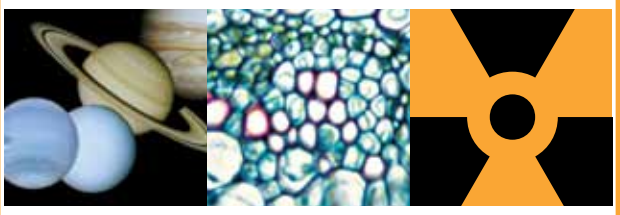


National Assessment Program – Science Literacy Technical Report

2012

acara
AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

NAP–SL 2012 Project Staff

Dr Sofia Kesidou from Educational Assessment Australia (EAA) was the Project Director of NAP–SL 2012. Jennifer Cowing (EAA) was the Project Manager. The test development team was led by Joe Merlino (EAA). The Public Report was written by Dr Sofia Kesidou, Dr Rassoul Sadeghi and Nicholas Marosszeky. The School Release Materials were written by Dr Sofia Kesidou, Jennifer Cowing, Joe Merlino and Dr Rassoul Sadeghi (EAA).

The sampling and data analysis tasks were undertaken by Dr Rassoul Sadeghi, Nicholas Marosszeky and Dr Edward Li from EAA and Dr Margaret Wu and Mark Dulhunty from Educational Measurement Solutions (EMS). The Technical Report was written by Dr Rassoul Sadeghi, Nicholas Marosszeky and Dr Sofia Kesidou from EAA and Dr Margaret Wu and Mark Dulhunty from EMS.

Cover Photographs
Second and third from top: courtesy of Auburn West Public School and Our Lady of Fatima, Catholic Primary School, © Educational Assessment Australia

ACARA Copyright Administration, ACARA
Level 10, 255 Pitt Street
Sydney
NSW 2000
Email: info@acara.edu.au

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

Contents

List of Tables

List of Figures

Chapter 1
2012 National Assessment Program – Science Literacy: Overview

1.1 Introduction

In July 2001, the then Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA, now the Standing Council on School Education and Early Childhood (SCSEEC)) agreed to the development of assessment instruments and key performance measures for reporting on student skills, knowledge and understandings in primary science. It directed the newly established Performance Measurement and Reporting Taskforce (PMRT), a nationally representative body, to undertake the national assessment program. The PMRT commissioned the assessment in July 2001 for implementation in 2003. The Primary Science Assessment Program (PSAP) – as it was then known – tested a sample of Year 6 students in all states and territories. The second cycle of the assessment was conducted as the National Assessment Program – Science Literacy (NAP–SL) in October 2006, the third cycle was conducted in October 2009 and the fourth cycle was conducted in October 2012.

NAP–SL was the first assessment program designed specifically to provide information about performance against the National Goals for Schooling in the Twenty-First Century (now the Educational Goals for Young Australians). Subsequently, Ministers for Education also endorsed similar sample assessment programs to be conducted for Civics and Citizenship (CC) and Information and Communications Technology Literacy (ICTL). Each sample assessment program is repeated every three years so that performance in these areas of study can be monitored over time. The first cycle of each program was intended to provide the baseline data against which future performance could be compared.

In January 2011, the Australian Curriculum, Assessment and Reporting Authority (ACARA) awarded the contract for the fourth cycle of NAP–SL to Educational Assessment Australia (EAA). This report provides detailed information on test development, sampling, item analysis, equating and scaling of the fourth cycle of the science literacy assessment conducted in 2012.

1.2 Purposes of the Technical Report

This Technical Report aims to provide detailed information with regard to the conduct of the 2012 NAP–SL assessment so that valid interpretations of the 2012 results can be made, and future cycles can be implemented with appropriate linking information from past cycles. Further, a fully documented set of NAP–SL procedures can also provide information for researchers who are planning assessments of this kind. The methodologies used in the 2012 NAP–SL assessment can inform researchers of the current developments in large-scale assessments. They can also highlight the limitations and suggest possible improvements in the future. Consequently, it is of great importance to provide technical details on all aspects of the assessment.

1.3 Organisation of the Technical Report

This report is divided into nine chapters.

Chapter 2 provides an outline of the test development and test design processes, including trialling and item selection, and the assessment domain of scientific literacy.

The sampling procedures across jurisdictions, schools and classes are discussed in Chapter 3.

Chapter 4 includes information about how the tests were administered and marked, including coding for student demographic data and participation or non-inclusion. It also provides an explanation of the reporting processes.

Chapter 5 details the processes involved in computing the sampling weights.

Chapter 6 details the processes undertaken to analyse data obtained from the final test.

Chapter 7 provides an outline of the scaling procedures followed as part of the data analysis.

The equating procedures which were followed so that the 2012 results could be reported against the baseline established in 2006 are discussed in Chapter 8.

Chapter 9 provides a brief overview of the cut-off points at each proficiency level and information on the performance of the items on the proficiency scale.

Appendices 1 – 8 provide further elaboration and exemplification of the information in the body of the Technical Report.

Chapter 2
Test Development and Test Design

2.1 Assessment domain

The National Assessment Program – Science Literacy (NAP–SL) measures scientific literacy. This is the application of broad conceptual understandings of science to make sense of the world, understand natural phenomena and interpret media reports about scientific issues. It also includes asking investigable questions, conducting investigations, collecting and interpreting data and making decisions. The construct evolved from the definition of scientific literacy used by the Organisation for Economic Co-operation and Development (OECD) – Programme for International Student Assessment (PISA):

> *... the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity.*
>
> (OECD 1999, p. 60)

This definition has been adopted for the purpose of monitoring primary science in NAP–SL (Ball et al. 2000). The science items and instruments assess outcomes that contribute to scientific literacy, including conceptual understandings, rather than focusing solely on scientific knowledge. They also assess student competence in carrying out investigations in realistic situations.

A scientific literacy Progress Map (see Appendix 1) was developed in the first assessment cycle based on this construct of scientific literacy and on an analysis of the state and territory curriculum and assessment frameworks. The Progress Map describes the development of scientific literacy across three strands of knowledge which are inclusive of Ball et al.'s concepts and processes and the elements of the OECD–PISA definition.

As in the previous three cycles of NAP–SL, three main areas of scientific literacy were assessed in 2012:

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.

In addition, the items drew on four major scientific concept areas: Earth and Space; Energy and Force; Living Things; and Matter. These concept areas, found most widely in state and territory curriculum documents, were used by item developers to guide item and test development. The list of endorsed examples for each of these concept areas is in Table A1.2 of Appendix 1.

A conscious effort was made to develop assessment items that related to everyday contexts. The intention was to ensure that all Year 6 students were familiar with the materials and experiences to be used in NAP–SL and so avoid any systematic bias in the instruments being developed.

2.2 Test blueprint

In response to the Australian Curriculum, Assessment and Reporting Authority's (ACARA's) Request for Tender for the 2012 NAP–SL assessment, Educational Assessment Australia (EAA) proposed the following assessment specifications:

> *It is anticipated that the 2012 final test forms will contain approximately 110 items in total (including link items from 2003, 2006 and 2009) providing sufficient assessment items for up to two hours of testing for each student in the national sample. This number of items will also provide items to form part of the School Release Materials for subsequent teacher use and items to be held secure for 2015.*
>
> *... Three types of items will be developed: multiple-choice items, short constructed response items (requiring one or two word responses from students); and extended constructed response items requiring students to provide an extended response. For Year 6 students an extended response might reasonably be expected to be of the order of one or two sentences – up to a short paragraph – if in text form or a diagram or constructed data table of equivalent detail.*
>
> *Consistent with previous assessments, the balance of item types within the trial item pool is proposed to be: 50% multiple-choice; 10% short constructed response: 40% extended constructed response. This balance is*

proposed on the basis that it is acknowledged that Year 6 students may be reluctant to provide overly lengthy written explanations to test questions. However, in order to assess the higher order skills demanded by upper levels of the framework it will be necessary to include some extended response items.

Due to the contextualised nature of the pencil-and-paper item sets and practical tasks, it is expected that the majority of item sets will contain a mix of item types.

These specifications were approved at the first meeting of the Science Literacy Review Committee (SLRC). In addition, it was confirmed that the balance between process items (Strands A and B) and conceptual items (Strand C) would be approximately in the proportion 50 per cent process and 50 per cent conceptual items.

2.2.1 Test design

In order to cover a wide range of content areas in science, but at the same time not to place too much burden on each student, the Balanced Incomplete Block rotation design was implemented. A rotation design allows a greater number of items to be assessed by using several booklets with different items rotated across them. It minimises the effect of biased item parameters caused by varying item positions arising from the placement of an item in a test booklet. Items were placed in 'clusters' and the clusters were rotated through the test forms, each appearing three times, each time in a different location ('block') in the test form. Seven test forms were developed for the final assessment. Table 2.1 shows the rotation design used in the 2012 NAP–SL final assessment.

Table 2.1 Rotation design used in the 2012 NAP–SL final assessment

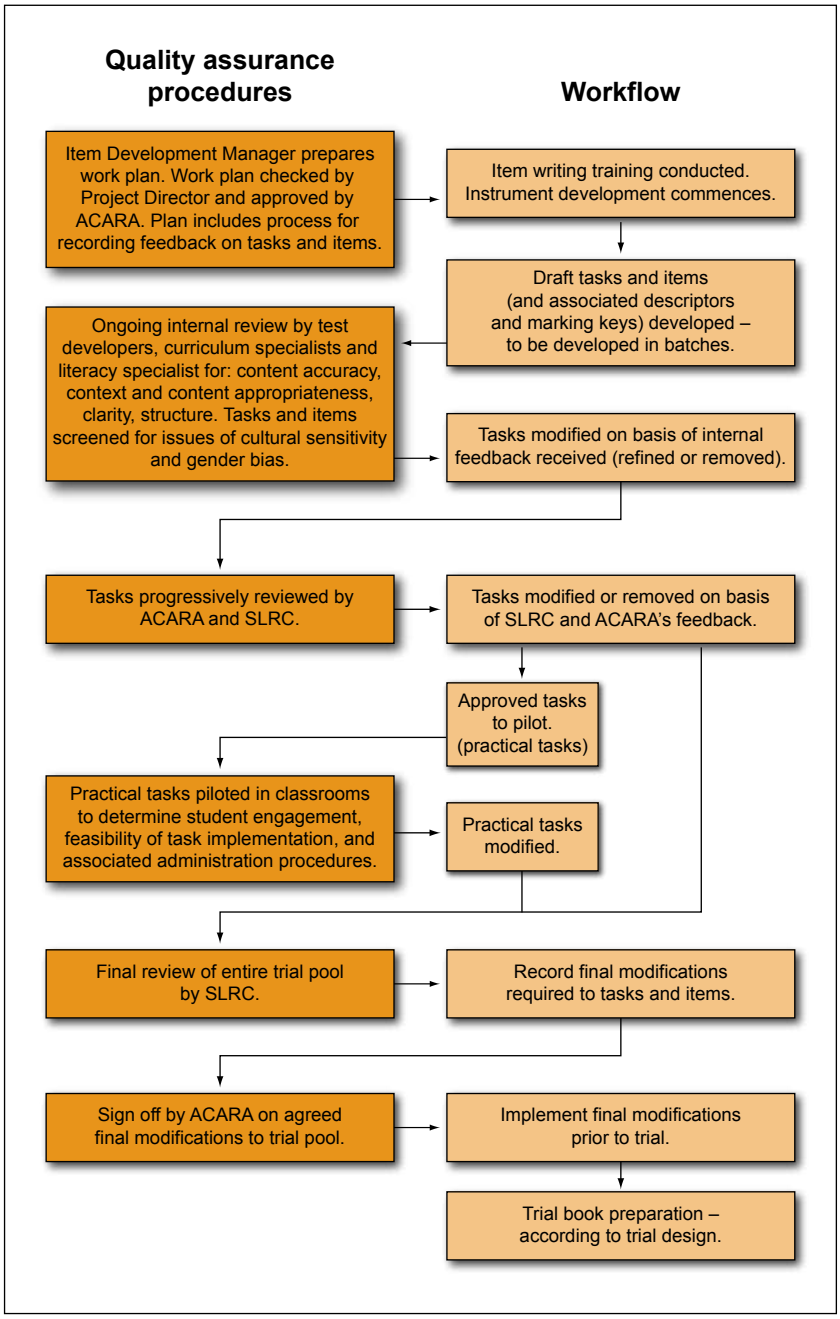
Booklet	Block 1	Block 2	Block 3
1	Cluster 1	Cluster 2	Cluster 4
2	Cluster 2	Cluster 3	Cluster 5
3	Cluster 3	Cluster 4	Cluster 6
4	Cluster 4	Cluster 5	Cluster 7
5	Cluster 5	Cluster 6	Cluster 1
6	Cluster 6	Cluster 7	Cluster 2
7	Cluster 7	Cluster 1	Cluster 3

2.3 Item development process

2.3.1 Item development

Item development was undertaken by EAA. A process was developed to facilitate item writing in prescribed batches. The following flow chart (Figure 2.1) outlines the workflow and associated quality assurance procedures implemented at each stage of test development. As illustrated, the significant and explicit involvement of the SLRC was essential for the acceptance of items for trial. The progressive review of batches of items allowed many opportunities for the SLRC to provide input to the test development process.

Figure 2.1 Test development work flow and quality assurance procedures



Draft marking guides and item descriptors (identifying item demands by reference to the strands and levels of the assessment domain) were developed at the same time as the items themselves, and were reviewed accordingly.

EAA held review panels in-house prior to releasing materials for ACARA and SLRC reviews. Items and draft marking guides were reviewed for content accuracy, context and literacy demand. The purpose of the literacy demand review was to ensure that the language used in the items would be accessible to all students and that the use of unfamiliar and difficult vocabulary would be avoided, except where such use was necessary for subject-specific outcomes.

All developed test materials were reviewed by SLRC members via the Online Test Item Collection and Review System (OTICRS), a secure reviewing application developed by ACARA. Items were released progressively in batches between May and August 2011. Specific criteria were developed to guide the SLRC review (Table 2.2). SLRC members were asked to judge each item against the criteria and justify their judgments. Procedures were also established for recording and responding to feedback on tasks and items as the review processes proceeded. In addition, associated documentation was prepared including unit templates and a spreadsheet tracking each batch of items against the assessment specifications.

