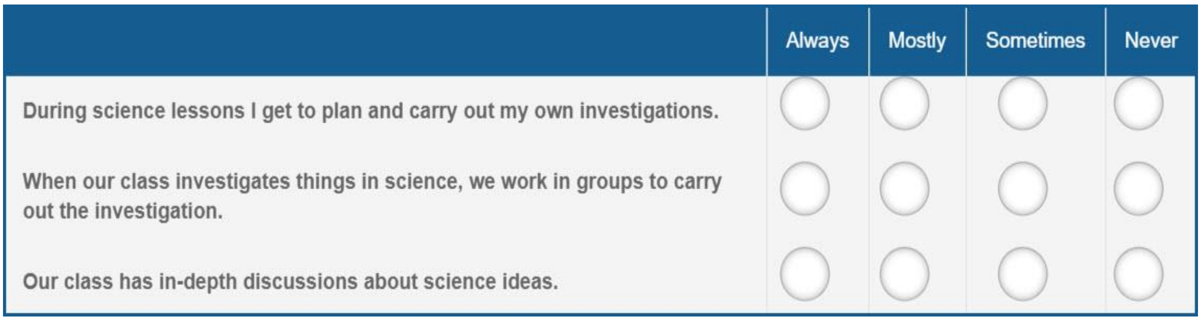
	Strongly agree	Agree	Disagree	Strongly disagree
I learn science topics quickly.	○	○	○	○
I can understand new ideas about science easily.	○	○	○	○
I can usually give good answers to science questions.	○	○	○	○
It is important that all students learn science.	○	○	○	○

Value of Science – Years 6 and 10



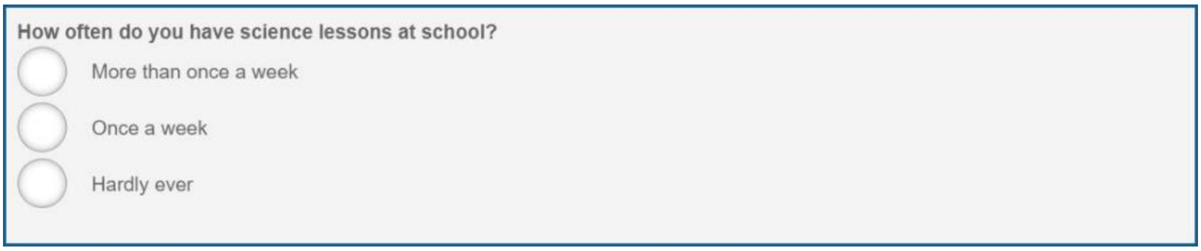
	Strongly agree	Agree	Disagree	Strongly disagree
Science is part of my everyday life.	○	○	○	○
Science is important for lots of jobs.	○	○	○	○
Scientific information helps people make informed decisions.	○	○	○	○
Our scientific knowledge is constantly changing.	○	○	○	○
Science can help us understand global issues that impact on people and the environment.	○	○	○	○

Science Teaching 1 – Years 6 and 10



	Always	Mostly	Sometimes	Never
During science lessons I get to plan and carry out my own investigations.	○	○	○	○
When our class investigates things in science, we work in groups to carry out the investigation.	○	○	○	○
Our class has in-depth discussions about science ideas.	○	○	○	○

Time Spent on Science – Year 6



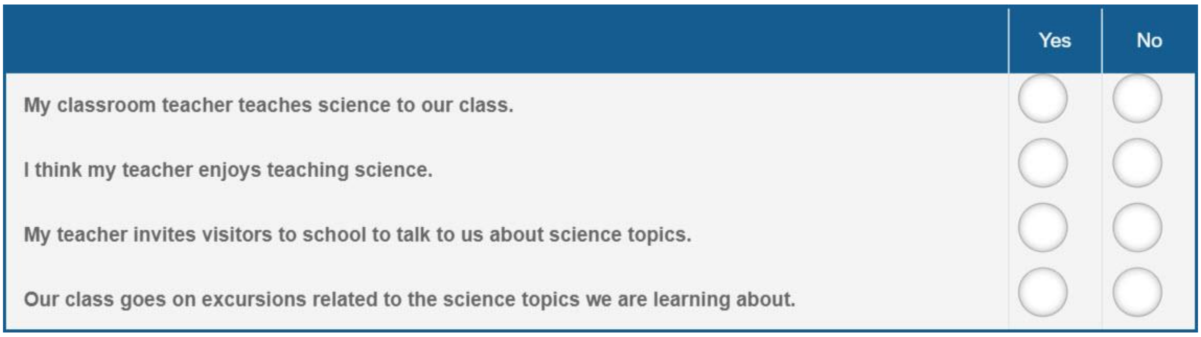
How often do you have science lessons at school?

○ More than once a week

○ Once a week

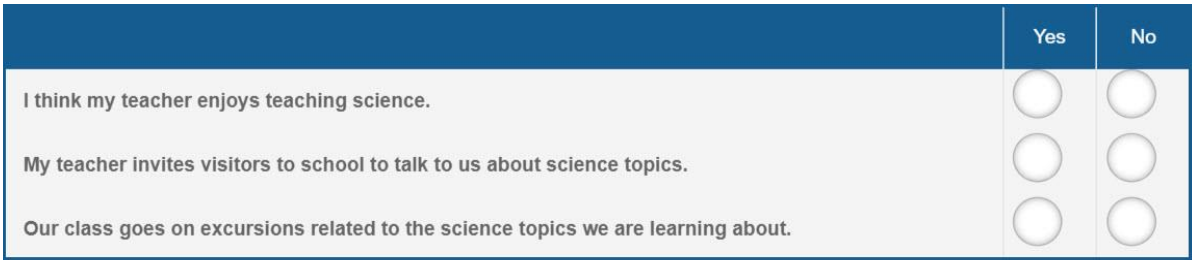
○ Hardly ever

Science Teaching 2 – Year 6



	Yes	No
My classroom teacher teaches science to our class.	○	○
I think my teacher enjoys teaching science.	○	○
My teacher invites visitors to school to talk to us about science topics.	○	○
Our class goes on excursions related to the science topics we are learning about.	○	○

Science Teaching 2 – Year 10



	Yes	No
I think my teacher enjoys teaching science.	○	○
My teacher invites visitors to school to talk to us about science topics.	○	○
Our class goes on excursions related to the science topics we are learning about.	○	○

APPENDIX 2 QUALITY MONITOR REPORT TEMPLATE

NAP-SL Main Study 2018 – QUALITY MONITOR REPORT

1. Staff Present

Who was present for the assessment session? (please check all that apply and indicate whether they were present for all or part of the test session)

Staff Member	Present for all of session	Present for part of session
Test Administrator		
School Contact		
School Technical Support Officer		
Principal		
Other (please specify) ___________________		

2. Timing

Room Set Up and Logging in

How long did it take for the computers to be switched on and logged into? _____ (mins)

Did the STSO or other school staff member assist the TA in setting up the computers?

☐ No ☐ Yes

Was the room suitably set up for the assessment and for students' optimal participation?

☐ No ☐ Yes

If No, please provide further comment.

____

____

Instructions

How long did it take the TA to lead students through the assessment instructions and practice questions? ______ (mins)

Please provide further comment if actual time was significantly different to the expected time of 10 mins.

____

____

Assessment Session

Students are given a set time allowance to complete the assessment (60 mins for Year 6 and 75 mins for Year 10). For the majority of students in this test session, was this time allowance:

☐ Too generous ☐ Just right ☐ Too short

How many students were unable to complete the assessment in the allocated time? ______

Survey **(untimed)**

How long did it take most of the students to complete the survey? ______ (mins)

How long did it take the slowest student to complete the survey? ______ (mins)

3. Test Instructions

Was the script followed according to the Test Administrator Handbook?

☐ No ☐Yes

If changes were made, were they

☐ Major ☐Minor

Why do you think the TA made changes to the script?

____

____

Do you think the variation to the script affected the performance of students?

☐ No ☐Yes

If Yes, please provide further comment.

____

____

4. Assistance Given

Were there any particular test questions that required clarification for the students?

☐ No ☐Yes

Please provide a general description of the item (e.g. Glacier ice block shapes) and a brief description of the issue/clarification given:

____

____

In your opinion, did the Test Administrator address students' questions adequately?

☐ No ☐Yes

If No, please provide further comment.

____

____

Was any extra assistance given to any students with special needs?

☐ No ☐Yes

If Yes, please provide further comment.

____

____

5. Technical Issues

Were any technical issues experienced at this school before or during the assessment session?

☐ No ☐Yes

If Yes, were they:

☐ Major ☐Minor

If technical issues were experienced, please describe what they were.

____

____

Do you think the technical issues affected the performance of students?

☐ No ☐Yes

If Yes, please provide further comment.

____

____

6. Student Behaviour

	No students	Some students	Most students
a) How many students talked and distracted other students during the assessment session?	☐	☐	☐
b) How many students made noise or moved around, causing disruption to other students during the session?	☐	☐	☐
c) How many students attempted to access their mobile phones or other personal electronic devices during the session?	☐	☐	☐
d) How many students became restless towards the end of the session?	☐	☐	☐
e) How many students appeared to be engaged in the test material?	☐	☐	☐
f) How many students appeared to struggle with understanding how to navigate the test interface?	☐	☐	☐

7. Outside Interruptions

Were the students distracted or impacted by any outside interruptions? For example:

- Announcements over the PA or intercom system
- Noise from other classes in the school
- Distractions from other students not participating in the test session within the classroom
- Students or teachers visiting the testing room

Please specify:

____

____

Questions 8 and 9 only need to be completed if the school communicates relevant information to you. It is not expected that you ask schools these questions.

8. School involvement in other national online assessments

Schools may be involved in other national online assessments.

For example:

- NAP-SL Mode-effect study (15 Oct - 2 Nov)
- SRT (20 Aug – 12 Oct)
- PRT (22 Oct – 2 Nov)

Please document any feedback you may receive from schools in relation to these other online events.

For example:

- Overlap with NAP-SL main study preparatory activity
- Identification of students for each event

____

____

9. School Receptiveness

How receptive was the school towards participating in NAP-SL? What do you perceive to be the school's overall attitude and level of commitment towards the assessment?

____

____

As a visitor, were you made to feel welcome by the school?

____

____

10. Other Comments

Please provide any other comments that you feel would help us improve this assessment and its administration.

____

____

____

Thank you very much for recording these observations.

Please transpose your observations to the online ACER Questionnaire as soon as possible following the assessment session.