The OTICRS system allowed SLRC members to examine each item, provide detailed feedback and then recommend whether the item should progress with or without modification, or whether the item should be removed. All feedback was then collated and responded to by the test development team. The refined items were released to ACARA for sign off prior to the trial.

Table 2.2 Criteria used in SLRC item reviews

Criterion	Explanation
Concept, Strand, Level	Concept, strand and level are appropriate.
Descriptor	Descriptor is appropriate.
Key/Scoring guide, Distractors, Scientific accuracy	Only one option is the possible answer. All distractors are plausible. Scientific content of the stimulus is accurate.
Language, Numeracy, Clarity	The language demand (e.g. sentence length and structure, word familiarity, etc) in the item is appropriate. The question or task is clearly stated, the wording in stem and options is clear and concise. The mathematics knowledge/skills needed to respond to the item are appropriate.
Interdependence	The item is not dependent on any other item in this item set.
Graphics	Graphics are appropriate.
Cultural appropriateness (gender, race, geopolitical sensitivities)	The content of the stimulus is culturally appropriate in the context of your jurisdiction/ sector.
Sequencing	Sequencing of the item is appropriate.
Engagement	Level of engagement is sufficient.
General comments	Comments about this item that are not covered by the criteria

2.3.2 Pilot studies

Each practical task was piloted with at least two classes of Year 6 students to ensure that the activities proposed and the associated administration procedures could be implemented with ease in Year 6 classroom settings. A total of five schools participated in the pilot study, with multiple classes in each school. The pilot study also established the degree to which the proposed tasks were engaging for students. All the materials required to carry out the practical tasks were relatively simple in nature and were provided to schools by EAA. Simple materials were used in order to ensure that students would not be disadvantaged by their potentially limited familiarity with specialist science equipment which is more likely to be found in secondary school laboratories. After students completed the practical tasks, EAA staff conducted whole classroom discussions to assess whether the items were clear to students and to obtain insight into the cognitive processes that students engaged when responding to the items. Two EAA staff members experienced in cognitive interviewing conducted the debriefing for each practical task session, with one staff member asking students questions and the other taking notes. In all classes observed, students completed the tasks and associated questions in the allocated time. In all practical tasks, students had no difficulty manipulating the materials. Students found the practical tasks and the debriefing sessions engaging, as evidenced by their focus while conducting the activities and their enthusiasm answering the interviewers' questions. Instructions and items in the practical tasks were modified based on the pilot findings.

2.3.3 Items delivered

A total of 221 items were released for review prior to the trial, including 21 link items from the 2009 NAP–SL assessment. The final pool of trial items was approved by ACARA and the SLRC in August 2011.

Table 2.3 Composition of the trial item pool (all released batches)

	Pencil-and-paper items	Practical task items	Released total pool
Major concept area: ES	40	0	40
Major concept area: EF	42	29	71
Major concept area: LT	47	14	61
Major concept area: M	35	14	49
Total	164	57	221
Strand A	25	14	39 (18%)
Strand B	67	35	102 (46%)
Strand C	72	8	80 (36%)
Total	164	57	221
Level 2	22	7	29 (13%)
Level 3	109	38	147 (67%)
Level 4	33	12	45 (20%)
Total	164	57	221
Multiple choice	69	9	78 (35%)
Short answer	22	9	31 (14%)
Extended response	73	39	112 (51%)
Total	164	57	221

EAA developed eight trial test forms comprising 164 objective items and four trial practical tasks. The items were placed into clusters that were arranged into the trial forms so that each cluster appeared twice. The trial forms contained one cluster (cluster 8) comprising link items drawn from the secure item pool from 2009.

2.3.4 Student Survey

A student survey instrument was developed for trialling that was guided by the items and results of the 2009 Student Survey as well as by recommendations from the SLRC members. Survey items with scoring parameters were submitted to and reviewed by the SLRC and ACARA. Following feedback, 42 items were selected for trial. The areas covered by the trial student survey were:

- students' interest in science
- students' self-concept in science
- students' perceived value of science
- students' perceptions of science
- frequency of science-related activities outside school
- frequency of science-related activities at school
- students' experiences related to science teaching and investigations
- science topics studied at school.

The trial survey was produced on a scannable form. It was conducted following the administration of the trial objective and practical tasks.

2.4 Field trial of test items and Student Survey

A convenience sample of 30 schools across ACT, NSW, QLD, SA and WA participated in the trial in October 2011. The trial schools were selected to reflect the range of educational contexts around the nation and included schools from government, Catholic and independent sectors; low and high socioeconomic areas; metropolitan and regional locations; large and small schools; and students from a variety of language backgrounds.

In total, 1057 students from the trial schools across the five selected states and territories participated in the trial. Each student completed one of the eight trial objective test forms and one of the four practical tasks. Within each class teachers were asked to evenly distribute the eight trial objective test forms amongst students. On completion of the trial objective test forms students within a class were asked to separate into groups of three (or groups of two where necessary) for completion of the practical task. Students within the same class completed the same practical task.

Classroom teachers were provided with a Test Administrator's Manual in advance of the trial to allow them to familiarise themselves with the test administration procedures. A trained invigilator was sent to each trial school to deliver and collect the trial assessment materials (to ensure the security of the materials) and to also observe and support the classroom teacher throughout the assessment and Student Survey. At the completion of each assessment and Student Survey session the invigilator and the classroom teacher each completed a session report form to provide feedback about various aspects of the trial administration. This feedback, in conjunction with a range of other sources of feedback, informed the selection and refinement of items for the final pool of assessment items and for the final student survey.

2.4.1 Marking process

A team of experienced markers was engaged for a one-week period. The Test Development Manager and Project Director trained the markers and remained on-site to oversee the marking process. On completion of marking each cluster of items or practical task, a debriefing session was held with the test developers, amendments were made to the marking guides as necessary, and illustrative correct and incorrect responses obtained from the trial assessment were added to the marking guides.

2.4.2 Data analysis

The trial scores were data entered and then analysed by EAA's data analysis team using both ConQuest and RUMM software. The data were also sent to an external contractor, Educational Measurement Solutions (EMS), for parallel processing. The results of the parallel analyses were identical.

Key criteria for judging the performance of items were measures of item fit statistics (weighted MNSQ) and item performance illustrated by Item Characteristic Curves (ICCs). Percentage correct and point-biserial correlation were noted, but only informed decisions to eliminate items if other indices were poor. Ten items were removed due to poor fit statistics. The analysis was then repeated using the remaining 211 items (second analysis).

Differential Item Functioning analyses (DIF)[1] for gender and language background (LBOTE) were carried out for all remaining items. However, DIF analysis results for LBOTE could not be considered due to the small sample size and the lack of information about specific language backgrounds provided by students who participated in the trial.

1 By definition, DIF refers to groups of students responding to an item differently, after adjusting for the groups' overall abilities. For example, if a group of boys and a group of girls have the same mean ability, but the probability of success on an item for the girls is higher (or lower) than the probability of success for the boys, then the item exhibits gender DIF. DIF does not refer to the difference in raw percentages correct for the groups, since these differences could be due to the fact that the groups have varying abilities. In other words, DIF examines the performance of a group on an item relative to the group's performance on other items. In this respect, a study of DIF shows the relative differences in performance on items in one test. DIF does not show 'absolute' differences in performance between two groups of students.

The DIF analyses for gender were carried out using ConQuest by fitting a facets model, where the interaction between an item and the gender group is estimated. When the interaction term is significantly different from zero at the 95 per cent confidence level, an item is deemed as showing DIF. An additional criterion applied was that a difference in item difficulty between boys and girls should be larger than 0.4 logits before the item is deemed to show large gender DIF.

In cases where items exhibited large gender DIF, content experts inspected the reasons for the observed bias. The items were flagged but not automatically removed simply based on statistical evidence of bias. Items were discarded only where there was agreement between the psychometric evidence and the content experts' review[2]. Inclusion of these items in the final assessment pool was monitored to ensure there was no gender DIF imbalance in the final assessment.

The remaining 211 items were further investigated for reverse thresholds. Two further items were eliminated in this process.

Figure 2.2 (Item-person map produced using ConQuest) illustrates diagrammatically the distribution of the 211 items (indicated by item identifiers) that were part of the second analysis of the trial data. The figure illustrates diagrammatically the distribution of all trialled items. It provides 'at a glance' the range of difficulty of the items and how they align with the ability of the students in the trial pool. The left-hand side of the X axis represents persons, the right-hand side represents items and the Y axis represents the logit scale. Each 'X' represents 6.7 students.

As can be seen from Figure 2.2, the 2012 trial NAP–SL assessment achieved an excellent spread of item difficulties and was appropriately matched to the ability distribution of the Year 6 cohort tested. There were a number of items that all students found to be very easy, a number of items that were challenging (even for the most able students), and many items in the middle range. The average test difficulty in terms of item facility was between 50 per cent and 55 per cent.

Shown in red are the items that were eliminated because of reverse thresholds. Items highlighted in yellow (36 items) showed less optimal item fit (items with a weighted MNSQ smaller than 0.9 or greater than 1.1) or displayed large gender DIF. Items highlighted in green (173 items) were the best performing items (items with a weighted MNSQ larger than 0.9 or smaller than 1.1 and small or no gender DIF) and were given priority for inclusion in the final assessment item pool.

2 For example, it may well be the case that girls and boys do not perform in the same way across content areas in a subject domain, and such differential performance may be expected. Thus, judgments should be made based on the importance of the skills tested in the specific items, and whether the inclusion of items showing gender DIF will bias the results in ways that are not consistent with the aims of the assessment.

Figure 2.2 Item-person map for the 211 trial items in the second analysis

2.4.3 Reports to trial schools

Reports were developed and provided to schools that had participated in the trial. The reports were received in schools in December 2011. They contained a number of A4 sheets, one for each of the eight trial test booklets used in the assessment. Individual students' results were given for the test booklet which they completed in the assessment. In addition there was a school report for each of the practical tasks conducted by the school. An information sheet providing advice on interpreting the reports was also included.

2.5 Item selection process for the final test

2.5.1 Item selection for the objective and practical test

All trial items were provided to the SLRC to view on a refined version of OTICRS. Reviewers were invited to view the stimulus and item images as well as the associated metadata. These metadata included the key or marking guide and acceptable responses for constructed response items, and the following psychometric data obtained from the trial:

- facility (per cent correct)
- discrimination
- weighted MNSQ.

The results from the trial were discussed at a meeting with the SLRC in Sydney in March 2012. The pool of the 173 best performing items from the trial (see Figure 2.2) was approved for use in the 2012 assessment.

The following changes to the test administration were recommended (based on invigilator feedback) and agreed upon at the meeting:

- the teacher should read the introduction to the practical task to the students before distributing the practical task materials
- the students should be allocated reading time prior to starting the practical task.

In addition, where there were several consecutive items which had the same key, the SLRC recommended changing the order of distractors.

EAA developed a draft final list of preferred test items for 2012 based on feedback by the SLRC. The final pool of 112 test items was developed to reflect the best balance of items against the original assessment specifications. The final pool of test items was presented to ACARA and approved for use in the 2012 assessment. The pool included six link items from the 2003 assessment, 14 items from 2006, and 16 of the 21 items from the 2009 assessment that had been included in the 2012 trial.

2.5.2 Item selection for the Student Survey

Students' responses to the trialled survey questions were scanned and analysed. All items and results were presented to the SLRC in a secure document for their review. SLRC members were invited to comment on the items and prioritise their inclusion in the final Student Survey form. At the SLRC meeting in March 2012, members discussed the results and agreed on a final list of 34 survey items that would comprise the final Student Survey.

2.6 Characteristics of the final assessment instrument

The actual distribution of items across the assessment domain for scientific literacy (strands and major scientific concept areas) is shown in Table 2.4. There were 112 items distributed across the seven pencil-and-paper tests and two practical tasks. Each student had to sit one pencil-and-paper test and one practical task.

Table 2.4 Composition of the final item pool

Domain	Item type and number of items			
	Multiple choice	Short answer	Extended response	Total
Distribution of items by strand				
Strand A	8	0	9	17
Strand B	17	6	19	42
Strand C	27	5	21	53
Total	52	11	49	112
Distribution of items by major scientific concept area				
Earth and Space (ES)	13	2	5	20
Energy and Force (EF)	10	5	21	36
Living Things (LT)	17	1	11	29
Matter (M)	12	3	12	27
Total	52	11	49	112

The final composition of the items (112) included in the 2012 assessment instrument is shown in Tables 2.5 to 2.9.

Table 2.5 Breakdown of concept areas across the final objective and practical papers

Paper type	Concept area				Total
	ES	EF	LT	M	
Objective	25	20	19	27	91
Practical	11	0	10	0	21
Total	36	20	29	27	112

Table 2.6 Breakdown of strands across the final objective and practical papers

Paper type	Strand			Total
	A	B	C	
Objective	14	26	51	91
Practical	3	16	2	21
Total	17	42	53	112

Table 2.7 Breakdown of targeted levels across the final objective and practical papers

Paper type	Level			Total
	2 and below	3	4 and above	
Objective	11	56	24	91
Practical	2	16	3	21
Total	13	72	27	112

Table 2.8 Breakdown of item types across the final objective and practical papers

Paper type	Item type			Total
	Multiple choice	Short answer	Extended response	
Objective	46	10	35	91
Practical	6	1	14	21
Total	52	11	49	112

Table 2.9 Breakdown of logit scale location ranges (based on trial statistics) across the final objective and practical papers

Paper type	Logit scale location ranges												Total
	−2.5 to −2.0	−2.0 to −1.5	−1.5 to −1.0	−1.0 to −0.5	−0.5 to 0.0	0.0 to 0.5	0.5 to 1.0	1.0 to 1.5	1.5 to 2.0	2.0 to 2.5	2.5 to 3.0	3.0 to 4.0	
Objective	3	4	4	5	4	14	6	7	5	1	1	1	55
Practical	2	1	4	2	3	3	2	1	0	1	1	1	21
Total	5	5	8	7	7	17	8	8	5	2	2	2	76

Note: Link items are not included in this table.