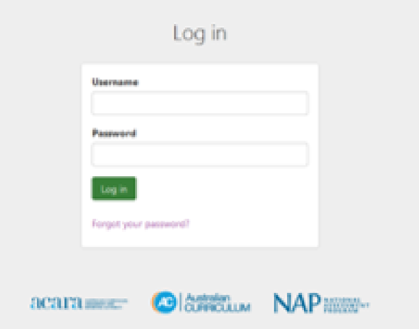
APPENDIX 3 SCHOOL SUMMARY REPORT INSTRUCTIONAL GUIDE FOR SCHOOLS

NAP-Science Literacy 2018 School Summary Report: Instructional Guide

Accessing the report

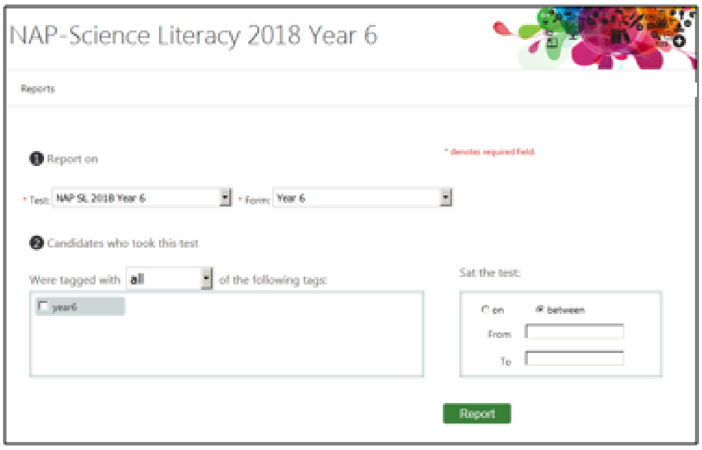
1. Navigate to the school report webpage for the required year level (Year 6 or Year 10):
   - Year 6 reports: https://oars.acer.edu.au/nap-sl-2018-year-6
   - Year 10 reports: https://oars.acer.edu.au/nap-sl-2018-year-10

2. Enter your username and password, and then click on the green 'Log in' button. Please note: your designated username and password are provided in the email to which these instructions were attached.



Login page

3. On the next page, click on the green 'Report' button. You can ignore the other text and check boxes on this page.



NAP-Science Literacy 2018 Year 6

Reports

❶ Report on

* denotes required field.

* Test: NAP SL 2018 Year 6 * Form: Year 6

❷ Candidates who took this test

Were tagged with all of the following tags:

□ year6

Sat the test:

○ on ● between

From

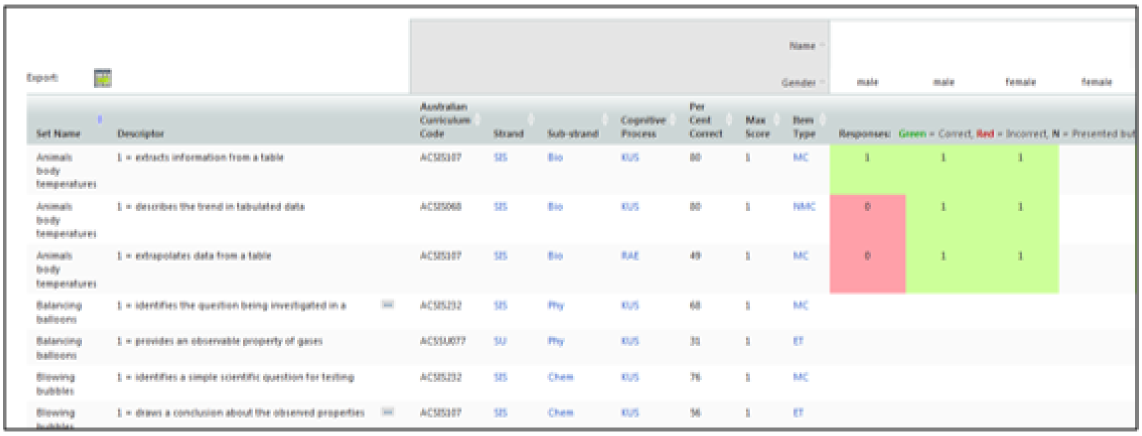
To

Report confirmation page – Year 6 view used as example

Viewing the school report

As this reporting site is currently undergoing re-development, you have two options to view the report. You can export the data as an Excel file by clicking the icon above the table.

Alternatively, you can view the interactive report online, as shown below. The report shows the results for all students in your school on all tasks included in the NAP-Science Literacy assessment. An example is given in the following screenshot. Moving horizontally on the screen can be achieved using the left and right arrow keys.



Export:

Set Name	Descriptor	Australian Curriculum Code	Strand	Sub-strand	Cognitive Process	Per Cent Correct	Max Score	Item Type	Responses: Green = Correct, Red = Incorrect, N = Presented but			
								Name				
								Gender	male	male	female	female
Animals body temperatures	1 = extracts information from a table	ACSIS107	SIS	Bio	KUS	80	1	MC	1	1	1	
Animals body temperatures	1 = describes the trend in tabulated data	ACSIS068	SIS	Bio	KUS	80	1	NMC	0	1	1	
Animals body temperatures	1 = extrapolates data from a table	ACSIS107	SIS	Bio	RAE	49	1	MC	0	1	1	
Balancing balloons	1 = identifies the question being investigated in a …	ACSIS232	SIS	Phy	KUS	68	1	MC				
Balancing balloons	1 = provides an observable property of gases	ACSSU077	SU	Phy	KUS	31	1	ET				
Blowing bubbles	1 = identifies a simple scientific question for testing	ACSIS232	SIS	Chem	KUS	76	1	MC				
Blowing bubbles	1 = draws a conclusion about the observed properties …	ACSIS107	SIS	Chem	KUS	56	1	ET				

School Report

Below is a brief description of the contents of each of the columns shown in this report.

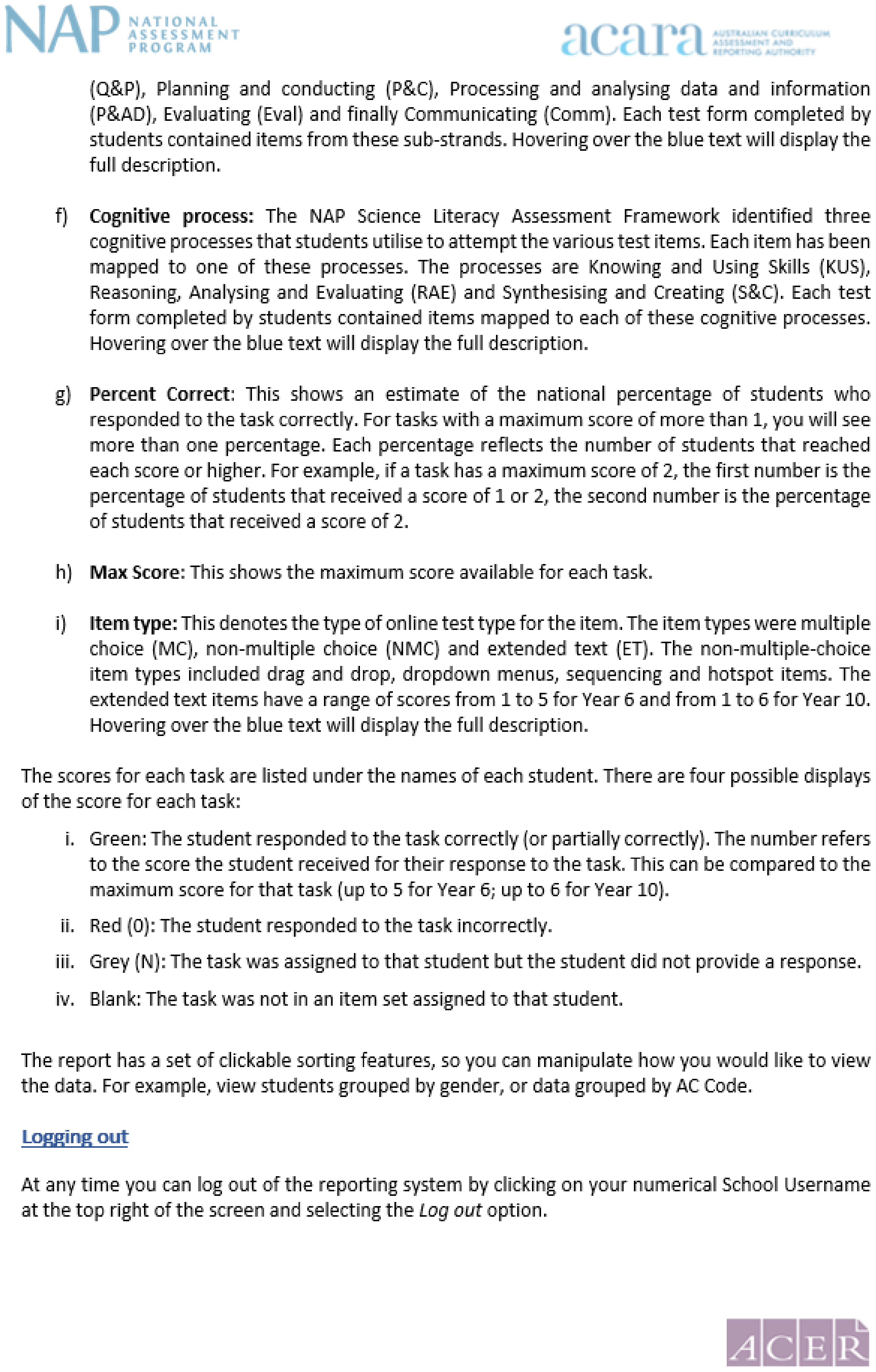
a) **Set name:** This denotes the name of the item set to which the item belongs. Each item set has a central theme and a variety of related tasks. Each student completed a combination of the various item sets.

b) **Descriptor:** This contains a brief description of what students needed to do in order to answer an item correctly. Each row refers to a single item in the assessment. You can click on the blue ellipsis (…) to expand the text for each item descriptor. For items with a maximum score of more than 1, the descriptor for each mark is provided on a separate line.

c) **AC Code:** This is the Australian Curriculum code to which the item has been mapped.

d) **Strand:** The Australian Curriculum for Science has been divided into Science Understanding (SU), Science as a Human Endeavour (SHE) and Science Inquiry Skills (SIS). Each test form completed by students contained items from these three strands. Hovering over the blue text will display the full description.