2.7 Reports to schools

Reports were developed and provided to schools that participated in the 2012 assessment. Schools received the reports in December 2012. The reports contained seven A4 sheets, one for each of the seven test booklets used in the final assessment. Individual students' results were given for the test booklet which they had completed in the assessment. In addition, there was a school report for the practical task conducted by the school. An information sheet providing advice on interpreting the reports was also included.

A sample school report can be found in Appendix 2. The school report includes a report for each objective booklet and a report for the practical task 'Reaction time'.

Chapter 3
Sampling Procedures

3.1 Overview

The desired (target) population for the National Assessment Program – Science Literacy (NAP–SL) consisted of all students enrolled in Year 6 in Australian schools in 2012.

As defined in the project specifications, the number of students sampled in each jurisdiction was determined with the following considerations in mind.

It was desirable that the estimated mean scores for all jurisdictions were of similar precision. While this was an ultimate goal, it was recognised that, in practice, reduced sample sizes would be needed for the smaller jurisdictions (i.e. ACT, NT and TAS). This is because most schools in the smaller jurisdictions would need to participate to form a large enough sample. As there are a number of national and international assessment projects implemented in Australia, many schools from the smaller jurisdictions would need to participate in multiple assessment projects, consequently putting more administrative burden on these schools, particularly the smaller schools.

The nationwide achieved sample was to be approximately 13 000 students located within approximately 600 schools throughout Australia.

The sample design for NAP–SL was a two-stage stratified[1] cluster sample. Stage 1 consisted of selecting schools that had Year 6 students. In this stage, schools were selected with probabilities proportional to their measure of size[2]. This selection procedure is referred to as 'probability proportional to size' (PPS) sampling. Stage 2 involved the random selection of one Year 6 intact class (or for some schools two or three classes) from each of the sampled schools selected in Stage 1.

For the first time, NAPLAN data were also included as an implicit stratification variable in the 2012 assessment (see section 3.5 for further information).

1 Stratification involves ordering and grouping schools according to different school characteristics (e.g. state, sector, geolocation) which helps ensure adequate coverage of all desired school types in the sample.
2 The school measure of size is related to estimated enrolment size of Year 6 students at the school.

3.2 Target population

The operational definition of the target population was a sampling frame which consisted of a list of all Australian schools and their 2011 Year 6 enrolment sizes as supplied by the Australian Curriculum, Assessment and Reporting Authority (ACARA).

Generally, large scale sample assessments of this type include provisions for excluding schools before the sampling of schools takes place. Schools might be excluded for reasons such as the school being located in a geographically remote location or of extremely small size. This approach was taken in 2003. However in 2006, 2009 and 2012, it was deemed desirable to include as many schools in the defined population as possible. Essentially this meant there were no school-level exclusions from the supplied sampling frame prior to sample selection. As such, the nationally defined population for the 2012 NAP–SL assessment was more inclusive than the 2003 defined population. However, the inclusion of schools that would previously have been excluded was expected to result in an increased non-response rate for the 2006, 2009 and 2012 assessments compared to the 2003 assessment. Consequently, a slightly inflated sample size was required to deal with this expected increase in non-response rate at the school level, so that the actual achieved numbers of schools and students in the sample were adequate.

In line with the procedures adopted in 2006 and 2009, if a small school (fewer than five students) was selected, then this school was only required to complete the pencil-and-paper tasks. In this way, very small schools were not excluded from the sample.

Table 3.1 shows the 2012 estimate of the number of educational institutions and students in the sampling frame for each jurisdiction, as provided by ACARA.

Table 3.1 Estimated 2012 Year 6 enrolment figures as provided by ACARA

State / Territory	Institutions	Students	Percentage of students
ACT	97	4503	1.6
NSW	2364	87709	32.1
NT	158	3198	1.2
QLD	1392	59231	21.7
SA	605	19326	7.1
TAS	221	6643	2.4
VIC	1766	62916	23.0
WA	883	29486	10.8
AUST	7486	273012	100.0

Note: Some percentages may not add up to 100 due to rounding.

3.3 School and student non-participation

In large scale assessments of this kind it is important to document reasons for non-participation so that interpretations of the main findings from the study can be appropriately made within the contexts of the assessment. As in the 2009 assessment, the 2012 study made provisions to document the reasons for school and student non-participation. Figure 3.1 lists the non-participation categories documented in the 2012 study whilst Table 3.2 details the exemption and refusal categories for non-participating schools and students.

Figure 3.1 NAP–SL non-participation categories

exemptions: exercise of principals' prerogative, subject to guidelines provided; and

refusals: specific parent objection to this form of assessment and consequential withdrawal of students from the program.

Table 3.2 NAP–SL exemption and refusal codes

Code	Category description
11	Not included; functional disability. Student has a moderate to severe permanent physical disability such that he/she cannot perform in the testing situation. Functionally disabled students who can respond to the assessment should be included.
12	Not included; intellectual disability. Student has a mental or emotional disability and is cognitively delayed such that he/she cannot perform in the testing situation. This includes students who are emotionally or mentally unable to follow even the general instructions of the assessment. Students should NOT be excluded solely because of poor academic performance or disciplinary problems.
13	Not included; limited assessment language proficiency. The student is unable to read or speak any of the languages of the assessment in the country and would be unable to overcome the language barrier in the testing situation. Typically a student who has received less than one year of instruction in the languages of the assessment may be excluded.
14	Not included; parent/caregiver requested that student not participate OR student refusal.

3.4 Sampling size estimations

To estimate the required sample size for each state and territory, the key consideration is the required degree of precision for the mean estimate of scientific literacy for each state and territory. As with many international studies of this kind, the stipulated precision for the estimated mean score for each state and territory is that the 95 per cent confidence interval around the estimated mean score should be within +/– 0.1s, where s is the standard deviation of the scientific literacy ability distribution in each jurisdiction. This degree of precision for the mean score corresponds to an effective sample size of 400 students. That is, if a simple random sample is taken, the required precision will be achieved with a sample size of 400. With assessments of this kind, simple random samples are usually not used because of logistical difficulties in administering tests in potentially 400 different locations. Consequently, less efficient sampling methods are used, and the required sample size needs to be larger than 400. More specifically, when the design effect[3] of the sample design is taken into account, the required sample size for each state and territory is given by:

$$n_c = n^* \times deff \qquad (1)$$

where n_c is the required sample size, n^* is the effective sample size, and *deff* is the design effect.

The NAP–SL specifications set the target sample at 13 000 students. The achieved precision of the statistics reported in the 2009 NAP–SL was analysed in order to establish whether the sample size of approximately 13 000 students enabled the stipulated precision to be achieved. Table 3.3 contains a summary of the achieved standard error (SE) and confidence interval for each state and territory as well as the value of the desired confidence intervals that correspond to the stipulated precision of +/– 0.1s. The table below also contains the desired sample size that complies with the stipulated precision for each jurisdiction.

Table 3.3 Empirical design effect observed in 2009

State/ Territory	Mean	SE	Confidence interval	Sample size	Desired confidence interval	Empirical design effect	Desired sample size
ACT	415	5.4	10.6	1199	9.3	4.1	1627
NSW	396	6.2	12.1	2092	9.8	8.3	3311
NT	326	14.6	28.6	743	12.8	9.7	3868
QLD	385	4.5	8.9	2043	9.3	4.9	1961
SA	380	5.3	10.4	1848	9.2	6.2	2478
TAS	386	6.9	13.5	1167	9.6	6	2407
VIC	398	4.7	9.2	2040	8.7	5.9	2346
WA	393	4.9	9.6	2030	9.5	5.4	2161
Total				13162			20158

3 The design effect is the ratio of the sampling variance, under the method used, to the sampling variance if a simple random sample had been chosen. That is, design effect is a measure of the loss of sampling efficiency.

As can be seen from Table 3.3, the analysis showed that the 2009 sample size for each state and territory was underestimated relative to the sample size that has the capacity to provide the stipulated precision (except for QLD). In order to rectify this problem it was proposed that the 2012 sample size be increased according to the magnitude of the design effect empirically established in 2009. However, this proposition was not approved by ACARA and the brief was issued that the 2012 sample size should approximate that of 2009. Consequently the proposed target sample size for 2012 was set equal to that for 2009 and is shown in Table 3.4.

Table 3.4 Proposed 2012 sample sizes for drawing samples

State / Territory	Students	Schools
ACT	1400	58
NSW	2100	92
NT	950	50
QLD	2100	91
SA	2100	94
TAS	1400	63
VIC	2100	92
WA	2100	93
Total	14250	633

3.5 Stratification

The sample methodology for the 2009 cycle of the NAP–SL was finalised in 2008. Since then, National Assessment Program – Literacy and Numeracy (NAPLAN) data have been collected annually for the full cohort of Australian schools. It was proposed that NAPLAN results be included as an additional implicit stratification variable for sampling in the 2012 NAP–SL assessment cycle. It was anticipated this would lead to increased precision in the estimates of Science Literacy without increasing the sample size, provided NAPLAN achievement was associated with achievement in Science Literacy.

The sampling frame was partitioned into 24 separate school lists with each list being a unique combination of state and territory (8) and school type (3 – government, Catholic and other). This explicit stratification was performed to ensure that an adequate number of students were sampled from each school type in each jurisdiction. Within each of the separate strata, schools were ordered (implicitly stratified) firstly according to their 2010 NAPLAN quintile rank, then by geographic location[4] and then according to their measure of size – which was related to the estimated number of Year 6 enrolments[5]. For most schools, the Measure of Size (MOS) was set to the 2011 Year 6 enrolment size (ENR) of the school. A school's MOS was adjusted if the school had a small or, alternatively, a very large number of Year 6 students.

4 As per MCEECDYA's definition.

5 The original Year 6 (gr06) variable was used to estimate the total number of students overall and per stratum. For the sample selection, the Year 6 estimated enrolment size (gr06) was initially rounded to the nearest whole number for each school.

Whilst sampling methods for both these school types are described in more detail in the subsequent sections, in general small schools had their MOS adjusted so that their selection in the sample would not result in excessively large sampling weights. In addition, very large schools had their MOS reduced so that they were not selected more than once.

The sample selection procedures were based on the Target Cluster Size (TCS) which was an estimate of the average class size in Australia. The TCS was set at 25 which was the same as for the 2009 NAP–SL assessment (National Assessment Program – Science Literacy Technical Report 2009, Section 3.5). Schools with an enrolment size less than the TCS had a MOS set to the average enrolment size of the same category of small schools within each jurisdiction. This was performed to prevent excessively large sampling weights and was only applied after stratification had occurred.

3.5.1 Small schools

If a large number of schools that were sampled had an ENR less than the TCS, then the actual number of students sampled could be less than the overall target sample. Schools with enrolment sizes less than the TCS are classified as small schools in both OECD (2012) and IEA (2009); however these studies have different approaches for the treatment of small schools within the sampling frame. In the 2012 NAP–SL, OECD (2012) guidelines were utilised for classifying and stratifying small schools, whilst an adapted version of IEA's (2009) treatment of small school MOS values was used.

As a preliminary exercise, schools were classified into different sizes according to OECD (2012, p. 68) classification rules: Large (MOS >= 25) and Small schools which were sub-divided into either Moderately Small (TCS/2 <= MOS < TCS) or Very Small (MOS < TCS/2) schools.

Table 3.5 shows the proportions of Large, Moderately Small and Very Small schools within each jurisdiction. It can be seen that there are many small schools in each jurisdiction. As such, it was important that an appropriate strategy was utilised to prevent an over-selection of small schools, which would have resulted in a lower sample size than the desired target sample size.

OECD (2012) guidelines were used for classifying and stratifying small schools, which involved deliberately under-sampling small schools and slightly over-sampling large schools. This ensured that small schools were represented in the sample while still achieving an adequate overall student sample size without substantially increasing the total number of schools sampled (see OECD 2012, pp. 68–74).

The MOS for a small school was set to the average ENR of all schools within the same explicit stratum and school size category. This strategy was adapted from the IEA (2009) approach to ensure that selection of very small schools would not result in excessively large sampling weights (see IEA 2009, pp. 85–87, section 5.4.2).

Table 3.5 Proportions of schools by school size and jurisdiction

State / Territory	School size	No. Schools	Percentage of schools	No. Students	Percentage of students
ACT	Large	77	79.4	4190	93.0
	Moderately small	14	14.4	268	6.0
	Very small	6	6.2	45	1.0
	Total	97	100.0	4503	100.0
NSW	Large	1386	58.6	77362	88.2
	Moderately small	379	16.0	7001	8.0
	Very small	599	25.3	3346	3.8
	Total	2364	100.0	87709	100.0
NT	Large	55	34.8	2355	73.6
	Moderately small	26	16.5	454	14.2
	Very small	77	48.7	389	12.2
	Total	158	100.0	3198	100.0
QLD	Large	799	57.4	53691	90.6
	Moderately small	171	12.3	3161	5.3
	Very small	422	30.3	2379	4.0
	Total	1392	100.0	59231	100.0
SA	Large	324	53.6	15923	82.4
	Moderately small	132	21.8	2464	12.7
	Very small	149	24.6	939	4.9
	Total	605	100.0	19326	100.0
TAS	Large	118	53.4	5453	82.1
	Moderately small	40	18.1	747	11.2
	Very small	63	28.5	443	6.7
	Total	221	100.0	6643	100.0
VIC	Large	1021	57.8	54229	86.2
	Moderately small	328	18.6	6131	9.7
	Very small	417	23.6	2556	4.1
	Total	1766	100.0	62916	100.0
WA	Large	518	58.7	25847	87.7
	Moderately small	119	13.5	2163	7.3
	Very small	246	27.9	1476	5.0
	Total	883	100.0	29486	100.0

Note: Some percentages may not add up to 100 due to rounding

3.5.2 Very large schools

Selecting schools with a probability proportional to size (PPS) can result in a school being sampled more than once if its ENR is sufficiently large. This can occur when the school enrolment size is larger than the explicit stratum sampling interval. To overcome this, very large schools had their MOS set equal to the size of the sampling interval of the explicit stratum that the school belonged to (an option that was utilised in IEA 2009, pp. 85–87, section 5.4.2).

3.6 Replacement schools

Replacement schools were included in the sample to help overcome problems in relation to school non-participation. There is a risk in any sample assessment that, if the non-participation rate is high, then the target sample sizes will not be achieved. Further, if non-participating schools tend to be lower performing schools, then a bias in the estimated achievement levels will likely occur.