e) **Sub-strand:** Each strand in the Australian Curriculum is further divided into sub-strands. In Science Understanding, the sub-strands are Biological Sciences (Bio), Chemical Sciences (Chem), Earth and Space Sciences (E&S) and Physical Sciences (Phys). Science as a Human Endeavour is divided into Nature and development of science (N&D) and Use and influence of science (U&I). Science Inquiry Skills has the sub-strands of Questioning and predicting



(Q&P), Planning and conducting (P&C), Processing and analysing data and information (P&AD), Evaluating (Eval) and finally Communicating (Comm). Each test form completed by students contained items from these sub-strands. Hovering over the blue text will display the full description.

f) **Cognitive process:** The NAP Science Literacy Assessment Framework identified three cognitive processes that students utilise to attempt the various test items. Each item has been mapped to one of these processes. The processes are Knowing and Using Skills (KUS), Reasoning, Analysing and Evaluating (RAE) and Synthesising and Creating (S&C). Each test form completed by students contained items mapped to each of these cognitive processes. Hovering over the blue text will display the full description.

g) **Percent Correct**: This shows an estimate of the national percentage of students who responded to the task correctly. For tasks with a maximum score of more than 1, you will see more than one percentage. Each percentage reflects the number of students that reached each score or higher. For example, if a task has a maximum score of 2, the first number is the percentage of students that received a score of 1 or 2, the second number is the percentage of students that received a score of 2.

h) **Max Score:** This shows the maximum score available for each task.

i) **Item type:** This denotes the type of online test type for the item. The item types were multiple choice (MC), non-multiple choice (NMC) and extended text (ET). The non-multiple-choice item types included drag and drop, dropdown menus, sequencing and hotspot items. The extended text items have a range of scores from 1 to 5 for Year 6 and from 1 to 6 for Year 10. Hovering over the blue text will display the full description.

The scores for each task are listed under the names of each student. There are four possible displays of the score for each task:

i. Green: The student responded to the task correctly (or partially correctly). The number refers to the score the student received for their response to the task. This can be compared to the maximum score for that task (up to 5 for Year 6; up to 6 for Year 10).

ii. Red (0): The student responded to the task incorrectly.

iii. Grey (N): The task was assigned to that student but the student did not provide a response.

iv. Blank: The task was not in an item set assigned to that student.

The report has a set of clickable sorting features, so you can manipulate how you would like to view the data. For example, view students grouped by gender, or data grouped by AC Code.

Logging out

At any time you can log out of the reporting system by clicking on your numerical School Username at the top right of the screen and selecting the *Log out* option.

APPENDIX 4 ORDERED MAP OF NAP–SL 2018 ITEMS

Below are all the items that appeared in the NAP–SL 2018 tests ranked according to their scale score and proficiency level. The table also indicates the cognitive process that was required for each item. The three cognitive processes are:

KUS = Knowing and using skills

RAE = Reasoning, analysing and evaluating

S&C = Synthesising and creating.

Year level	Scale score	Level	Item Descriptor	Cognitive process
Year 10	973	Above 5	explains how forces and energy contribute to maglev trains travelling very fast	RAE
Year 10	895	Above 5	uses data from both investigations to explain the benefits for the villagers of changing glacier shapes	S&C
Year 6	827	Above 5	explains how the investigation models how a heron catches its food	RAE
Year 10	827	Above 5	describes the processes occurring at two tectonic features which provide evidence for plate tectonics	RAE
Year 10	819	Above 5	justifies the steps in a valid investigation	S&C
Year 10	796	Above 5	uses a diagram to sequence the processes that occur at a subduction zone	KUS
Year 6	789	Above 5	describes an environmental impact of biodegradable plastic bottles	RAE
Year 6	787	Above 5	identifies two variables for a presented investigation	KUS
Year 10	777	Above 5	uses diagrams to compare the circulatory systems of humans and fish	RAE
Year 10	758	Above 5	explains why it is important to use the ruler in this investigation	S&C
Year 10	756	Above 5	completes a diagram showing the position of Earth during various seasons	KUS
Year 10	753	Above 5	describes an environmental impact of biodegradable plastic bottles	RAE
Year 10	751	Above 5	describes the processes occurring at one tectonic feature and states the type of plate boundary of another tectonic feature	RAE
Year 6	740	Above 5	explains how three basic human needs could be satisfied on Mars	S&C
Year 10	738	Above 5	uses data from both investigations to support the ball recommendation	RAE
Year 10	733	Above 5	describes an accurate and reliable investigation	RAE
Year 10	732	Above 5	uses diagrams to extract information about isotopes of uranium	RAE
Year 10	711	Above 5	identifies the two diagrams representing isotopes of carbon	RAE
Year 6	707	Above 5	recognises that heat can move from one object to another	KUS
Year 10	698	5	selects two benefits for society from space research	RAE
Year 10	698	5	identifies two variables for a presented investigation	KUS

Year 10	697	5	classifies example of heat transfer that occurs when cooking on a BBQ	KUS
Year 10	685	5	classifies three examples of heat transfer that can occur in space	KUS
Year 10	685	5	draws conclusions from tabulated data	KUS
Year 6	684	5	applies knowledge of the properties of materials in an everyday situation	S&C
Year 10	683	5	identifies the electrical components needed to build a circuit with an electromagnet	KUS
Year 10	683	5	describes an impact of feral goats on the herbivores and carnivores in the food web	RAE
Year 10	681	5	identifies two correct pieces of information from a graph	RAE
Year 6	669	5	applies knowledge of electrical circuits as a means of transforming electricity	RAE
Year 10	666	5	describes that hovering leads to no contact between the train and the tracks and hence no friction	RAE
Year 10	656	5	identifies the type of plate boundary of two tectonic features	RAE
Year 6	652	5	suggests a method to increase the accuracy of measurements	KUS
Year 10	652	5	explains the role of energy and forces in determining bounce height of balls	S&C
Year 10	647	5	classifies the variables in a given investigation	KUS
Year 10	646	5	uses data from one investigation as evidence for the decision to change glacier shape	RAE
Year 10	645	5	identifies two changes to water particles during evaporation	KUS
Year 6	644	5	identifies two factors to consider when placing solar panels on a roof	KUS
Both	643	5	explains why a syringe is the most suitable measuring device for a given investigation	RAE
Year 10	642	5	interprets a diagram of the life cycle of a star	KUS
Year 6	635	5	identifies two variables to be controlled in an experiment	KUS
Year 10	633	5	classifies and defines fuels used in BBQs	KUS
Year 6	628	5	explains how two basic human needs could be satisfied on Mars	S&C
Year 6	626	5	draws conclusions from tabulated data	KUS
Year 10	625	5	describes an accurate or reliable investigation	RAE
Year 6	622	5	identifies variables held constant in a given investigation	KUS
Year 10	621	5	correctly compares the orbit lengths for Earth and Earth's Moon	KUS
Year 6	619	5	selects the variables that will be held constant in the investigation	KUS
Year 10	618	5	matches each force acting on a maglev train with the source of the force	KUS
Year 10	618	5	states that the ruler is needed to ensure investigation is fair	KUS
Year 6	617	5	provides a reason why an investigation may not be fair	KUS
Year 10	609	5	links the decrease in barn owls to the reduction in the population of mice and crickets	S&C

Year 10	608	5	describes how two adaptations help a barn owl to hunt	RAE
Year 10	606	5	suggests a method to increase the accuracy of measurements	KUS
Year 6	605	5	explains how the adaptations of a kangaroo help it to jump	RAE
Year 10	604	5	uses a diagram to sequence the steps to form an artificial glacier	RAE
Year 10	601	4	sequences the steps in the process of natural selection	S&C
Year 10	597	4	identifies three components of a fair test	KUS
Year 10	592	4	uses a graph to complete the equation for speed	RAE
Both	592	4	explains how the leaves and roots of spinifex grass help it survive in the desert	RAE
Year 10	590	4	identifies that convection transfers heat in Earth's mantle	KUS
Both	590	4	uses scales to identify the diagram that is consistent with the sowing instructions	KUS
Year 6	587	4	uses a scale to make a calculation	RAE
Year 10	586	4	describes a model of a compound	KUS
Year 10	579	4	classifies the variables in a given investigation	KUS
Year 10	578	4	describes an impact of feral goats on the herbivores or carnivores in the food web	RAE
Both	574	4	selects two ways to increase recycling of plastic bags	RAE
Both	573	4	provides relevant data as evidence that the prediction was supported	RAE
Year 6	571	4	sequences the steps in an investigation	KUS
Both	568	4	extracts information from a graph	KUS
Year 6	566	4	selects factors which are relevant when assessing the risk posed by a tsunami	RAE
Year 6	563	4	extracts information from the table	KUS
Year 10	562	4	uses diagrams to show the arrangement of magnets for a maglev train and tracks	RAE
Year 10	562	4	identifies the type of plate boundary of one tectonic feature	RAE
Year 10	560	4	uses data from one investigation to support the ball recommendation	RAE
Year 10	560	4	identifies the independent variable in the investigation	KUS
Year 10	559	4	provides evidence of a chemical reaction	KUS
Year 6	559	4	uses experimental data to describe a property of wood	RAE
Year 6	558	4	states that the prediction was not supported by the results	RAE
Year 6	557	4	identifies that light is refracted as it moves through a lens	KUS
Year 6	557	4	explains how one basic human need could be satisfied on Mars	S&C
Both	551	4	identifies variables to be controlled in a given investigation	KUS
Both	551	4	draws conclusions from tabulated results	RAE
Year 10	549	4	identifies two conclusions from the described scenario about seat belts	RAE
Year 6	544	4	classifies rubbish into recyclable and non-recyclable	RAE
Year 10	543	4	describes a property common to glass and plastic bottles	KUS
Year 6	540	4	states that plant-based plastic bottles decompose faster than other plastic bottles	KUS