If a school elected not to participate for some reason, then a replacement school was selected for inclusion in the sample. Replacement schools were assigned as per PISA procedures (OECD 2012, p. 67). That is, for a sampled school, the school immediately following it in the sampling frame was assigned as the first replacement school for it, and the school immediately preceding it was assigned as the second replacement school.

3.7 Class selection

One class containing Year 6 students was sampled per school. In some schools where there were several Year 6 classes, each with a small number of Year 6 students, the classes were combined to create a pseudo-class, where possible. Classes generally had equal probabilities of selection. The overall procedure for class selection was as follows:

1. Small classes were combined to create a pseudo-class.
2. Each natural or pseudo-class (referred to as a cluster) was assigned a random number.
3. The clusters in a school were ordered by the assigned random numbers.
4. The first cluster on each school's ordered list was chosen for the sample.

3.7.1 Small classes

In a number of cases, schools had multi-level or remedial classes that contained small numbers of Year 6 students. If many of these small classes are selected, the total sample size will likely be less than the original target sample size, as the class size for these classes is much smaller than the average class size of 25. Twenty-five was determined as the basis for the estimation of the number of schools and classes to be selected.

To overcome this problem, a strategy was employed that built on the procedures used by IEA (2009). Classes with fewer than 15 students were combined with another Year 6 class at the same school. The resulting pseudo-class was considered a single class for sampling purposes.

3.8 The 2012 NAP–SL sampling frame

Table 3.6 outlines the sampling frame for the number of schools by stratum to be sampled using the procedures outlined in the previous sections. Further details on the characteristics of the schools actually sampled are included in Appendix 3.

Table 3.6 Number of schools by stratum to be sampled according to the sampling frame

State / Territory	Proposed target sample size for 2012	Number of schools by stratum					Total
		Very small	Moderately small	Large Catholic	Large govt	Large other	
ACT	1400	1	4	15	30	7	58
NSW	2100	7	9	15	52	9	92
NT	950	12	8	4	23	3	50
QLD	2100	8	6	14	54	10	91
SA	2100	8	15	14	44	13	94
TAS	1400	7	9	8	35	4	63
VIC	2100	7	11	17	48	10	92
WA	2100	9	9	13	51	11	93
Total	14250	59	71	100	336	67	633

Note: Numbers may not add up to the total due to rounding.

3.9 The 2012 NAP–SL sample

Table 3.7 provides a breakdown of the sample according to jurisdiction. The target sample is the number of Year 6 students *enrolled at the time of testing* in the sampled schools. The achieved sample is the number of Year 6 students who participated (attempted the test).

Table 3.7 2012 NAP–SL target and achieved sample sizes by jurisdiction

State / Territory	Number of students enrolled at the time of testing		Number of students who participated in the test	
	Students	Percentage of students	Students	Percentage of students
ACT	1305	8.9	1242	9.4
NSW	2246	15.3	2060	15.6
NT	959	6.5	710	5.4
QLD	2207	15.0	2052	15.5
SA	2082	14.2	1926	14.6
TAS	1420	9.7	1259	9.5
VIC	2112	14.4	1854	14.0
WA	2344	16.0	2133	16.1
Total	14675	100.0	13236	100.0

Note: Numbers may not add up to 100 due to rounding.

The numbers of non-participating students are provided in Table 3.8, broken down by jurisdiction and reason for non-participation.

Table 3.8 Student non-participation by jurisdiction

State / Territory	Non-inclusion code					Total
	Absent	Functional disability	Intellectual disability	Limited language proficiency	Student or parent refusal	
ACT	54	1	6	2	0	63
NSW	181	0	3	0	2	186
NT	227	1	4	13	4	249
QLD	124	5	16	7	3	155
SA	141	3	6	3	3	156
VIC	143	2	8	5	3	161
TAS	215	1	9	7	26	258
WA	190	3	9	8	1	211
AUST	1275	16	61	45	42	1439

The 2012 NAP–SL results were analysed in order to assess the magnitude of the misalignment between the sample size and the precision requirement of +/– 0.1s because the analysis of the design effect for 2009 showed that the planned sample size was not large enough to provide the stipulated precision. As can be seen in Table 3.9, the 2012 sample size did not include a sufficient number of students in order to comply with the stipulated precision for some jurisdictions. The design effect observed in 2012 was smaller than observed in 2009 (Table 3.3) for some jurisdictions. This may be due to the inclusion of NAPLAN results as an implicit stratification variable in 2012.

Table 3.9 Empirical design effect observed in 2012

State / Territory	Mean	SE	Confidence interval	Sample size	Desired confidence interval	Empirical design effect	Desired sample size
ACT	429	6.7	13.2	1242	9.5	6.2	2486
NSW	395	5.1	9.9	2060	9.9	5.4	2145
NT	319	15.9	31.1	710	13.3	10.1	4051
QLD	392	3.3	6.4	2052	9.4	2.5	1012
SA	392	4.0	7.9	1926	9.3	3.6	1425
TAS	395	6.3	12.3	1259	10.1	4.9	1967
VIC	393	5.0	9.7	1854	9.3	5.3	2115
WA	406	4.9	9.5	2133	10.1	4.9	1956
Total				13236			17157

Additional technical specifications can be found in Appendices 5 and 6.

Chapter 4
Test Administration Procedures and Data Preparation

4.1 Registration of classes and students

For most jurisdictions, School Contact Officers nominated by the sample schools were informed that they were to register their students using the templates provided by EAA. In some jurisdictions the student registration task was completed centrally. The student registration procedures were designed so that student information could be collected, coded and then used for further analysis. Registration of students prior to the test day also allowed EAA to overprint the test booklets with individual student details and provide students with the practical task allocated to their classes. These steps also ensured that every student received the correct practical task materials and that student details could be cross-checked.

4.2 Administering the tests to students

The final assessments were administered to the sampled students in October 2012. The participating schools were sent the following materials: a School Contact Officer's Manual (sent on behalf of ACARA in June 2012 along with a brochure for teachers and a brochure for parents explaining the assessment), the Test Administrator's Manual and the assessment instruments, together with the appropriate practical materials for the particular task being undertaken and the Student Survey. Detailed instructions were also given in relation to the participation or exclusion of students with a disability and students from non-English speaking backgrounds.

The teachers were able to review the Test Administrator's Manual before the assessment date and raise questions with EAA or the coordinators of the National Assessment Program – Science Literacy (NAP–SL) in their jurisdiction. EAA provided a toll-free telephone number and an email address so that any queries from teachers could be quickly addressed.

The assessment instruments were administered to a sample consisting of 4.85 per cent of the total Australian Year 6 student population. Tests were administered on the following dates:

- 17 October 2012 – New South Wales, Northern Territory, Queensland, South Australia, Tasmania and Victoria
- 24 October 2012 – Australian Capital Territory and Western Australia

Students' regular class teachers administered the tests in order to minimise disruption to the normal class environment. Teachers followed the standardised test administration procedures in the Test Administrator's Manual.

Teachers were required to complete a Student Participation Form, confirming details about any student who had not participated in or had been excluded from the assessment (see Appendix 6 Student Participation Form).

A quality-monitoring program was established to gauge the extent to which class teachers followed the specified administration procedures. This involved trained observers monitoring the administration of the assessment in a representative sample of classes in 32 (approximately five per cent) of the participating schools.

The test observers were required to complete a report for each assessment they observed. They recorded the timing of the assessment, variations in administration procedures and any problems or disturbances which occurred. Their reports indicate a high degree of conformity by schools with the administration procedures. In particular:

- All test observers reported that the timing for the objective part of the assessment was appropriate. In all schools but one, it was reported that all students finished the objective part of the assessment within the allocated time. One test observer noted that four students did not finish the objective assessment within the allocated time. Four test observers noted that more time was needed for the introduction, instructions and materials distribution for the practical tasks. One observer commented that the teacher had to give instructions at a slower pace to accommodate the language requirements of students with special needs.
- The teachers' test instructions to classes in four schools were noted as having varied from the script in the Test Administrator's Manual. In all cases these variations were considered to have been minor (e.g. re-reading instructions for the practical task after one student returned late from a break, and a short interruption from reading the introduction script due to a brief administrative phone call). Test observers reported that these variations to the test administration script did not affect student performance.
- All test observers reported that the locations of the assessment sessions met the requirements set out in the Test Administrator's Manual. Low levels of disruptive student behaviour were recorded by three test observers (e.g. students murmuring, laughing or disturbing each other). Other minor disruptions were recorded during the administration of the assessment. These included brief announcements, noise from other classrooms and visits to the class by other teachers.

4.3 Marking procedures

Each multiple-choice item had only one correct answer. The open-ended items required students to construct their own responses. The open-ended items were further categorised into items that required a single-word or short-sentence response and those that required a more substantive response (referred to as 'extended-response' items). Some open-ended items had polytomous scores. That is, students could score either one mark or two marks depending on the achievement level demonstrated by their response.

Over half of the items were open-ended and required marking by trained markers.

Marking guides were prepared by EAA and refined during the trialling process. The marking team included experienced teacher markers employed by EAA.

The markers participated in a one-and-a-half day training session led by the Test Development Manager. The session involved formal presentations followed by hands-on practice with pre-marked sample student answer booklets. Presentations included leading markers through an overview of each cluster or practical task and discussing the marking criteria and illustrative answers for correct and incorrect student responses exemplified in the marking guides. In the hands-on practice, markers practised marking with a pre-marked sample of items and discussed the scores assigned to each item to help clarify distinctions between score levels. At the end of the session, all markers were asked to mark the same set of student answer booklets. The scores were compared to the scores agreed to by expert scorers (the Project Director, the Test Development Manager and the group leaders). Trainers discussed with markers agreements and disagreements between their scores and the scores given by expert scorers. Additional practice was provided to markers for items on which consistency and accuracy were low.

Markers were monitored constantly for reliability by having samples of their student answer booklets check-marked by group leaders. In cases where there were differences between markers and group leaders, the scoring was reconciled jointly in consultation with the Test Development Manager. In addition, once a day all markers were asked to mark the same set of student answer booklets. The scores were compared to the scores agreed to by expert scorers and differences were discussed and reconciled.

In addition, approximately five per cent of the 2009 trend item responses were re-marked by the 2012 markers to ensure the reliability of marking. These procedures, coupled with the intensive training at the beginning of the marking exercise, ensured that markers applied the scoring criteria consistently and accurately.

4.4 Data entry procedures

The multiple-choice responses and teacher marked scores were data processed. A validation of the data processing ensured accuracy in data capture.

Scanning software was used to capture images of all the student responses. The resulting image files have been indexed and provided to ACARA for future reference.

Demographic information and information collected to determine student inclusion in the testing population was obtained from participating schools using the Student Participation Form (SPF). The SPF consisted of two parts: Part A was designed to collect information about the school (including information about the number of students enrolled in Year 6 and the number of classes in Year 6) and Part B was designed to collect relevant information about individual students. A sample of the SPF can be found in Appendix 6.

4.4.1 Data coding rules

Data coding rules for collecting student inclusion information in the SPF are explained in full on pages 8 to 10 of the Test Administrator's Manual. Table 4.1 contains codes that were used and their explanation.

Table 4.1 Codes used in the Student Participation Form

Special education needs codes
0 = No special education needs
1 = Functional disability
2 = Intellectual disability
3 = Limited test language proficiency
Non-inclusion codes
10 = Absent
11 = Not included; functional disability
12 = Not included; intellectual disability
13 = Not included; limited test language proficiency
14 = Student or parent refusal
Indigenous codes
1 = Aboriginal but not Torres Strait Islander origin
2 = Torres Strait Islander but not Aboriginal origin
3 = Both Aboriginal and Torres Strait Islander origin
4 = Neither Aboriginal nor Torres Strait Islander origin
9 = Not stated/unknown

Chapter 5
Computation of Sampling Weights

The sampling weights calculated for the National Assessment Program – Science Literacy (NAP–SL) were based on procedures detailed in the TIMSS 2007 Technical Report (IEA 2009). The procedures outlined in this document were designed for several different sampling scenarios. Only the procedures relevant to the NAP–SL context are presented here.

5.1 School weight

5.1.1 School base weight

School level base weight for school i

$$BW_{sc}^{i} = \frac{M}{n \cdot m_i} \qquad (2)$$

where n was the total number of schools sampled within each explicit stratum and m_i was the Measure of Size (MOS) assigned to the i^{th} school, and

$$M = \sum_{i=1}^{N} m_i \qquad (3)$$

where N was the total number of schools (i.e. both sampled and not sampled) in the explicit stratum.

For small school strata, schools were assigned equal MOS values. Small school sampling weights, using the above equations, can be given by:

$$BW_{sc}^{i} = \frac{N \cdot m_i}{n \cdot m_i} \qquad (4)$$

This can be simplified to:

$$BW_{sc}^{i} = \frac{N}{n} \qquad (5)$$

5.1.2 School non-participation adjustment

School level base weights were calculated for all sampled and replacement schools that satisfied the condition that more than 50 per cent of the eligible students actually participated in the study. In total, 633 schools were sampled of which there

were 16 schools that did not participate in the testing (and could not be replaced). Three schools were found to be ineligible in that there were no Year 6 students enrolled at the school at the time of testing. The remaining 13 schools were either exempted from testing or did not participate for some other reason.

A school-level non-response adjustment was calculated separately for each explicit stratum to account for schools that were sampled but did not participate. Such an adjustment means that the final school weights will be representative of the whole population of Year 6 students rather than the population directly represented by the participating schools.

Specifically, the non-response adjustment was calculated as:

$$A_{sc} = \frac{n_s + n_{r1} + n_{r2} + n_{nr}}{n_s + n_{r1} + n_{r2}} \qquad (6)$$

where:

n_s was the number of originally sampled schools that participated

n_{r1} and n_{r2} was the number of first and second replacement schools, respectively, that participated, and

n_{nr} was the number of schools that did not participate.

Note that the three ineligible schools were not included in the calculation of this adjustment[1].

5.1.3 Final school weight

The final school weight was then the product of the school base weight and non-participation adjustment:

$$FW_{sc}^{i} = BW_{sc}^{i} \cdot A_{sc} \qquad (7)$$

5.2 Class weight

Typically, when a class is selected at random, the probability of selection for the class is $1/n$, where n is the total number of eligible classes in that school. Consequently, the class weight is n.

However, it should be noted that, while an average class size of 25 students is assumed, a considerable number of classes have around 13–15 students. Pseudo-classes were created prior to class selection using the process described in

1 see PISA 2009 Technical Report, Chapter 8: Survey Weighting and the Calculation of Sampling Variance (OECD 2012, p. 121) and TIMSS 2007 Technical Report, Chapter 9: Sampling Weights and Participation Rates (IEA 2009, p. 167)

Chapter 3. Each natural class or pseudo-class within a school was then allocated a cluster ID. Each cluster had an equal probability of being selected. Consequently, class weights were simply equal to the number of clusters at a particular school.

5.2.1 Class base weight

When classes/clusters were selected with equal probability, the base class weight is given by:

$$BW_{cl1}^{i} = \frac{C^{i}}{c^{i}} \qquad (8)$$

where C^i is the total number of classes for the i^{th} school and c^i is the total number of sampled classrooms. For NAP–SL only one class/cluster was selected per school, so the base class weight is simply equal to the number of unique clusters at the school:

$$BW_{cl1}^{i} = C^{i} \qquad (9)$$

5.2.2 Final class weight

The final class weight is equal to the base class weight since classes were selected with equal probabilities.

$$FW_{cl}^{i} = BW_{cl}^{i} \qquad (10)$$

5.3 Student weight

5.3.1 Student base weight

Each student in the sampled class was certain of selection at the student level. The student base weight was therefore equal to 1 for all students.

$$BW_{st}^{i} = 1.0 \qquad (11)$$

5.3.2 Student non-participation adjustment

A student non-participation adjustment was calculated for any school that had at least one student who was eligible to do the test but did not participate for some reason[2]. This was given by:

$$A_{st}^{i} = \frac{s_{rs}^{i} + s_{nr}^{i}}{s_{rs}^{i}} \qquad (12)$$

where s^i_{rs} was the number of eligible students that participated, and s^i_{nr} was the number of eligible students that did not participate, at the i^{th} school.

2 These are the absent and refusal students and does not include exclusions, such as functionally disabled.

5.3.3 Final student weight

The final student weight is then equal to the product of the student base weight and non-participation adjustment.

$$FW_{st}^{i} = BW_{st}^{i} \cdot A_{st}^{i} \qquad (13)$$

This simplifies to:

$$FW_{st}^{i} = A_{st}^{i} \qquad (14)$$

That is, the student final weight is equal to the student non-participation adjustment.

5.4 Final weight

In summary, the final weight is the product of the final school, class and student weights:

$$W^{i} = FW_{sc}^{i} \cdot FW_{cl}^{i} \cdot FW_{st}^{i} \qquad (15)$$

Chapter 6
Item Analysis of the Final Test

6.1 Item analyses

This chapter presents the item analyses of the 2012 National Assessment Program – Science Literacy (NAP–SL) main assessment data.

6.1.1 Sample size

In all, 13 236 students participated in at least one of the two components of the 2012 NAP–SL assessment: the paper-and-pencil test and the practical task. Table 6.1 shows the number of participating students by state and territory.

Table 6.1 Number of participating students by state and territory

State / Territory	Number of students
ACT	1242
NSW	2060
NT	710
QLD	2052
SA	1926
TAS	1259
VIC	1854
WA	2133
Total	13236

6.1.2 Number of students by booklet

Seven test booklets with link items were rotated in each class (see Section 6.2 for the test design). Each student completed only one test booklet. Table 6.2 shows the number of students that completed each test booklet. It can be seen that the test rotation scheme worked well, as the number of students per booklet is approximately equal across the seven booklets. As each objective item appears in three test booklets, the number of students who took each objective item is approximately 5650. As each student completed one of two practical tasks, the number of students who took each practical task item is approximately 6500.

Table 6.2 Number of students by test booklet

Booklet	Number of students
1	1870
2	1875
3	1914
4	1909
5	1900
6	1871
7	1897
Total	13236

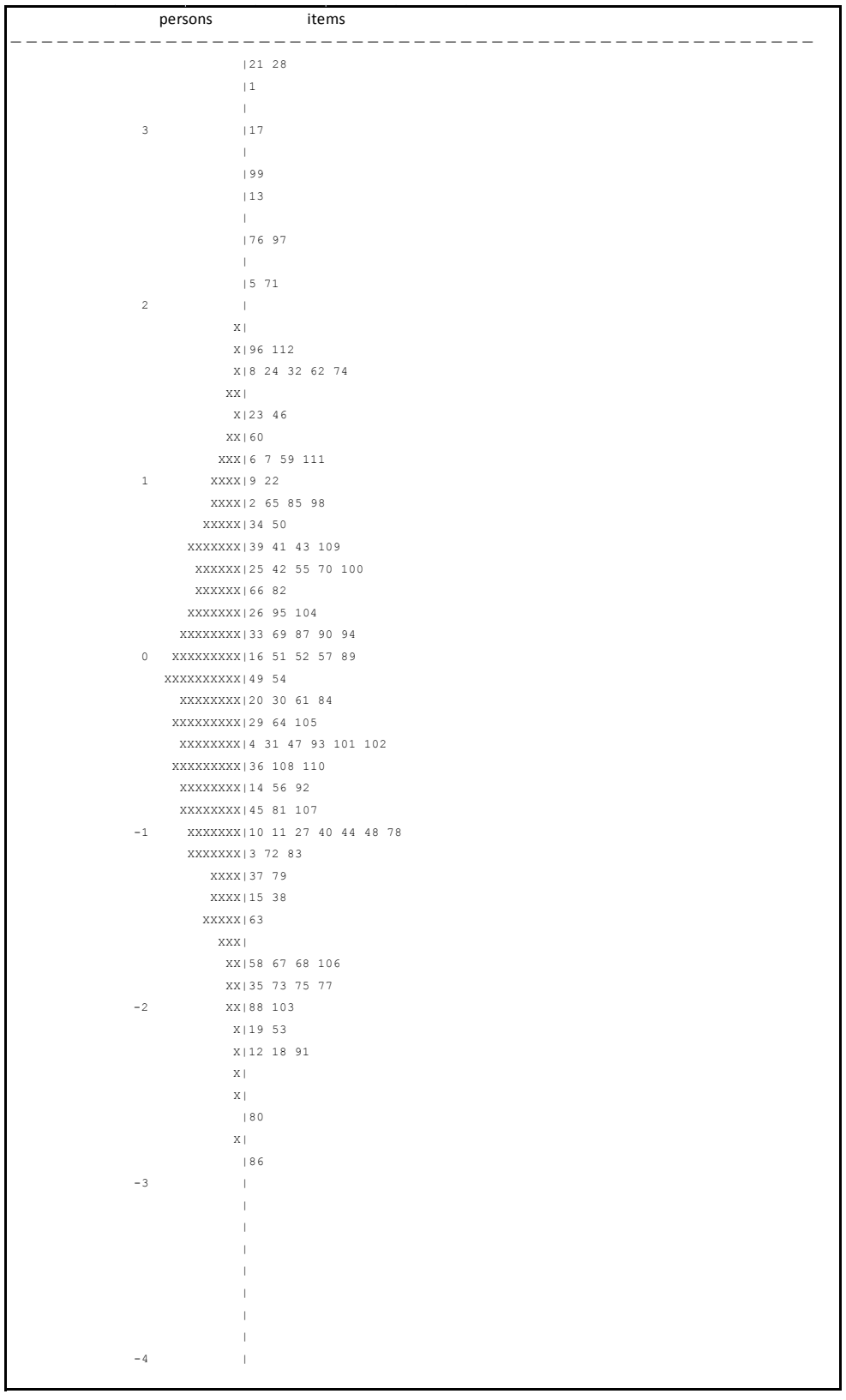
6.1.3 Initial item analysis

The first item analysis carried out was on all data records. No sampling weights were used. This analysis aimed to detect any items that did not function well. In this analysis, all trailing missing item responses were treated as not administered, except for the first item following the last non-missing item. Embedded missing responses were treated as incorrect. A complete list of items and their codes can be found in Table 6.3.

6.1.3.1 Item–person map

Figure 6.1 shows an item–person map from this analysis.

Figure 6.1 Item–person map



The vertical scale in Figure 6.1 shows increasing proficiency, with student ability distribution shown in the left panel (indicated by an 'X'). The items are placed in the right panel (indicated by item numbers) in item difficulty order, where items at the top are most difficult.

Figure 6.1 shows that the items cover a wide range of difficulty levels. The average item difficulty is zero logit, while the average ability is -0.289 logit, showing that the match between item difficulties and person abilities is quite good overall.

Items falling outside parameters of discrimination 0.25–0.5 and fit 0.85–1.15 were checked by test developers and the decision was made to include all of the items in the final data analysis. Item Characteristic Curves (ICCs) from RUMM can be found in the file *NAPSL2012_CheckStateLocations.xlsx* (refer to Section 6.3 for details on how to obtain access to this file).

6.1.3.2 Summary item statistics

Table 6.3 shows summary item statistics for each of the 112 items.

Table 6.3 Summary item statistics in 2012

Item label	Item reference number	Number of students	Percentage correct	Discrimination index	Fit mean square
PAQ01	1	6570	2.59	0.17	1.04
PAQ02	2	6570	27.58	0.34	1.00
PAQ03	3	6570	66.44	0.38	0.98
PAQ04	4	6570	53.27	0.36	1.01
PAQ05	5	6570	11.37	0.29	0.97
PAQ06	6	6570	23.04	0.40	0.94
PAQ07	7	6570	26.37	0.43	0.99
PAQ08	8	6570	15.81	0.35	0.95
PAQ09	9	6570	24.93	0.15	1.13
PAQ10	10	6570	64.81	0.41	0.95
PAQ11	11	6570	62.97	0.46	0.92
PBQ01	12	6498	83.90	0.19	1.06
PBQ02	13	6498	6.82	0.16	1.03
PBQ03	14	6498	59.39	0.33	1.02
PBQ04	15	6498	71.27	0.43	0.91
PBQ05	16	6498	44.40	0.20	1.14
PBQ06	17	6498	5.37	0.16	1.01
PBQ07	18	6498	83.81	0.36	0.93
PBQ08	19	6498	82.69	0.37	0.93
PBQ09	20	6498	50.03	0.48	0.91
PBQ10	21	6498	3.51	0.16	1.01
IDOB532	22	5662	23.77	0.29	1.01
IDOB533	23	5662	19.09	0.17	1.07
IDOB534	24	5662	15.82	0.30	0.98
IDOB483	25	5662	32.39	0.45	0.91

Table 6.3 (Cont.) Summary item statistics in 2012

Item label	Item reference number	Number of students	Percentage correct	Discrimination index	Fit mean square
IDOB484	26	5662	37.85	0.42	0.95
IDOB486	27	5662	62.13	0.37	0.99
IDOB487	28	5662	3.66	0.18	1.00
IDOB489	29	5662	50.90	0.48	0.91
IDOB490	30	5662	48.27	0.50	0.89
IDOB491	31	5662	53.18	0.43	0.95
IDOB470	32	5662	15.26	0.19	1.04
IDOB471	33	5662	38.48	0.31	1.04
IDOB473	34	5662	28.49	0.34	1.00
IDOB097	35	5647	78.20	0.36	0.96
IDOB098	36	5647	55.22	0.35	1.01
IDOB149	37	5647	67.88	0.38	0.96
IDOB150	38	5647	69.36	0.40	0.93
IDOB021	39	5647	29.64	0.34	0.99
IDOB022	40	5647	62.16	0.37	0.98
IDOB023	41	5647	30.95	0.22	1.09
IDOB177	42	5647	33.86	0.38	1.11
IDOB178	43	5647	29.64	0.36	0.98
IDOB084	44	5647	60.70	0.25	1.10
IDOB085	45	5647	59.98	0.36	1.00
IDOB086	46	5647	17.87	0.21	1.07
IDOB087	47	5647	51.83	0.39	0.98
IDOB088	48	5647	59.48	0.40	0.95
IDOB457	49	5668	45.85	0.44	0.93
IDOB458	50	5668	28.93	0.21	1.10
IDOB459	51	5668	42.82	0.26	1.09
IDOB460	52	5668	43.60	0.30	1.06
IDOB492	53	5668	81.05	0.34	0.96
IDOB493	54	5668	44.53	0.24	1.10
IDOB494	55	5668	33.86	0.36	0.99
IDOB496	56	5668	58.91	0.35	1.02
IDOB564	57	5668	42.81	0.35	1.12
IDOB565	58	5668	75.48	0.39	0.94
IDOB566	59	5668	22.35	0.34	0.98
IDOB568	60	5668	21.26	0.32	0.99
IDOB569	61	5668	47.48	0.41	0.96
IDOB570	62	5668	14.71	0.28	1.00
IDOB559	63	5660	72.84	0.41	0.94
IDOB561	64	5660	49.36	0.49	0.89
IDOB562	65	5660	27.00	0.39	0.95
IDOB563	66	5660	34.54	0.36	1.00
IDOB451	67	5660	75.21	0.39	0.95
IDOB452	68	5660	75.39	0.16	1.13
IDOB453	69	5660	40.97	0.26	1.09

Table 6.3 (Cont.) Summary item statistics in 2012

Item label	Item reference number	Number of students	Percentage correct	Discrimination index	Fit mean square
IDOB454	70	5660	33.59	0.38	0.98
IDOB551	71	5660	10.48	0.25	1.00
IDOB552	72	5660	62.93	0.34	1.02
IDOB553	73	5660	76.27	0.38	0.94
IDOB554	74	5660	14.95	0.17	1.08
IDOB503	75	5660	75.44	0.35	0.97
IDOB506	76	5660	8.60	0.27	0.97
IDOB041	77	5599	78.75	0.24	1.05
IDOB044	78	5599	62.76	0.31	1.04
IDOB173	79	5599	68.24	0.22	1.10
IDOB474	80	5599	86.57	0.20	1.03
IDOB475	81	5599	61.48	0.39	0.96
IDOB184	82	5599	37.59	0.38	1.08
IDOB185	83	5599	64.74	0.20	1.13
IDOB186	84	5599	47.31	0.25	1.10
IDOB368	85	5599	25.56	0.30	1.00
IDOB370	86	5599	87.27	0.28	0.96
IDOB371	87	5599	39.83	0.40	0.96
IDOB461	88	5672	81.22	0.31	0.99
IDOB462	89	5672	43.69	0.39	0.96
IDOB517	90	5672	41.59	0.43	0.93
IDOB518	91	5672	84.03	0.35	0.93
IDOB519	92	5672	57.81	0.34	1.01
IDOB521	93	5672	52.50	0.27	1.08
IDOB522	94	5672	40.71	0.38	0.97
IDOB529	95	5672	38.29	0.52	0.86
IDOB530	96	5672	14.32	0.34	0.95
IDOB531	97	5672	8.87	0.25	0.98
IDOB444	98	5672	25.93	0.12	1.14
IDOB446	99	5672	9.06	0.39	0.93
IDOB417	100	5677	34.56	0.46	0.91
IDOB418	101	5677	53.43	0.33	1.02
IDOB419	102	5677	55.06	0.24	1.09
IDOB360	103	5677	79.95	0.34	0.96
IDOB363	104	5677	37.66	0.39	0.96
IDOB362	105	5677	51.66	0.32	1.03
IDOB308	106	5677	76.18	0.39	0.91
IDOB309	107	5677	61.71	0.38	0.95
IDOB310	108	5677	57.30	0.30	1.04
IDOB313	109	5677	32.31	0.29	1.05
IDOB315	110	5677	55.87	0.30	1.04
IDOB405	111	5677	22.95	0.29	1.00
IDOB406	112	5677	14.34	0.24	1.01

6.1.3.3 Test reliability

The person separation index (the proportion of variance of the estimated person measures and the total variance including error) for the 2012 NAP–SL tests was 0.88, which is very high[1].