Year 6	539	4	describes one factor about people and buildings that are considered in a management plan	RAE
Year 10	536	4	defines an energy transformation	KUS
Year 10	535	4	identifies the gravitational force on an astronaut	KUS
Year 6	534	4	draws an inference from data in a table and information provided	RAE
Year 10	534	4	uses a diagram to calculate available energy	KUS
Both	530	4	identifies the role of leaves	KUS
Year 10	527	4	identifies the body systems shown in a diagram	KUS
Year 10	526	4	sequences the events during a crash test	KUS
Year 10	526	4	identifies the role of energy or forces in determining the bounce height of balls	KUS
Year 10	522	4	classifies environmental features as abiotic or biotic	RAE
Year 10	522	4	recognises that telescopes use refraction	KUS
Year 10	519	4	identifies two components of a fair test	KUS
Year 6	517	4	provides an observable property of gases	KUS
Year 6	514	4	identifies natural disasters caused by geological events	KUS
Year 10	514	4	identifies the formula for calcium carbonate	KUS
Year 10	510	4	identifies the reason for mosquitos laying large numbers of eggs	KUS
Both	510	4	describes an advantage of a parasite not killing its host	RAE
Year 6	509	4	uses data from the investigation to support the ball recommendation	RAE
Year 10	509	4	identifies the testable question for the investigation	KUS
Year 6	508	4	recognises a step required to make an investigation valid	KUS
Year 10	507	4	identifies that friction is reduced for maglev trains	RAE
Year 10	506	4	states that the decision to change glacier shapes was supported by the results	RAE
Year 10	506	4	indicates on a map locations most likely to experience geological activity	RAE
Year 10	502	4	identifies the independent variable in the investigation	KUS
Year 10	499	4	states that plant-based plastic bottles decompose faster than other plastic bottles	KUS
Year 10	497	4	classifies the management strategies to reduce malaria	RAE
Year 10	493	3	identifies the unit of force	KUS
Year 6	493	3	draws a conclusion based on evidence provided	RAE
Year 10	492	3	defines a non-renewable resource	KUS
Year 10	491	3	calculates missing data in a table	KUS
Year 10	490	3	sequences the steps in an investigation	KUS
Year 6	490	3	provides an advantage and a disadvantage using a solar powered toy	RAE
Year 6	489	3	classifies variables in an investigation	KUS
Year 6	489	3	describes a property common to glass and plastic bottles	KUS
Year 10	488	3	classifies a substance as a mixture	KUS
Year 10	484	3	selects two adaptations of the described animal	KUS
Year 10	482	3	identifies a way to improve reliability in the investigation	KUS
Year 6	482	3	identifies that a state of change occurs during melting	KUS
Year 6	479	3	identifies that Earth spinning on its axis causes night and	KUS

			day	
Both	479	3	identifies the relationship between two organisms as parasitic	KUS
Year 10	479	3	identifies the cause of night and day	KUS
Year 6	478	3	links a structure with the role it plays when a kangaroo jumps	RAE
Year 6	478	3	explains why variables should be changed and measured in fair tests	KUS
Both	477	3	describes why a syringe allows more accurate measurements	RAE
Year 6	474	3	extrapolates data from a table	RAE
Year 10	474	3	suggests the cause for an abnormal result in a trial	KUS
Year 10	473	3	matches each atomic particle with its charge	KUS
Year 10	471	3	identifies that newton is the unit of force	KUS
Year 6	470	3	identifies the variable that will be changed in the investigation	KUS
Year 10	463	3	identifies the number of electrons in a neutral atom of carbon	RAE
Both	462	3	explains how the leaves or roots of spinifex grass help it survive in the desert	RAE
Year 6	459	3	identifies an observable feature that can be used to group living things	KUS
Year 6	457	3	describes that large numbers of eggs increases chance of hatching	KUS
Year 10	456	3	recommends the best type of ball for handball (based on the provided results)	KUS
Both	455	3	identifies the question to be tested in an investigation	KUS
Year 10	454	3	recognises that government websites are the most reliable sources of information	KUS
Year 10	453	3	describes how one adaptation helps a barn owl to hunt	RAE
Year 10	453	3	converts centimetres to millimetres	KUS
Year 10	450	3	completes the energy transformation that occurs during a dive	KUS
Year 10	450	3	identifies a specific section of a distance-time graph	KUS
Year 10	449	3	states that mice and crickets would have less food if crops were no longer grown	RAE
Both	445	3	identifies a suitable change to an experimental design	RAE
Year 6	444	3	identifies why an investigation is a fair test	KUS
Year 6	444	3	interprets provided text and a diagram	RAE
Year 10	444	3	identifies that feral goats compete with the herbivores	KUS
Year 6	442	3	identifies a point on a graph	KUS
Year 6	441	3	describes objects that form shadows as opaque	KUS
Year 6	437	3	selects two adaptations of the described animal	RAE
Year 6	437	3	identifies two reasons why measurements may be difficult in a given investigation	KUS
Year 6	436	3	describes the effect of a parasite killing its host	KUS
Year 6	433	3	describes how long back legs help fleas to survive	KUS
Both	430	3	states that the prediction was supported by the results	KUS
Year 10	425	3	extracts information from text	KUS

Year 6	421	3	identifies a way to improve the quality of the results	RAE
Year 10	418	3	classifies substances as elements or compounds	KUS
Year 6	415	3	matches the parts of a tsunami early warning system with their function	RAE
Year 6	415	3	interprets information provided in a table	KUS
Year 6	414	3	draws a conclusion about the observed properties of liquids in an investigation	RAE
Year 6	414	3	identifies the most appropriate graph for the experiment results	KUS
Year 10	413	3	extracts information from a diagram	KUS
Year 10	412	3	uses a graph to extrapolate data	KUS
Year 10	410	3	uses a graph to rank the results	KUS
Year 10	408	3	calculates the missing average in a data table	KUS
Year 10	406	3	uses a diagram to describe the malaria microorganism	KUS
Year 6	406	3	selects a point on a graph	KUS
Year 10	404	3	identifies a reason for multiple trials in an investigation	KUS
Year 6	392	2	describes one factor about people or buildings that are considered in a management plan	RAE
Year 6	391	2	uses a diagram to identify and predict an outcome	RAE
Year 6	390	2	identifies and defines an irreversible change	KUS
Year 10	390	2	identifies one component of a fair test	KUS
Year 6	388	2	uses information in a diagram to identify an object shown in a photograph	KUS
Both	387	2	extracts information from a life cycle diagram	KUS
Both	387	2	indicates the benefits of using plant-based plastic bags	RAE
Year 6	387	2	uses simple column graphs to represent data	KUS
Year 10	383	2	correctly compares the sizes of the sun, Earth and Earth's moon	KUS
Year 6	380	2	identifies a requirement when planning an appropriate investigation method	KUS
Both	379	2	selects two environmental impacts of plastic bags	KUS
Year 6	377	2	identifies that astronomers study the solar system	KUS
Year 6	377	2	identifies one variable to be controlled in an experiment	KUS
Year 10	377	2	identifies a point on a graph	KUS
Year 6	376	2	locates a point on a line graph	KUS
Year 10	374	2	uses a diagram to complete a flowchart of blood circulation	KUS
Year 10	374	2	selects most suitable measuring device for an investigation	KUS
Year 10	373	2	identifies that distances in space are measured in light years	KUS
Year 6	372	2	identifies that seismometers can detect earthquakes	KUS
Year 10	371	2	identifies that gravity acts on all objects in the universe	KUS
Both	369	2	identifies an environmental change caused by nature	KUS
Year 10	369	2	defines kinetic energy	KUS
Year 6	368	2	extrapolates information from the text	RAE
Year 10	367	2	applies experimental data to a real-life situation	RAE
Year 6	367	2	interprets a graph to determine the trend	RAE
Year 6	364	2	recommends the best type of ball for handball (based on	KUS

			the provided results)	
Year 10	363	2	identifies a situation with similar energy changes to a pool dive	KUS
Year 10	363	2	selects the most accurate piece of equipment to measure volume	KUS
Year 10	360	2	identifies a specific section of a distance-time graph	RAE
Year 6	357	2	identifies a component in an electrical circuit that performs a given function	KUS
Year 10	357	2	identifies a combustion reaction	KUS
Year 10	357	2	identifies the interaction of forces during a rocket launch	KUS
Year 6	356	2	identifies that today's model of the solar system has the Sun in the centre	KUS
Year 6	353	2	provides an advantage or disadvantage using a solar powered toy	RAE
Year 6	353	2	identifies a method of separating steel waste from glass waste	KUS
Year 6	350	2	identifies an accurate way to measure time	KUS
Year 6	349	2	shows awareness that science involves using evidence to develop explanations of events	RAE
Year 6	347	2	identifies a testable question for a given investigation	KUS
Year 6	346	2	identifies the question being investigated in a scientific investigation	KUS
Year 10	345	2	describes why scientists change and accept new theories	KUS
Year 6	344	2	identifies three planets in our Solar System	KUS
Year 6	342	2	describes the energy transformation in a given electrical circuit	KUS
Year 6	341	2	predicts the change of state when temperature is deceased	KUS
Year 6	340	2	describes the change of state during freezing	KUS
Year 6	337	2	ranks the sun, Earth and Earth's moon from largest to smallest	KUS
Year 10	330	2	identifies that sound cannot be heard in the vacuum of space	KUS
Year 6	327	2	identifies an example of a reversible reaction	KUS
Year 10	326	2	defines photosynthesis	KUS
Year 10	324	2	identifies that the Big Bang is the current theory about the origins of the universe	KUS
Year 10	319	2	classifies the advantages and disadvantages of maglev trains	RAE
Both	321	2	converts centimetres to metres	KUS
Year 6	319	2	identifies correct observations from a diagram	KUS
Year 6	316	2	interprets data in a table to make a comparison	RAE
Year 6	316	2	lists two or three basic human needs	KUS
Year 10	316	2	uses diagram to identify that earthquakes occur at plate boundaries	KUS
Year 6	314	2	compares the properties of a solid and a liquid	KUS
Year 10	311	2	identifies prey of a barn owl in the food web	RAE
Year 6	311	2	identifies that boiling water changes into a gas and evaporates	KUS
Year 6	307	2	determines the position of the sun to form a shadow	KUS