6.1.4 Booklet effect

'Booklet effect' refers to the differences in booklet difficulties after equating of the booklets has been carried out. That is, students may be advantaged or disadvantaged by taking a particular test booklet, even after booklets have been equated. Table 6.4 shows the booklet difficulty estimates. The estimation of booklet adjustments was carried out through a ConQuest analysis with the model statement:

*booklet + item + item*step*

Table 6.4 Booklet difficulty parameters

Booklet number	Booklet parameter (logit)	Error
1	-0.028	0.005
2	-0.051	0.006
3	-0.003	0.005
4	-0.029	0.006
5	0.055	0.006
6	0.010	0.006
7	0.044	0.014

The booklet parameters shown in Table 6.4 are very close to zero, indicating that booklet effect was not a serious issue for this assessment. It is noted that booklet 2 seems to be somewhat easier and booklet 5 appears to be more difficult than the other booklets. However, in estimating the student Proficiency Levels, the booklet effect was taken into account. In doing so, the booklet effect was set as one of the model parameters in estimating the student parameters in ConQuest.

6.1.5 Item statistics by state and territory

While the items worked quite well in general for the overall sample, it is important to check if the items performed well within each state and territory, and whether the item difficulties are similar across states and territories. For a few items, the discrimination index falls below 0.2 for some states and territories. In particular, item IDOB444 is the least discriminating item. The discrimination index of this item for VIC and NT is 0.04 and 0.07 respectively. Detailed item statistics for this item are shown in Figure 6.2. It can be seen from Figure 6.2 that option 4 of this item strongly attracted students in VIC and NT. The item required students to identify the experimental setup that will help answer a scientific question, i.e. 'which box material affects the temperature of the air inside the box'.

1 In comparison, the reported reliability for PISA 2009 science was 0.89. Reported reliability for TIMSS 2007 Grade 6 and Grade 9 was 0.80 and 0.84 respectively.

Approximately 48 per cent of students in VIC and 45 per cent of students in NT chose the incorrect option 4; 'Box 5 and Box 7'. This misunderstanding may be due to students taking into consideration the type of outside surface rather than the box material.

Figure 6.2 Item analysis for item IDOB444 for VIC and NT

VIC: item: (IDOB444)

Cases for this item 800 Discrimination 0.04

Item Threshold(s): 0.89

Item Delta(s): 0.89

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
1	0.00	71	8.88	-0.11	-3.00 (.003)	-0.64	0.92
2	0.00	111	13.88	-0.12	-3.50 (.000)	-0.57	1.00
3	1.00	216	27.00	0.04	1.05 (.293)	-0.13	0.92
4	0.00	387	48.38	0.14	3.90 (.000)	-0.22	0.80
7	0.00	2	0.25	0.04	1.00 (.318)	0.25	0.95
9	0.00	12	1.50	-0.12	-3.48 (.001)	-1.27	0.98
A	0.00	1	0.13	0.02	0.62 (.536)	0.05	0.00

NT: item: (IDOB444)

Cases for this item 283 Discrimination 0.07

Item Threshold(s): 1.18

Item Delta(s): 1.18

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
1	0.00	33	11.66	-0.21	-3.63 (.000)	-0.72	1.10
2	0.00	49	17.31	-0.13	-2.25 (.025)	-0.47	0.96
3	1.00	67	23.67	0.07	1.19 (.234)	-0.06	0.93
4	0.00	128	45.23	0.21	3.62 (.000)	-0.06	0.91
7	0.00	1	0.35	-0.08	-1.43 (.154)	-0.82	0.00
9	0.00	5	1.77	-0.09	-1.55 (.122)	-0.88	0.98

6.1.6 Comparison of item difficulty parameters across states and territories

Figure 6.3 shows a comparison of item difficulties calibrated for each state and territory separately, using ConQuest. For each state and territory, the average item difficulty was set to zero, so that each item difficulty shows the deviation from the average item difficulty within that state and territory. In this way, the item difficulties across different states and territories can be compared, as the overall ability level of students for each state and territory is controlled for. If an item has very different difficulty values across states and territories, then there is evidence of differential item functioning. Figure 6.3 shows that the calibrated item difficulties are very similar across states and territories. That is, there is little evidence of differential item functioning. Similarly, there is no significant difference in the item discrimination indices across states and territories, as shown in Figure 6.4.

Further analyses using RUMM software show that for most items the locations are similar across states and territories. However, when comparing the state and territory location to the whole sample location, a few items fall outside of the confidence interval. For further details please refer to the spreadsheet *NAPSL2012_CheckStateLocations.xlsx* (refer to Section 6.3 for details on how to obtain access to this file).

Figure 6.3 Comparison of item difficulty parameters across states and territories

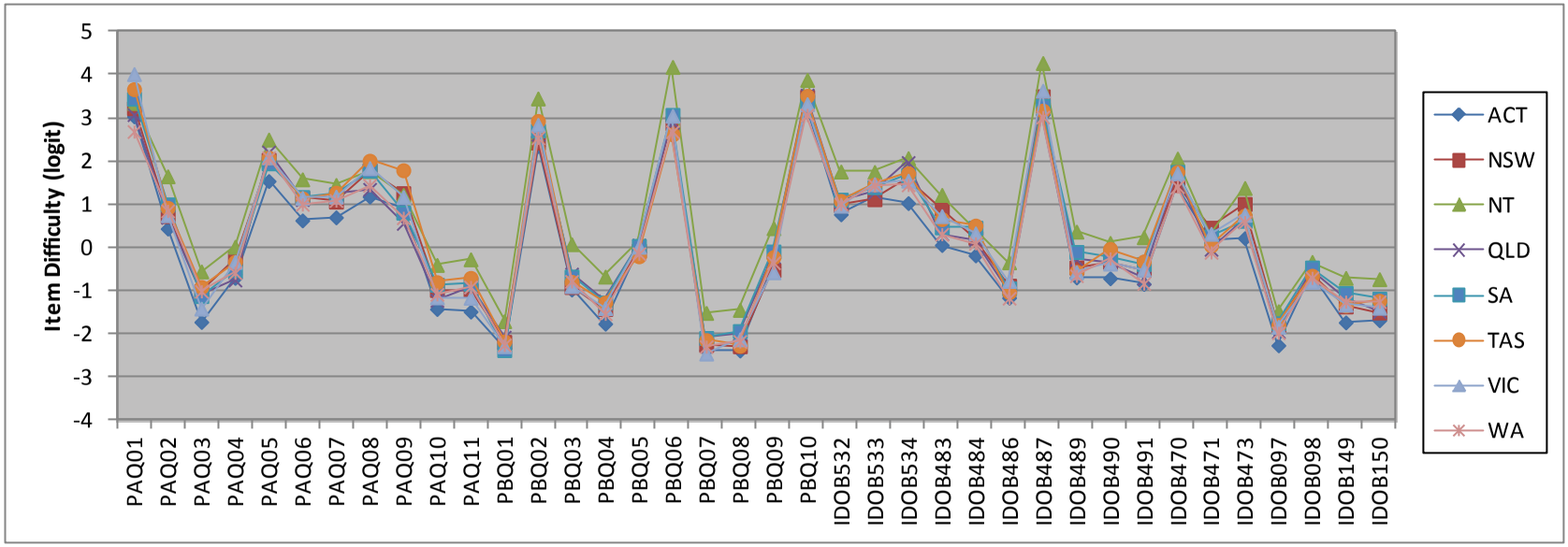


Figure 6.3 (Cont.) Comparison of item difficulty parameters across states and territories

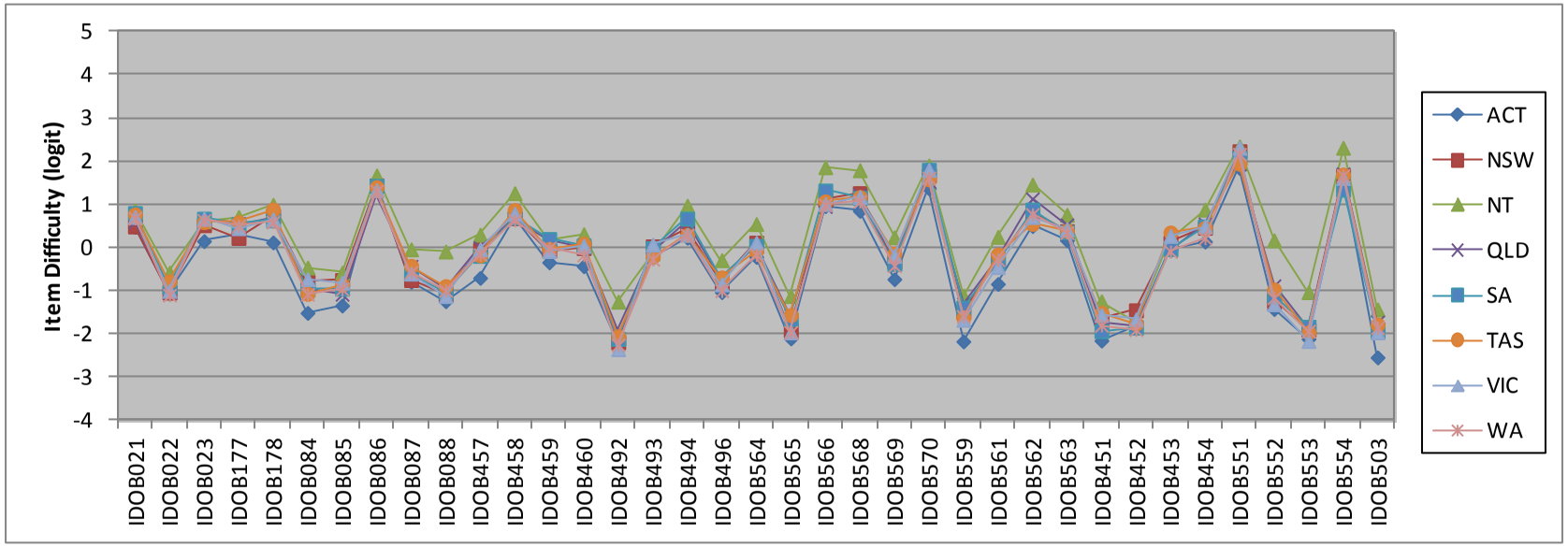


Figure 6.3 (Cont.) Comparison of item difficulty parameters across states and territories

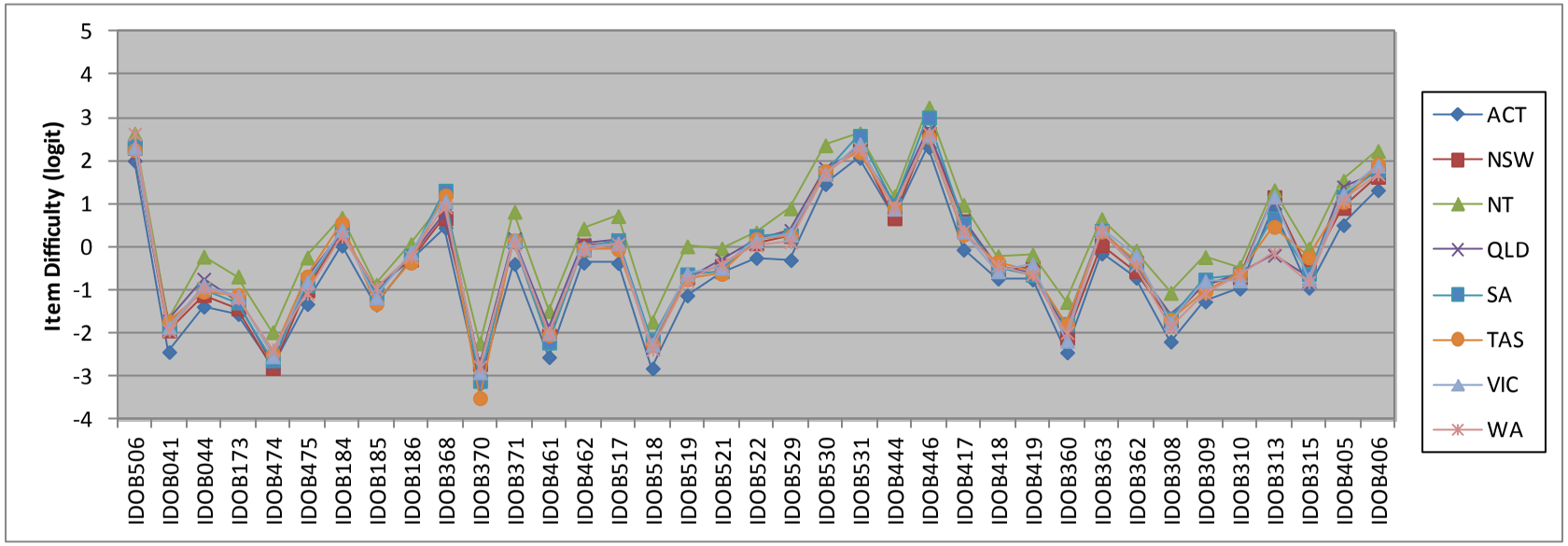


Figure 6.4 Discrimination index by state/territory

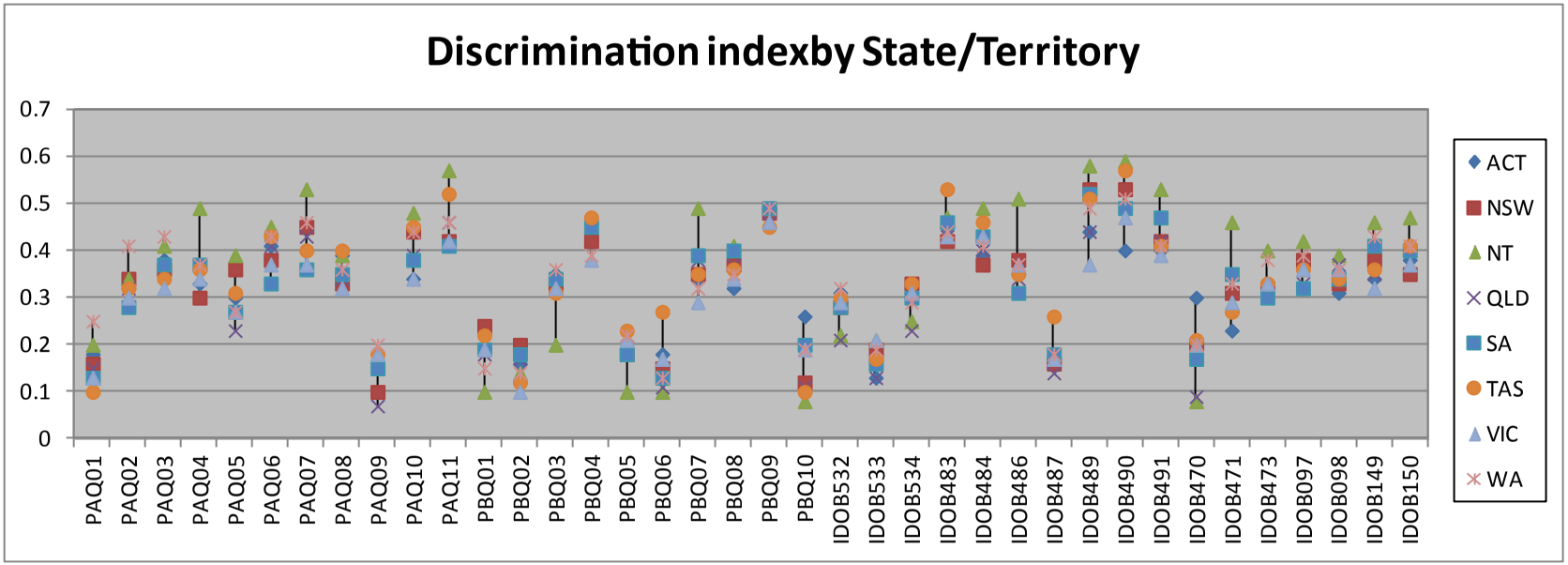


Figure 6.4 (Cont.) Discrimination index by state/territory

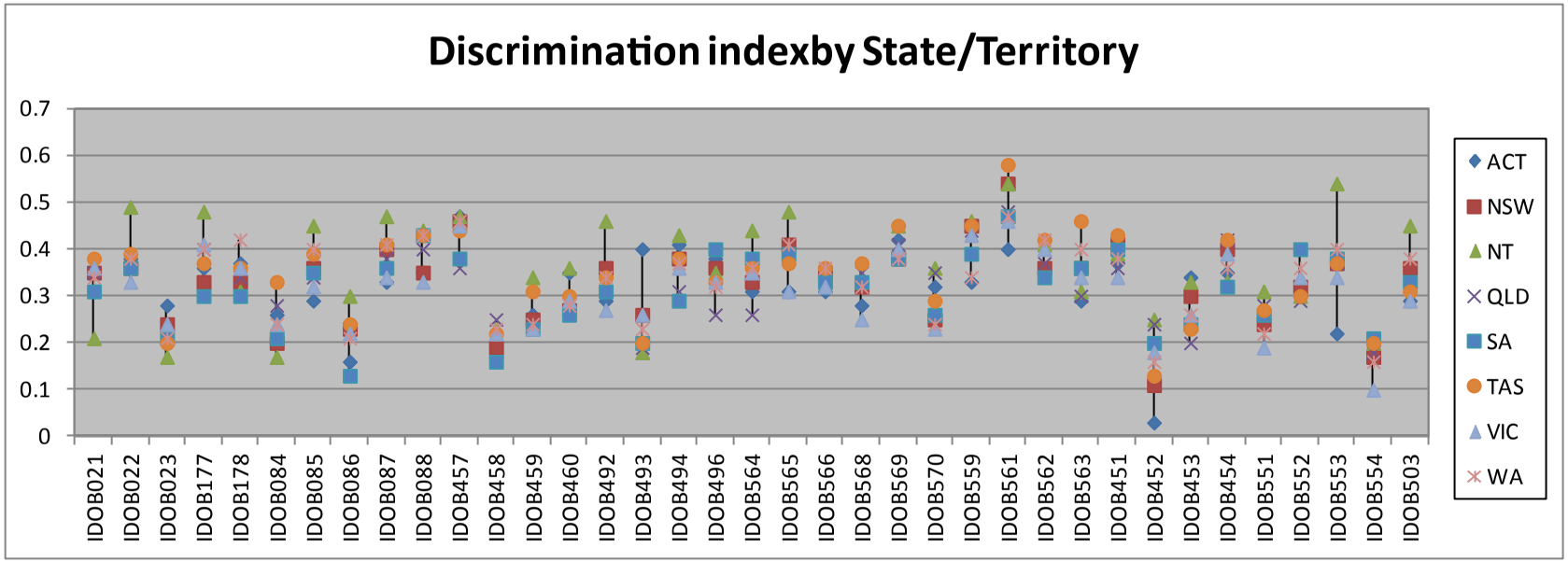
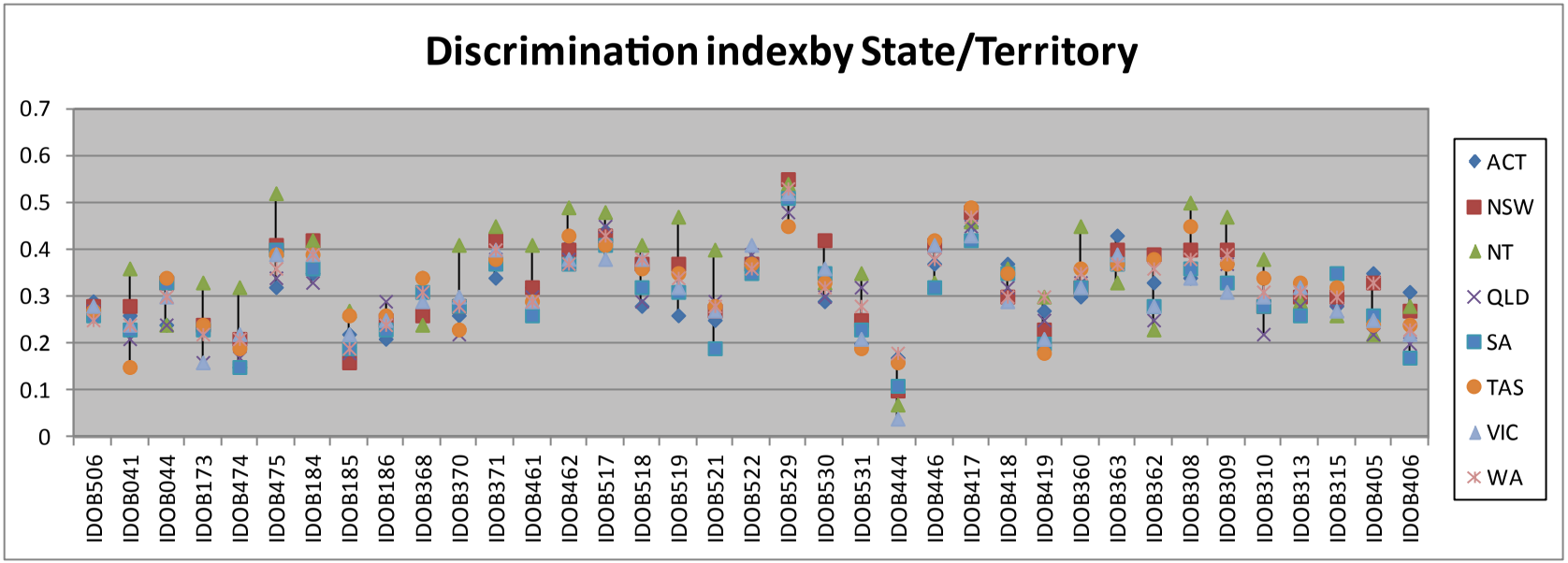


Figure 6.4 (Cont.) Discrimination index by state/territory



6.1.7 Item difficulty by gender

Table 6.5 shows item parameters calibrated separately for gender groups, arranged in order of the difference between the item difficulty parameters. The left side of the table shows items where boys performed better, and the right side of the table shows items where girls performed better. For most items, the difference in item difficulty parameters is small. If one takes 0.5 logits as a cut-off value for identifying a relatively large gender difference, then only six items fall in this category: boys performed better on item IDOB453, item IDOB451, item IDOB483 and IDOB568, and girls performed better on item IDOB563 and item PBQ06. These six items were retained in the analysis owing to the fact that the estimation model included gender as a regression term and was thus able to absorb the observed gender DIF for these six items. Item IDOB529 showed the smallest difference (0.004 logits) in item difficulty between boys and girls.

Table 6.5 Item difficulty parameters for gender groups

Boys performed better				Girls performed better			
Code	**Girls**	**Boys**	**Diff**	**Code**	**Girls**	**Boys**	**Diff**
IDOB453	0.460	-0.260	0.720	IDOB563	0.176	0.702	-0.526
IDOB451	-1.416	-2.010	0.594	PBQ06	2.710	3.222	-0.512
IDOB483	0.831	0.311	0.520	IDOB474	-2.787	-2.343	-0.444
IDOB568	1.476	0.962	0.514	IDOB370	-3.095	-2.653	-0.442
IDOB452	-1.500	-1.954	0.454	PAQ03	-1.300	-0.892	-0.408
IDOB446	2.961	2.521	0.440	PAQ09	0.832	1.210	-0.378
IDOB489	-0.166	-0.604	0.438	PBQ09	-0.447	-0.099	-0.348
PBQ05	0.210	-0.200	0.410	IDOB566	0.966	1.304	-0.338
IDOB085	-0.710	-1.096	0.386	IDOB534	1.483	1.809	-0.326
IDOB084	-0.742	-1.110	0.368	PBQ01	-2.354	-2.030	-0.324
IDOB149	-1.059	-1.401	0.342	PAQ02	0.697	1.017	-0.320
IDOB473	0.899	0.581	0.318	PAQ06	0.973	1.293	-0.320
IDOB457	0.037	-0.241	0.278	PAQ04	-0.589	-0.273	-0.316
IDOB470	1.797	1.521	0.276	IDOB561	-0.450	-0.142	-0.308
IDOB173	-1.071	-1.347	0.276	IDOB486	-1.066	-0.772	-0.294
IDOB405	1.258	0.992	0.266	IDOB461	-2.163	-1.875	-0.288
IDOB462	0.144	-0.114	0.258	IDOB559	-1.652	-1.370	-0.282
IDOB517	0.240	0.002	0.238	IDOB313	0.457	0.717	-0.260
IDOB044	-0.796	-1.032	0.236	IDOB564	-0.077	0.175	-0.252
IDOB518	-2.135	-2.367	0.232	IDOB371	0.035	0.281	-0.246
PAQ05	2.190	1.964	0.226	PAQ10	-1.122	-0.888	-0.234
IDOB363	0.426	0.210	0.216	IDOB570	1.570	1.804	-0.234
IDOB021	0.793	0.591	0.202	IDOB097	-1.954	-1.728	-0.226
IDOB493	0.040	-0.158	0.198	IDOB086	1.303	1.525	-0.222
IDOB503	-1.772	-1.964	0.192	PAQ08	1.550	1.764	-0.214
IDOB533	1.483	1.295	0.188	IDOB308	-1.780	-1.580	-0.200
IDOB521	-0.324	-0.502	0.178	IDOB177	0.393	0.591	-0.198
IDOB362	-0.281	-0.457	0.176	PBQ08	-2.164	-1.988	-0.176
IDOB460	0.080	-0.084	0.164	IDOB309	-0.959	-0.783	-0.176

Table 6.5 (Cont.) Item difficulty parameters for gender groups

Boys performed better			
Code	**Girls**	**Boys**	**Diff**
IDOB023	0.680	0.542	0.138
IDOB310	-0.587	-0.721	0.134
IDOB454	0.532	0.402	0.130
IDOB041	-1.790	-1.914	0.124
IDOB490	-0.204	-0.324	0.120
IDOB531	2.418	2.310	0.108
PBQ04	-1.272	-1.378	0.106
IDOB496	-0.719	-0.819	0.100
IDOB554	1.716	1.622	0.094
IDOB459	0.081	-0.005	0.086
IDOB491	-0.476	-0.558	0.082
IDOB506	2.383	2.303	0.080
IDOB551	2.171	2.095	0.076
IDOB532	1.107	1.037	0.070
IDOB569	-0.239	-0.297	0.058
IDOB098	-0.540	-0.596	0.056
IDOB565	-1.720	-1.766	0.046
IDOB492	-2.047	-2.091	0.044
IDOB315	-0.580	-0.622	0.042
IDOB418	-0.427	-0.465	0.038
PBQ10	3.433	3.401	0.032
IDOB530	1.785	1.755	0.030
PBQ07	-2.146	-2.172	0.026
IDOB444	0.935	0.911	0.024
IDOB519	-0.659	-0.675	0.016
PAQ01	3.214	3.208	0.006
IDOB529	0.272	0.268	0.004

Girls performed better			
Code	**Girls**	**Boys**	**Diff**
PBQ03	-0.802	-0.634	-0.168
IDOB562	0.772	0.940	-0.168
PAQ07	1.090	1.256	-0.166
IDOB185	-1.137	-0.973	-0.164
IDOB475	-0.938	-0.782	-0.156
IDOB417	0.406	0.556	-0.150
PAQ11	-0.983	-0.845	-0.138
IDOB022	-1.014	-0.878	-0.136
IDOB186	-0.243	-0.133	-0.110
IDOB471	0.146	0.246	-0.100
IDOB484	0.229	0.325	-0.096
IDOB087	-0.581	-0.487	-0.094
IDOB368	0.904	0.994	-0.090
IDOB494	0.434	0.520	-0.086
IDOB178	0.633	0.717	-0.084
IDOB088	-0.994	-0.922	-0.072
IDOB458	0.730	0.792	-0.062
IDOB406	1.738	1.800	-0.062
IDOB522	0.124	0.178	-0.054
IDOB487	3.324	3.374	-0.050
IDOB552	-1.064	-1.020	-0.044
IDOB184	0.354	0.396	-0.042
IDOB150	-1.318	-1.304	-0.014
IDOB553	-1.875	-1.863	-0.012
PBQ02	2.677	2.687	-0.010
IDOB360	-1.934	-1.924	-0.010
IDOB419	-0.527	-0.521	-0.006

6.1.8 Impact of item type on student performance

Percentages of students omitting responses by item type and gender are shown in Table 6.6. It can be seen that the omit rate for extended response items is higher than that of short answer items. Similarly, the omit rate for short answer items is higher than that of multiple-choice items. These trends are consistent across jurisdictions.