Year 6	303	2	describes how buildings form shadows	KUS
Year 6	302	2	identifies a simple scientific question for testing	KUS
Year 6	301	2	selects most suitable measuring device for an investigation	KUS
Year 6	300	2	identifies that geologists study earthquakes and tsunamis	KUS
Year 10	298	2	uses diagrams to describe lunar and solar eclipses	KUS
Both	297	2	selects the most suitable graph to display the results	KUS
Year 6	296	2	classifies objects as solids, liquids or gases	KUS
Year 6	296	2	describes cause and effect in the context of changes due to natural processes	RAE
Both	295	2	identifies the trend in graphical data	RAE
Both	295	2	locates a point on a graph	KUS
Year 10	292	2	identifies the formula for carbon dioxide	KUS
Year 6	290	2	extracts information from a table	KUS
Year 10	289	2	describes the function of the lungs	KUS
Year 10	287	1	identifies that a dog has a similar heart to humans	KUS
Year 6	286	1	uses a table to draw conclusions about relationships in data	RAE
Year 10	282	1	recognises that reflection of light forms an image	KUS
Year 6	282	1	describes the trend in tabulated data	KUS
Year 6	279	1	selects two reasons why a given circuit may stop working	KUS
Year 10	278	1	selects the correct data from a table	KUS
Year 6	278	1	lists one basic human need	KUS
Both	275	1	uses a table to order the results for a series of trials	RAE
Year 6	270	1	identifies an impact of an earthquake	KUS
Year 10	269	1	identifies that the immune system responds to diseases	KUS
Year 6	269	1	identifies the scientific question being tested	KUS
Year 6	263	1	interprets tabulated data	RAE
Year 6	261	1	describes advantages and disadvantages of burning rubbish	RAE
Year 6	259	1	describes the change of state when liquids are cooled	KUS
Year 6	259	1	identifies that wings help animals to fly	KUS
Year 6	254	1	identifies the purposes of different types of animal feet	RAE
Year 6	253	1	extracts information from a graph	RAE
Year 10	251	1	identifies the producers in the food web	KUS
Year 6	250	1	selects two animals that can shelter under spinifex	RAE
Year 6	250	1	classifies objects as solids or liquids	KUS
Year 6	247	1	provides labels on a column graph	KUS
Year 10	237	1	classifies predators and prey included in the food web	KUS
Year 6	235	1	defines an irreversible reaction	KUS
Year 10	235	1	identifies that physicists study forces and motion	KUS
Year 6	233	1	uses a diagram to order the planets from closest to furthest from the Sun	KUS
Year 6	224	1	selects appropriate equipment for the investigation	KUS
Both	221	1	describes the role of fertilisers	RAE
Both	220	1	extracts information from a table	KUS
Year 6	211	1	uses a diagram to determine the number of wires in a circuit	KUS

Year 6	208	1	makes a comparison based on data provided in a table	RAE
Both	198	1	provides a reason for wearing safety goggles	KUS
Year 6	171	Below 1	identifies that batteries were the energy source in a given electrical circuit	KUS
Year 6	154	Below 1	interprets information provided in a diagram	KUS
Year 6	133	Below 1	identifies an indicator that an electrical circuit is operating	KUS
Year 10	114	Below 1	shows awareness that animals depend on each other and the environment to survive	KUS
Year 6	93	Below 1	identifies two electrical devices in the home	KUS
Year 6	79	Below 1	uses a life cycle to complete a flowchart	KUS
Year 6	79	Below 1	classifies waste products into different categories	KUS
Year 6	68	Below 1	identifies structures that are involved in jumping	RAE
Year 6	-5	Below 1	identifies that decreasing temperatures will change a liquid to a solid	KUS
Year 6	-29	Below 1	identifies anatomical structures used for swimming or walking	KUS
Year 6	-29	Below 1	identifies that warmer temperatures can melt frozen solids	KUS

APPENDIX 5 ITEM DIFFICULTIES

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0001	x00121214	1	Year 6	No	-0.77	-1.14	-0.65	311	77		0.89	
I0002	x00121218	1	Year 6	No	-0.61	-0.99	-0.50	327	74		1.01	
I0003	x00120912	1	Year 6	No	-1.16	-1.53	-1.04	270	83		0.95	
I0004	x00120906	1	Year 6	No	-0.18	-0.56	-0.07	372	67		1.00	
I0005	x00120922	1	Year 6	No	-0.87	-1.25	-0.76	300	79		0.95	
I0006	x00121246	1	Year 6	No	-4.01	-4.39	-3.90	-29	97		1.76	
I0007	x00121245	1	Year 6	No	-0.47	-0.85	-0.36	341	73		0.99	
I0008	x00121247	1	Year 6	No	-0.49	-0.86	-0.37	340	74		1.08	
I0009	x00121244	1	Year 6	No	-1.49	-1.87	-1.38	235	86		0.86	
I0010	x00121156	1	Year 6	No	0.84	0.47	0.96	479	47		1.06	
I0011	x00121158	1	Year 6	No	1.59	1.22	1.71	558	31		1.10	
I0012	x00121157	1	Year 6	No	-0.33	-0.71	-0.22	356	71		0.96	
I0013	x00121161	1	Year 6	No	-0.13	-0.51	-0.02	377	67		1.04	
I0014	x00121336	1	Year 6	No	-0.80	-1.18	-0.69	307	78		1.01	
I0015	x00121338	1	Year 6	No	-0.84	-1.21	-0.72	303	77		0.99	
I0016	x00121341	1	Year 6	No	0.48	0.10	0.59	441	54		1.02	
I0017	x00121342	1	Year 6	No	0.00	-0.37	0.12	391	65		1.05	
I0018	x00121260	1	Year 6	No	-3.78	-4.16	-3.67	-5	97		1.45	
I0019	x00121263	1	Year 6	No	-0.42	-0.80	-0.31	347	72		0.88	
I0020	x00121448	1	Year 6	No	-1.32	-1.70	-1.21	253	84		0.87	
I0021	x00121380	1	Year 6	No	0.51	0.13	0.62	444	54		0.97	
I0022	x00118667	1	Year 6	No	-0.39	-0.77	-0.28	350	70		0.93	
I0023	x00119003	1	Year 6	No	0.29	-0.09	0.40	421	55		1.17	
I0024	x00121668	1	Year 6	No	-1.72	-2.10	-1.61	211	89		0.94	
I0025	x00121669	1	Year 6	No	-2.10	-2.48	-1.99	171	92		0.87	
I0026	x00121672	1	Year 6	No	-2.46	-2.84	-2.35	133	93		0.98	
I0027	x00121586	1	Year 6	No	-0.97	-1.34	-0.85	290	79		0.95	
I0028	x00121590	1	Year 6	No	0.79	0.41	0.90	474	49		1.02	
I0029	x00120787	1	Year 6	No	0.75	0.38	0.87	470	49		1.06	
I0030	x00130155	1	Year 6	Yes	-0.91	-1.29	-0.80	296	76		1.10	
I0031	x00130158	1	Year 6	Yes	-0.40	-0.78	-0.29	349	67		1.02	
I0032	x00130159	1	Year 6	Yes	-1.16	-1.54	-1.05	269	79		0.90	
I0033	x00130165	1	Year 6	Yes	-0.85	-1.23	-0.74	302	75		0.94	
I0034	x00130172	1	Year 6	Yes	0.63	0.25	0.74	457	49		1.13	
I0035	x00130173	1	Year 6	Yes	-0.22	-0.59	-0.10	368	65		1.03	
I0036	x00130179	1	Year 6	Yes	-0.14	-0.51	-0.02	376	61		0.90	
I0037	x00000408	1	Year 6	Yes	0.51	0.13	0.62	444	48		1.05	
I0038	x00000410	1	Year 6	Yes	-1.00	-1.38	-0.89	286	78		0.95	

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0039	x00000412	1	Year 6	Yes	0.65	0.28	0.77	459	48		1.08	
I0040	x00000454	1	Year 6	Yes	-0.43	-0.81	-0.32	346	68		0.97	
I0041	x00000173	1	Year 6	Yes	-1.75	-2.12	-1.63	208	83		1.07	
I0042	x00000174	1	Year 6	Yes	-0.71	-1.09	-0.60	316	70		0.97	
I0043	x00000483	1	Year 6	Yes	0.23	-0.15	0.34	415	56		0.94	
I0044	x00000485	1	Year 6	Yes	-2.26	-2.64	-2.15	154	88		1.13	
I0045	x00000505	1	Year 6	Yes	-0.02	-0.40	0.09	388	61		1.06	
I0046	x00000476	1	Year 6	Yes	-0.10	-0.48	0.01	380	64		1.11	
I0047	x00000388	1	Year 6	Yes	-0.37	-0.74	-0.25	353	64		0.99	
I0048	x00000204	1	Year 6	Yes	-0.04	-0.41	0.08	387	62		0.96	
I0049	x00121206	1	Year 6	No	-0.91	-1.28	-0.79	296	79		0.90	
I0050	x00121221	1	Year 6	No	-0.01	-0.39	0.10	390	64		1.07	
I0051	x00120891	1	Year 6	No	1.67	1.29	1.78	566	29		0.97	
I0052	x00120896	1	Year 6	No	0.23	-0.14	0.35	415	60		0.97	
I0053	x00120902	2	Year 6	No	0.71	0.33	0.82	465	49		1.08	
I0054	x00120911	1	Year 6	No	1.18	0.80	1.29	514	38		1.07	
I0055	x00121241	1	Year 6	No	-1.35	-1.73	-1.24	250	85		0.95	
I0056	x00121243	1	Year 6	No	-0.73	-1.11	-0.62	314	77		0.96	
I0057	x00121155	1	Year 6	No	-0.45	-0.83	-0.34	344	73		0.96	
I0058	x00121154	1	Year 6	No	-1.51	-1.89	-1.40	233	87		0.91	
I0059	x00121159	1	Year 6	No	-0.52	-0.89	-0.40	337	73		1.03	
I0060	x00121181	1	Year 6	No	-1.26	-1.64	-1.15	259	84		0.98	
I0061	x00121184	1	Year 6	No	-4.01	-4.39	-3.90	-29	98		1.40	
I0062	x00121193	1	Year 6	No	-1.30	-1.68	-1.19	254	83		0.98	
I0063	x00119112	1	Year 6	No	2.21	1.83	2.32	622	24		0.95	
I0064	x00121259	1	Year 6	No	-1.26	-1.63	-1.14	259	85		1.05	
I0065	x00121255	1	Year 6	No	-0.68	-1.06	-0.57	319	78		1.04	
I0066	x00121264	1	Year 6	No	0.14	-0.24	0.25	406	62		0.93	
I0067	x00121516	1	Year 6	No	-1.34	-1.72	-1.23	250	85		0.89	
I0068	x00120800	1	Year 6	No	0.87	0.49	0.98	482	45		1.00	
I0069	x00121360	1	Year 6	No	1.47	1.09	1.58	544	31		1.13	
I0070	x00121383	1	Year 6	No	-1.25	-1.62	-1.13	261	82		0.85	
I0071	x00118666	1	Year 6	No	2.18	1.81	2.30	619	22		0.93	
I0072	x00118614	1	Year 6	No	-1.59	-1.97	-1.48	224	85		1.00	
I0073	x00118697	1	Year 6	No	-1.38	-1.75	-1.26	247	82		0.98	
I0074	x00118710	1	Year 6	No	-0.23	-0.60	-0.11	367	66		0.92	
I0075	x00119010	1	Year 6	No	0.94	0.57	1.06	490	43		1.01	
I0076	x00121670	1	Year 6	No	-0.47	-0.84	-0.35	342	72		1.11	
I0077	x00121673	1	Year 6	No	-1.07	-1.44	-0.95	279	80		0.88	
I0078	x00121674	1	Year 6	No	-2.85	-3.22	-2.73	93	94		1.10	
I0079	x00121593	5	Year 6	No	1.09	0.71	1.20	505	45		1.09	

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0080	x00121588	1	Year 6	No	-1.04	-1.42	-0.93	282	79		0.95	
I0081	x00121414	1	Year 6	No	1.88	1.50	1.99	587	26		1.08	
I0082	x00123620	1	Year 6	No	-2.98	-3.36	-2.87	79	95		1.14	
I0083	x00130154	2	Year 6	Yes	0.30	-0.08	0.41	422	56		1.14	
I0084	x00130163	1	Year 6	Yes	0.22	-0.16	0.33	414	56		1.08	
I0085	x00130181	1	Year 6	Yes	2.42	2.04	2.53	644	15		1.01	
I0086	x00130176	2	Year 6	Yes	1.10	0.72	1.21	506	40		1.14	
I0087	x00000386	1	Year 6	Yes	-2.98	-3.36	-2.87	78	95		1.01	
I0088	x00121196	3	Year 6	No	-0.07	-0.44	0.05	384	53		1.20	
I0089	x00119198	2	Year 6	No	0.44	0.06	0.55	437	56		1.16	
I0090	x00119001	1	Year 6	No	1.59	1.22	1.71	558	32		0.91	
I0091	x00119011	1	Year 6	No	2.16	1.78	2.27	617	22		1.01	
I0092	x00119016	2	Year 6	No	2.90	2.53	3.02	695	16		1.02	
I0093	x00130152	1	Year 6	Yes	-0.33	-0.70	-0.21	357	65		1.15	
I0094	x00130153	1	Year 6	Yes	2.66	2.28	2.77	669	13		1.00	
I0095	x00130168	1	Year 6	Yes	0.22	-0.16	0.34	414	56		0.92	
I0096	x00130169	1	Year 6	Yes	0.83	0.45	0.94	478	43		1.02	
I0097	x00130170	1	Year 6	Yes	0.43	0.06	0.55	436	51		0.99	
I0098	x00130171	1	Year 6	Yes	0.40	0.03	0.52	433	52		1.09	
I0099	x00130174	1	Year 6	Yes	-1.23	-1.60	-1.11	262	79		0.91	
I0100	x00130175	1	Year 6	Yes	1.37	0.99	1.48	534	29		1.11	
I0101	x00000453	1	Year 6	Yes	1.21	0.83	1.32	517	34		0.95	
I0102	x00000428	1	Year 6	Yes	0.98	0.60	1.09	493	41		0.86	
I0103	x00000475	1	Year 6	Yes	1.12	0.74	1.23	508	41		1.01	
I0104	x00000205	1	Year 6	Yes	3.02	2.64	3.13	707	10		0.94	
I0105	x00000207	2	Year 6	Yes	2.20	1.83	2.32	622	14		0.98	
I0106a	x00120562a	1	Year 6	No	-0.86	-1.23	-0.74	301	78		1.09	
I0106b	x00120562b	1	Year 10	No	-0.16	-0.54	-0.05	374		79		0.95
I0107	x00120564	1	Link	No	-0.67	-1.05	-0.56	321	76	82	1.05	1.24
I0108	x00119110	1	Link	No	0.62	0.24	0.73	455	49	64	1.16	1.07
I0109	x00119139	1	Link	No	-0.90	-1.28	-0.79	297	80	86	1.02	1.20
I0110a	x00119813a	1	Year 6	No	0.49	0.11	0.60	442	53		1.01	
I0110b	x00119813b	1	Year 10	No	-0.13	-0.51	-0.02	377		79		1.03
I0111	x00121440	1	Link	No	-0.92	-1.30	-0.81	295	80	85	1.04	1.11
I0112	x00121450	1	Link	No	-0.21	-0.59	-0.10	369	68	79	0.95	1.02
I0113	x00121530	1	Link	No	1.32	0.95	1.44	530	34	58	1.07	0.97
I0114	x00121533	1	Link	No	-1.62	-2.00	-1.51	221	86	92	0.91	0.92
I0115	x00121368	1	Link	No	-1.84	-2.21	-1.72	198	88	90	1.04	1.42
I0116	x00121381	1	Link	No	0.51	0.14	0.63	445	52	69	0.99	0.88
I0117	x00121420	1	Link	No	-0.03	-0.41	0.08	387	64	79	1.11	0.95
I0118	x00121426	1	Link	No	0.84	0.46	0.95	479	47	71	0.98	0.81

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0119	x00120773	1	Link	No	-1.63	-2.01	-1.52	220	86	94	0.93	0.66
I0120	x00120561	1	Link	No	1.53	1.16	1.65	551	33	48	1.03	1.02
I0121	x00119136	1	Link	No	0.44	0.06	0.55	437	56	69	0.94	0.94
I0122	x00119138	1	Link	No	-1.11	-1.48	-0.99	275	81	86	0.93	1.21
I0123	x00119150	1	Link	No	-0.92	-1.29	-0.80	295	78	91	1.06	0.82
I0124	x00121444	1	Link	No	1.69	1.32	1.81	568	31	51	1.04	0.92
I0125a	x00121445a	1	Year 6	No	0.44	0.06	0.55	437	55		0.93	
I0125b	x00121445b	1	Year 10	No	0.89	0.51	1.00	484		63		1.08
I0126	x00121454	2	Link	No	1.30	0.92	1.41	527	38	56	1.03	1.02
I0127	x00121525	1	Link	No	1.90	1.52	2.01	590	25	44	1.06	1.02
I0128a	x00122348a	1	Year 6	No	2.24	1.87	2.36	626	19		1.00	
I0128b	x00122348b	1	Year 10	No	2.80	2.43	2.92	685		26		0.99
I0129	x00120815	1	Link	No	-0.12	-0.49	0.00	379	66	74	0.94	1.15
I0130	x00120821	1	Link	No	-0.04	-0.41	0.08	387	64	75	0.97	1.01
I0131	x00120810	1	Link	No	1.75	1.37	1.86	574	28	42	1.04	1.13
I0132a	x00120785a	1	Year 6	No	1.72	1.34	1.83	571	29		1.00	
I0132b	x00120785b	1	Year 10	No	0.94	0.57	1.06	490		62		0.98
I0133	x00120774	1	Link	No	1.53	1.15	1.64	551	33	43	1.08	1.13
I0134a	x00119137a	1	Year 6	No	2.49	2.12	2.61	652	18		1.03	
I0134b	x00119137b	1	Year 10	No	2.06	1.68	2.17	606		40		0.99
I0135	x00119144	2	Link	No	1.06	0.68	1.17	502	42	67	0.95	0.87
I0136a	x00119145a	1	Year 6	No	3.78	3.41	3.90	787	8		1.10	
I0136b	x00119145b	1	Year 10	No	2.93	2.56	3.05	698		26		0.91
I0137a	x00130079a	1	Year 6	No	0.94	0.56	1.05	489	43		1.08	
I0137b	x00130079b	1	Year 10	No	1.46	1.08	1.57	543		51		1.14
I0138a	x00120809a	2	Year 6	No	2.62	2.24	2.73	665	19		0.96	
I0138b	x00120809b	2	Year 10	No	2.25	1.87	2.36	626		37		0.99
I0139	x00121411	1	Link	No	1.14	0.76	1.25	510	41	62	1.01	0.93
I0140	x00130077	2	Link	No	1.62	1.24	1.73	560	32	52	0.96	0.90
I0141	x00120703	1	Year 10	No	-0.44	-0.81	-0.32	345		83		0.97
I0142	x00120923	1	Year 10	No	-0.71	-1.09	-0.60	316		85		0.93
I0143	x00120723	1	Year 10	No	1.90	1.52	2.01	590		43		0.99
I0144	x00120624	1	Year 10	No	0.98	0.60	1.09	493		61		0.89
I0145	x00120678	1	Year 10	No	-1.49	-1.86	-1.37	235		91		0.99
I0146	x00121107	1	Year 10	No	-0.33	-0.70	-0.21	357		81		0.94
I0147	x00121109	1	Year 10	No	0.77	0.39	0.88	471		65		1.04
I0148	x00121111	1	Year 10	No	1.38	1.00	1.49	536		52		1.15
I0149	x00121113	1	Year 10	No	-1.04	-1.42	-0.93	282		89		0.96
I0150	x00121115	1	Year 10	No	-0.58	-0.95	-0.46	330		84		1.01
I0151	x00120840	1	Year 10	No	0.59	0.21	0.70	453		68		0.97
I0152	x00120845	1	Year 10	No	0.20	-0.17	0.32	412		74		0.96

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0153	x00120848	1	Year 10	No	-0.23	-0.61	-0.12	367		80		1.01
I0154	x00121166	1	Year 10	No	0.15	-0.23	0.26	406		75		0.99
I0155	x00121174	1	Year 10	No	1.14	0.76	1.25	510		58		1.13
I0156	x00121177	1	Year 10	No	-1.16	-1.54	-1.05	269		89		0.95
I0157	x00122029	1	Year 10	No	-0.62	-1.00	-0.51	326		84		0.98
I0158	x00120857	1	Year 10	No	0.69	0.32	0.81	463		66		0.97
I0159	x00120879	1	Year 10	No	0.61	0.23	0.72	454		68		1.09
I0160	x00120970	1	Year 10	No	-0.30	-0.68	-0.19	360		79		0.93
I0161	x00121623	1	Year 10	No	0.56	0.19	0.68	450		66		1.01
I0162	x00121011	1	Year 10	No	1.36	0.99	1.48	534		53		1.03
I0163	x00121364	Deleted from item analysis due to poor fit										
I0164	x00120414	1	Year 10	No	1.12	0.75	1.24	509		57		0.95
I0165	x00120453	1	Year 10	No	-0.27	-0.64	-0.15	363		80		0.96
I0166	x00120512	1	Year 10	No	-1.08	-1.45	-0.96	278		88		0.96
I0167	x00120514	1	Year 10	No	0.87	0.50	0.99	482		62		0.92
I0168	x00120517	1	Year 10	No	1.61	1.24	1.73	560		47		0.99
I0169	x00120519	1	Year 10	No	0.12	-0.25	0.24	404		74		0.94
I0170	x00120520	1	Year 10	No	0.16	-0.22	0.27	408		74		0.97
I0171	x00120524	1	Year 10	No	0.79	0.41	0.90	474		64		1.07
I0172	x00121121	1	Year 10	No	-0.17	-0.55	-0.06	373		80		1.06
I0173	x00121122	1	Year 10	No	1.25	0.88	1.37	522		55		1.03
I0174	x00121125	1	Year 10	No	-0.64	-1.01	-0.52	324		85		1.00
I0175	x00121126	1	Year 10	No	-0.19	-0.56	-0.07	371		80		1.05
I0176	x00120754	1	Year 10	No	0.33	-0.05	0.44	425		71		0.92
I0177	x00120756	1	Year 10	No	1.60	1.23	1.72	559		47		1.21
I0178	x00120755	1	Year 10	No	1.18	0.80	1.29	515		57		0.98
I0179	x00120760	1	Year 10	No	1.86	1.49	1.98	586		43		0.92
I0180	x00120762	1	Year 10	No	0.92	0.55	1.04	488		62		1.16
I0181	x00120986	1	Year 10	No	1.38	1.01	1.50	536		53		1.10
I0182	x00120987	1	Year 10	No	-0.21	-0.59	-0.10	369		81		1.04
I0183	x00120985	1	Year 10	No	0.95	0.58	1.07	491		62		1.02
I0184	x00120988	1	Year 10	No	-0.26	-0.64	-0.15	363		81		1.03
I0185	x00120786	1	Year 10	No	1.06	0.68	1.17	502		59		1.01
I0186	x00120784	1	Year 10	No	-0.94	-1.32	-0.83	292		87		0.98
I0187	x00121412	1	Year 10	No	-2.65	-3.02	-2.53	114		96		1.24
I0188	x00121094	1	Year 10	No	2.19	1.82	2.31	621		37		1.05
I0189	x00121095	1	Year 10	No	0.85	0.47	0.96	479		64		1.11
I0190	x00121052	1	Year 10	No	-1.00	-1.37	-0.88	287		89		0.99
I0191	x00120714	1	Year 10	No	0.21	-0.16	0.33	413		75		0.97
I0192	x00121632	1	Year 10	No	1.10	0.72	1.21	506		59		1.00
I0193	x00120733	1	Year 10	No	3.87	3.49	3.98	796		13		1.04

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0194	x00120734	4	Year 10	No	2.94	2.56	3.05	699		20		1.07
I0195	x00120671	1	Year 10	No	1.30	0.92	1.41	527		55		0.96
I0196	x00120652	1	Year 10	No	1.92	1.55	2.04	592		43		0.95
I0197	x00120676	1	Year 10	No	1.51	1.14	1.63	550		51		1.02
I0198	x00121108	1	Year 10	No	-0.33	-0.70	-0.21	357		81		0.96
I0199	x00121117	1	Year 10	No	2.81	2.43	2.92	685		24		1.07
I0200	x00120850	1	Year 10	No	1.80	1.42	1.91	579		44		1.12
I0201	x00120846	1	Year 10	No	0.19	-0.19	0.30	410		76		1.07
I0202	x00121170	1	Year 10	No	1.02	0.64	1.13	498		61		0.90
I0203	x00121178	1	Year 10	No	1.26	0.88	1.37	522		55		1.19
I0204	x00120583	1	Year 10	No	-0.76	-1.14	-0.65	311		86		0.95
I0205	x00120585	1	Year 10	No	-1.47	-1.85	-1.36	237		91		0.99
I0206	x00120589	1	Year 10	No	-1.33	-1.71	-1.22	251		90		0.97
I0207	x00120596	2	Year 10	No	1.33	0.96	1.45	531		55		1.18
I0208	x00120860	1	Year 10	No	3.06	2.68	3.17	711		23		0.94
I0209	x00120869	1	Year 10	No	3.26	2.88	3.37	732		21		0.95
I0210	x00120878	1	Year 10	No	0.79	0.41	0.90	473		64		0.96
I0211	x00119122	1	Year 10	No	2.45	2.07	2.56	648		34		0.87
I0212	x00120940	1	Year 10	No	1.64	1.26	1.75	562		46		1.05
I0213	x00120956	1	Year 10	No	-0.68	-1.06	-0.57	319		83		0.95
I0214	x00120957	1	Year 10	No	2.79	2.42	2.91	684		25		1.02
I0215	x00120965	1	Year 10	No	2.17	1.79	2.28	618		37		0.93
I0216	x00121598	6	Year 10	No	2.13	1.75	2.24	614		36		1.30
I0217	x00121009	1	Year 10	No	2.93	2.55	3.04	697		24		1.03
I0218	x00121015	1	Year 10	No	2.31	1.94	2.43	633		33		1.15
I0219	x00121037	1	Year 10	No	2.43	2.05	2.54	645		33		0.96
I0220	x00120408	1	Year 10	No	2.03	1.66	2.15	604		40		0.98
I0221	x00120536	1	Year 10	No	2.78	2.40	2.89	682		24		1.12
I0222	x00121127	1	Year 10	No	2.93	2.56	3.05	698		23		1.03
I0223	x00121128	1	Year 10	No	2.40	2.02	2.51	642		34		0.95
I0224	x00120757	1	Year 10	No	0.26	-0.11	0.38	418		72		0.96
I0225	x00120758	1	Year 10	No	0.97	0.59	1.08	492		61		1.05
I0226	x00120977	1	Year 10	No	0.57	0.19	0.68	450		69		1.18
I0227	x00120777	1	Year 10	No	2.01	1.63	2.12	601		42		0.84
I0228	x00121083	1	Year 10	No	-0.89	-1.26	-0.77	298		88		0.96
I0229	x00121097	1	Year 10	No	-0.07	-0.45	0.04	383		79		1.02
I0230	x00121099	1	Year 10	No	3.49	3.11	3.60	756		15		1.03
I0231	x00121048	1	Year 10	No	-0.97	-1.35	-0.86	289		88		0.99
I0232	x00121051	1	Year 10	No	-0.16	-0.54	-0.05	374		80		0.95
I0233	x00121056	1	Year 10	No	1.30	0.93	1.42	527		56		0.99
I0234	x00121057	1	Year 10	No	3.69	3.31	3.80	777		15		1.02

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty from 2018 free calibration	Difficulty on historical scale and mode effect adjustment			% Correct Year 6	% Correct Year 10	Weighted fit (MNSQ) Year 6	Weighted fit (MNSQ) Year 10
						RP=0.50	RP=0.62	SL scale				
I0235	x00130060	2	Year 10	No	1.32	0.94	1.43	529		55		0.95
I0236	x00120606	3	Year 10	No	1.69	1.31	1.80	568		46		1.26
I0237	x00121613	2	Year 10	No	1.89	1.51	2.00	589		42		1.02
I0238	x00119199	3	Year 10	No	1.86	1.48	1.97	585		45		1.01
I0239	x00120954	3	Year 10	No	3.10	2.72	3.21	716		28		0.92
I0240	x00120504	2	Year 10	No	2.84	2.46	2.95	688		23		1.03
I0241	x00120540	3	Year 10	No	2.79	2.41	2.90	683		30		0.91

Item code	Item name	Scores	Vertical link	Horizontal link	Difficulty on historical scale and mode effect adjustment													
					Mean		Threshold 1		Threshold 2		Threshold 3		Threshold 4		Threshold 5		Threshold 6	
					RP=0.50	SL scale	RP=0.50	SL scale	RP=0.50	SL scale	RP=0.50	SL scale	RP=0.50	SL scale	RP=0.50	SL scale	RP=0.50	SL scale
I0053	x00120902	2	Year 6	No	0.33	465	-0.04	426	0.71	504								
I0079	x00121593	5	Year 6	No	0.71	505	-0.64	363	-1.49	274	1.00	535	3.20	766	1.49	587		
I0083	x00130154	2	Year 6	Yes	-0.08	422	-0.41	388	0.24	456								
I0086	x00130176	2	Year 6	Yes	0.72	506	-0.34	395	1.78	617								
I0088	x00121196	3	Year 6	No	-0.44	384	-3.35	79	0.96	531	1.06	542						
I0089	x00119198	2	Year 6	No	0.06	437	-0.31	398	0.43	475								
I0092	x00119016	2	Year 6	No	2.53	695	1.30	566	3.76	824								
I0105	x00000207	2	Year 6	Yes	1.83	622	1.57	595	2.08	648								
I0126	x00121454	2	Link	No	0.92	527	0.60	493	1.25	561								
I0135	x00119144	2	Link	No	0.68	502	0.33	465	1.03	539								
I0138a	x00120809a	2	Year 6	No	2.24	665	1.22	558	3.26	772								
I0138b	x00120809b	2	Year 10	No	1.87	626	0.79	513	2.96	740								
I0140	x00130077	2	Link	No	1.24	560	0.73	507	1.75	614								
I0194	x00120734	4	Year 10	No	2.56	699	1.81	620	2.01	641	3.26	772	3.17	762				
I0207	x00120596	2	Year 10	No	0.96	531	0.58	491	1.33	570								
I0216	x00121598	6	Year 10	No	1.75	614	-0.11	418	1.24	560	1.63	601	1.74	612	2.23	664	3.78	826
I0235	x00130060	2	Year 10	No	0.94	529	0.43	476	1.45	582								
I0236	x00120606	3	Year 10	No	1.31	568	0.56	489	1.54	591	1.85	624						
I0237	x00121613	2	Year 10	No	1.51	589	1.27	563	1.76	615								
I0238	x00119199	3	Year 10	No	1.48	585	0.62	495	1.02	537	2.80	724						
I0239	x00120954	3	Year 10	No	2.72	716	0.95	530	2.12	653	5.09	964						
I0240	x00120504	2	Year 10	No	2.46	688	2.16	657	2.76	719								
I0241	x00120540	3	Year 10	No	2.41	683	1.05	540	1.93	633	4.25	876						

APPENDIX 6 PROFICIENCY LEVEL DESCRIPTIONS

Proficiency level	Level descriptors
Level 5 or above	Explains interactions that have been observed in terms of an abstract science concept. Summarises conclusions and explains the patterns in the data in the form of a rule and are consistent with the data. When provided with an experimental design involving multiple variables, can identify the questions being investigated.
Level 4	Applies knowledge of relationship between variables to explain a reported phenomenon. Extrapolates from an observed pattern to describe an expected outcome or event. Demonstrates awareness of the principles of conducting an experiment and controlling variables.
Level 3	Interprets information in a contextualised report by application of relevant science knowledge. Interprets data and identifies patterns in – and/or relationships between – elements of the data. Collates and compares data set of collected information. Gives reason for controlling a single variable.
Level 2	Selects appropriate reason to explain reported observation related to personal experience. Interprets simple data set requiring an element of comparison. Makes simple standard measurements and records data as descriptions.
Level 1 or below	Describes a choice for a situation based on a first-hand concrete experience, requiring an application of limited knowledge. Identifies simple patterns in the data and/or interprets a data set containing some interrelated elements. Makes measurements or comparisons involving information or stimulus in a familiar context.

APPENDIX 7 SUMMARY OF MODE EFFECT RESULTS FOR NAP-SL 2018

A mode-effect study was designed to investigate the effect of the change in delivery mode from a paper-based to a computer- based assessment in the NAP–SL context. The outcome of this study was intended to inform 1) comparability of online results in 2018 and 2) the effort needed to place the results of the online 2018 NAP–SL onto the historical scale. Forty schools from Australian Capital Territory, New South Wales, Queensland, South Australia, Tasmania, Victoria and Western Australia were selected to participate in the study. In each school, approximately 20 to 25 students participated in each school.

The mode effect test contained 36 items, divided into two parts: Part A and Part B. Part A was the first half of the test; Part B was the second half of the test. Each part had a paper and an online version. Schools were randomly assigned into two groups. Group 1 (n=397) sat Part A on computer and Part B on paper while Group 2 (n=366) took Part A on paper and Part B on computer.

The Rasch measurement model, using ACER ConQuest, was applied to calibrate items, perform DIF analysis and investigate the impact of mode effect at both test and item levels. Table A7.1 summarises the design of the mode effect study.

Table A7.1 Design of the mode effect study

	Test Part A	**Test Part B**
Group 1	Online	Paper
Group 2	Paper	Online

Item performance

First differential item functional (DIF) was examined for each part between the computer and paper-based version of the items. None of the items showed substantial DIF and therefore all items were treated as link items between modes. A scatterplot of relative item difficulties for Part A and Part B between the paper and computer version of the items is presented in Figure A7.1 and in Figure A7.1. The broken line is the identity line, which is the expected location for each of the items.

Figure A7.1 Evaluation of Part A link items

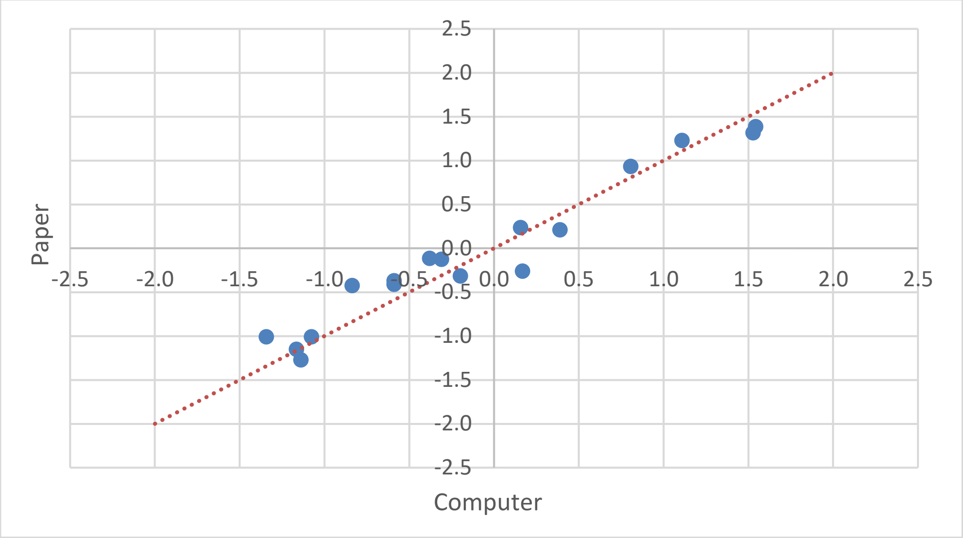
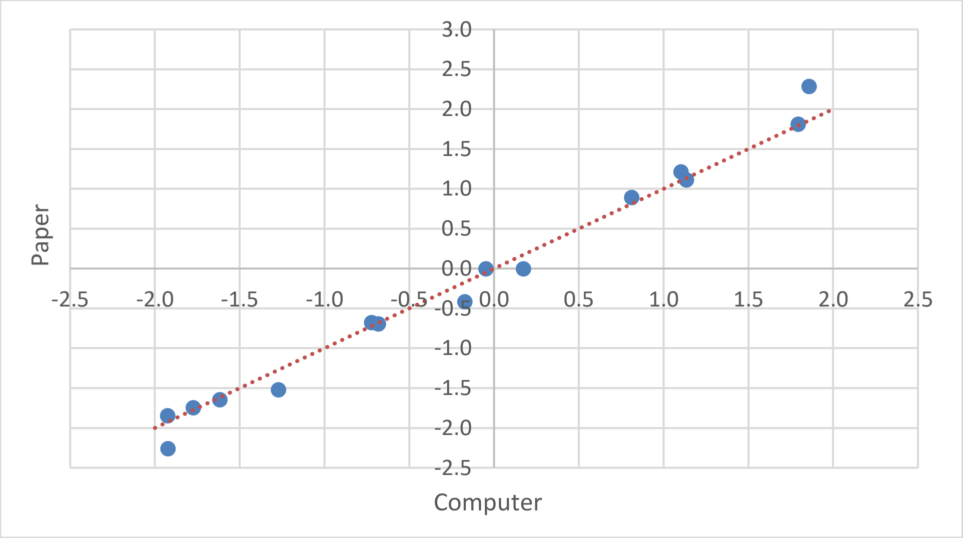


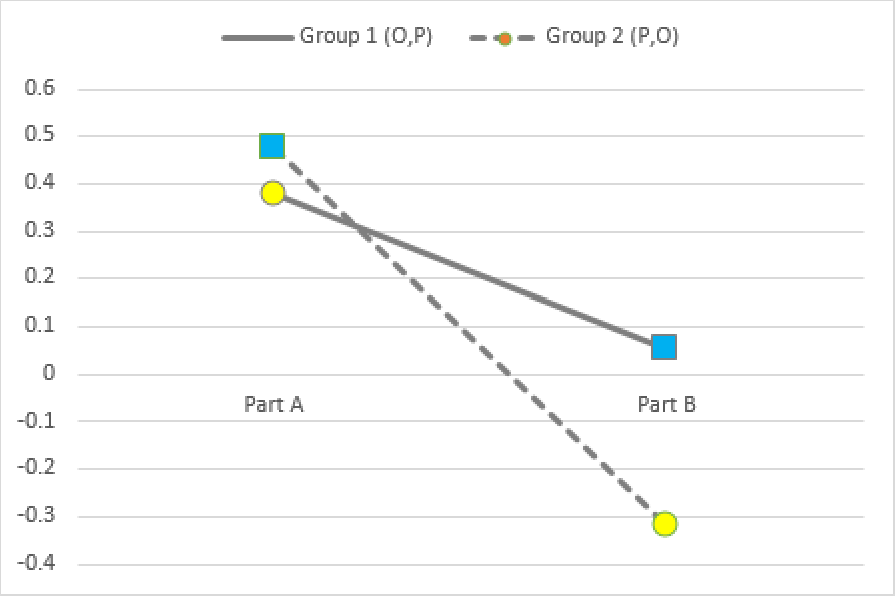
Figure A7.2 Evaluation of Part B link items



Student performance

The student results showed that they performed higher on paper than online, which is consistent with the finding in 2015. The difference was smaller for Part A than for Part B, see Figure A7.3.

Figure A7.3 Results of the mode effect study



The difference between the groups in Part A of the test was **0.098** of a logit (about one tenth of a standard deviation); the difference between the groups in Part B was **0.369** of a logit (about one third of a standard deviation). The average mode effect of Part A and B together was **0.234** ((0.098+0.369)/2).

The size of the difference in Part B of the test and the difference in the size of the mode effect between Part A and Part B are larger than expected. The size of the mode effect appears more realistic for Part A. Items in Parts A and B were visually inspected on paper and on the screen. No obvious differences were observed between the parts or mode. Further review of differences between Part A and Part B revealed that Part A consisted of 18 per cent constructed response items and Part B consisted of 67 per cent constructed response items.

These percentages were then compared with the distribution of item types in the NAP-SL 2018 Main Study test for Year 6. The results in Table A7.2 show that 24 per cent of the items in the test were constructed response items. This percentage was closer to the percentage in Part A (18%) of the mode effect test than Part B (67%).

Table A7.2 Distribution of item types in NAP-SL 2018 MS test for Year 6

Item type	Number	Proportion
Constructed Response	33	0.24
Multiple choice	77	0.55
Interactive	22	0.16
Other	8	0.06
Total	140	1.00

Table A7.3 shows the national and jurisdictional trends when applying the mode-effect as an additional shift to the equating process, placing the 2018 results onto the historical scale. The national mean did not significantly change between 2015 and 2018 and neither did the jurisdictional means, except for Queensland which showed an increase in performance of 28 NAP score points.

Table A7.3 Preliminary trends with interpolated mode effect adjustment (+0.131)

	Year 10		Year 6 (2018)		Year 6 (2015)		Difference (Year 6)		
	Mean	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Z-score**
NSW	486	6.0	397	5.4	411	4.4	-14	8.1	-1.68
Vic.	487	7.8	405	5.3	399	4.5	6	8.1	0.73
QLD	489	10.1	426	4.3	398	5.4	**28**	8.1	**3.45**
WA	515	9.5	415	7.4	408	3.8	7	9.3	0.78
SA	471	13.7	400	7.9	392	4.5	8	10.0	0.84
Tas.	483	28.3	405	7.6	414	6.0	-9	10.6	-0.83
ACT	545	12.8	427	9.0	414	6.2	13	11.7	1.07
NT	449	13.2	302	20.0	320	0.1	-18	20.4	-0.88
Aust.	490	3.7	407	2.5	403	2.2	4	5.4	0.80

It is recommended to shift the 2018 results up by 0.131 of a logit to correct for the effect the switch from paper-based testing to computer-based testing had on student performance.