Table 6.6 Percentages of students omitting responses by item type

State / Territory	Gender	Item type and per cent omits		
		Multiple-choice (MC)	Short answer (SA)	Extended response (ER)
ACT	Females	2.04	5.14	7.50
	Males	2.35	6.51	8.80
NSW	Females	2.17	5.21	7.63
	Males	2.49	5.64	9.27
NT	Females	6.37	11.30	15.82
	Males	6.40	13.66	18.26
QLD	Females	3.68	5.99	7.90
	Males	3.75	7.26	9.45
SA	Females	2.20	5.56	8.83
	Males	2.39	5.82	10.09
TAS	Females	2.77	5.57	8.44
	Males	3.20	7.54	11.25
VIC	Females	2.79	6.11	8.68
	Males	2.96	7.05	10.35
WA	Females	2.51	5.57	8.10
	Males	2.79	6.59	10.83
AUST	Females	2.83	5.93	8.61
	Males	3.03	6.93	10.40

It is interesting to note that the omit rates for extended response items appear to be higher for male students than female students.

6.2 Test design

6.2.1 Sample test design

Each test booklet contained an objective test and two practical tasks. Students were only required to complete the objective test and one of the two practical tasks. The objective tests were made up of item sets grouped into clusters. Each cluster appeared in three of the seven test booklets – once at the beginning of the paper (Block 1), once in the middle (Block 2) and once at the end of the paper (Block 3). The following table shows how each item was arranged within the booklets.

Table 6.7 List of item codes and details

Item Label	Paper	Block 1	Block 2	Block 3	SRM question number	Unit title
PAQ01	Practical	AQ01				
PAQ02	Practical	AQ02				
PAQ03	Practical	AQ03				
PAQ04	Practical	AQ04				
PAQ05	Practical	AQ05				
PAQ06	Practical	AQ06				
PAQ07	Practical	AQ07				
PAQ08	Practical	AQ08				
PAQ09	Practical	AQ09				
PAQ10	Practical	AQ10				
PAQ11	Practical	AQ11				
PBQ01	Practical	BQ01			P1	Reaction time
PBQ02	Practical	BQ02			P2	Reaction time
PBQ03	Practical	BQ03			P3	Reaction time
PBQ04	Practical	BQ04			P4	Reaction time
PBQ05	Practical	BQ05			P5	Reaction time
PBQ06	Practical	BQ06			P6	Reaction time
PBQ07	Practical	BQ07			P7	Reaction time
PBQ08	Practical	BQ08			P8	Reaction time
PBQ09	Practical	BQ09			P9	Reaction time
PBQ10	Practical	BQ10			P10	Reaction time
IDOB532	Objective	B4Q26	B6Q15	B7Q01		
IDOB533	Objective	B4Q27	B6Q16	B7Q02		
IDOB534	Objective	B4Q28	B6Q17	B7Q03		
IDOB483	Objective	B4Q29	B6Q18	B7Q04	21	Evaporating liquids
IDOB484	Objective	B4Q30	B6Q19	B7Q05	22	Evaporating liquids
IDOB486	Objective	B4Q31	B6Q20	B7Q06	23	Evaporating liquids
IDOB487	Objective	B4Q32	B6Q21	B7Q07	24	Evaporating liquids
IDOB489	Objective	B4Q33	B6Q22	B7Q08	25	Evaporating liquids
IDOB490	Objective	B4Q34	B6Q23	B7Q09	26	Evaporating liquids
IDOB491	Objective	B4Q35	B6Q24	B7Q10	27	Evaporating liquids
DOB470	Objective	B4Q36	B6Q25	B7Q11		
IDOB471	Objective	B4Q37	B6Q26	B7Q12		
IDOB473	Objective	B4Q38	B6Q27	B7Q13		
IDOB097	Objective	B1Q01	B5Q27	B7Q14		
IDOB098	Objective	B1Q02	B5Q28	B7Q15		
IDOB149	Objective	B1Q03	B5Q29	B7Q16		
IDOB150	Objective	B1Q04	B5Q30	B7Q17		
IDOB021	Objective	B1Q05	B5Q31	B7Q18		
IDOB022	Objective	B1Q06	B5Q32	B7Q19		
IDOB023	Objective	B1Q07	B5Q33	B7Q20		
IDOB177	Objective	B1Q08	B5Q34	B7Q21		
IDOB178	Objective	B1Q09	B5Q35	B7Q22		
IDOB084	Objective	B1Q10	B5Q36	B7Q23		
IDOB085	Objective	B1Q11	B5Q37	B7Q24		
IDOB086	Objective	B1Q12	B5Q38	B7Q25		

Table 6.7 (Cont.) List of item codes and details

Item Label	Paper	Block 1	Block 2	Block 3	SRM question number	Unit title
IDOB087	Objective	B1Q13a	B5Q39a	B7Q26a		
IDOB088	Objective	B1Q13b	B5Q39b	B7Q26b		
IDOB457	Objective	B2Q12	B3Q01	B7Q27	28	Seed dispersal
IDOB458	Objective	B2Q13	B3Q02	B7Q28	29	Seed dispersal
IDOB459	Objective	B2Q14	B3Q03	B7Q29	30	Seed dispersal
IDOB460	Objective	B2Q15	B3Q04	B7Q30	31	Seed dispersal
IDOB492	Objective	B2Q16	B3Q05	B7Q31	9	Food and energy
IDOB493	Objective	B2Q17	B3Q06	B7Q32	10	Food and energy
IDOB494	Objective	B2Q18	B3Q07	B7Q33	11	Food and energy
IDOB496	Objective	B2Q19	B3Q08	B7Q34	12	Food and energy
IDOB564	Objective	B2Q20	B3Q09	B7Q35		
IDOB565	Objective	B2Q21	B3Q10	B7Q36		
IDOB566	Objective	B2Q22	B3Q11	B7Q37		
IDOB568	Objective	B2Q23	B3Q12	B7Q38		
IDOB569	Objective	B2Q24	B3Q13	B7Q39		
IDOB570	Objective	B2Q25	B3Q14	B7Q40		
IDOB559	Objective	B3Q28	B5Q13	B6Q01		
IDOB561	Objective	B3Q29	B5Q14	B6Q02		
IDOB562	Objective	B3Q30	B5Q15	B6Q03		
IDOB563	Objective	B3Q31	B5Q16	B6Q04		
IDOB451	Objective	B3Q32	B5Q17	B6Q05	3	Light and shadows
IDOB452	Objective	B3Q33	B5Q18	B6Q06	4	Light and shadows
IDOB453	Objective	B3Q34	B5Q19	B6Q07	5	Light and shadows
IDOB454	Objective	B3Q35	B5Q20	B6Q08	6	Light and shadows
IDOB551	Objective	B3Q36	B5Q21	B6Q09		
IDOB552	Objective	B3Q37	B5Q22	B6Q10		
IDOB553	Objective	B3Q38	B5Q23	B6Q11		
IDOB554	Objective	B3Q39	B5Q24	B6Q12		
IDOB503	Objective	B3Q40	B5Q25	B6Q13		
IDOB506	Objective	B3Q41	B5Q26	B6Q14		
IDOB041	Objective	B1Q14	B2Q01	B6Q28		
IDOB044	Objective	B1Q15	B2Q02	B6Q29		
IDOB173	Objective	B1Q16	B2Q03	B6Q30		
IDOB474	Objective	B1Q17	B2Q04	B6Q31	1	Life in the desert
IDOB475	Objective	B1Q18	B2Q05	B6Q32	2	Life in the desert
IDOB184	Objective	B1Q19	B2Q06	B6Q33	15	Curtains
IDOB185	Objective	B1Q20	B2Q07	B6Q34	16	Curtains
IDOB186	Objective	B1Q21	B2Q08	B6Q35	17	Curtains
IDOB368	Objective	B1Q22	B2Q09	B6Q36	32	Recycling
IDOB370	Objective	B1Q23	B2Q10	B6Q37	33	Recycling
IDOB371	Objective	B1Q24	B2Q11	B6Q38	34	Recycling
IDOB461	Objective	B2Q26	B4Q14	B5Q01	7	Mixing liquids
IDOB462	Objective	B2Q27	B4Q15	B5Q02	8	Mixing liquids
IDOB517	Objective	B2Q28	B4Q16	B5Q03		
IDOB518	Objective	B2Q29	B4Q17	B5Q04		

Table 6.7 (Cont.) List of item codes and details

Item Label	Paper	Block 1	Block 2	Block 3	SRM question number	Unit title
IDOB519	Objective	B2Q30	B4Q18	B5Q05		
IDOB521	Objective	B2Q31	B4Q19	B5Q06		
IDOB522	Objective	B2Q32	B4Q20	B5Q07		
IDOB529	Objective	B2Q33	B4Q21	B5Q08	35	Changing rocks
IDOB530	Objective	B2Q34	B4Q22	B5Q09	36	Changing rocks
IDOB531	Objective	B2Q35	B4Q23	B5Q10	37	Changing rocks
IDOB444	Objective	B2Q36	B4Q24	B5Q11	38	Solar energy
IDOB446	Objective	B2Q37	B4Q25	B5Q12	39	Solar energy
IDOB417	Objective	B1Q25	B3Q15	B4Q01	18	Making jelly
IDOB418	Objective	B1Q26	B3Q16	B4Q02	19	Making jelly
IDOB419	Objective	B1Q27	B3Q17	B4Q03	20	Making jelly
IDOB360	Objective	B1Q28	B3Q18	B4Q04		
IDOB363	Objective	B1Q29	B3Q19	B4Q05		
IDOB362	Objective	B1Q30	B3Q20	B4Q06		
IDOB308	Objective	B1Q31	B3Q21	B4Q07		
IDOB309	Objective	B1Q32	B3Q22	B4Q08		
IDOB310	Objective	B1Q33	B3Q23	B4Q09		
IDOB313	Objective	B1Q34	B3Q24	B4Q10	13	Testing paper towels
IDOB315	Objective	B1Q35	B3Q25	B4Q11	14	Testing paper towels
IDOB405	Objective	B1Q36	B3Q26	B4Q12		
IDOB406	Objective	B1Q37	B3Q27	B4Q13		

Note: Unit titles are shown for items which appear in the 2012 School Release Materials (SRM) only. To maintain security of future link items, all other unit titles have not been listed.

6.3 Item analysis files

Access to the data files and output from the analyses can be made available to researchers or future contractors who want to replicate procedures on application for approval to ACARA at datarequest@acara.edu.au. Relevant data files are listed throughout this Report.

Chapter 7
Scaling of Test Data

7.1 Overview

The process of scaling refers to the estimation of student achievement distributions using information from students' responses to the test items. In the National Assessment Program – Science Literacy (NAP–SL), the scaling process involved two separate phases. Firstly, the item parameters were calibrated using a sample of the data. These item parameters were used as the basis for equating the 2012 results to the 2006 scale. Secondly, student Proficiency Levels were then calculated based on the full dataset.

7.1.1 Calibration of item parameters

The calibration of item parameters used a calibration sample which included equal numbers of respondents from each jurisdiction. Section 7.2 includes information on the selection of the calibration sample and the methodology for the calibration of item parameters.

7.1.2 Estimating student Proficiency Levels and producing plausible values

Once item parameters had been determined, student Proficiency Levels were estimated. As the main purpose of the study is to obtain profiles of student achievement at the population level, rather than at the individual student level, a methodology using plausible values (Wu, 2005) was adopted. The following sections describe in detail the two phases of the scaling process.

7.2 Calibration sample

7.2.1 Overview

To estimate item difficulty parameters, a subset of the responses, called the calibration sample, was used to ensure that each jurisdiction had an equal representation in the sample so that the larger states did not unduly influence the item parameter values. Since NT had the smallest number of responses, all 738 responses were included in the calibration sample. For each of the other jurisdictions, a random sample of 738 responses was selected. Consequently, the calibration sample consisted of 5904 (=738×8) responses.

7.2.2 Data files availability

Access to the data files and output from the analyses is available under specific circumstances on application to ACARA at datarequest@acara.edu.au.

7.2.2.1 *CalibrationSample.sav*

The file *CalibrationSample.sav* contains student background variables as well as item responses.

The variables with prefix 'IDOB' (e.g. IDOB177) are students' raw item responses, recoded with A, B, 9 and M. The following rules apply to the recoding:

- For the pencil-and-paper test, the first 'not reached' item is coded as 'A' with the remaining 'not reached' items as 'B', and embedded missing responses remain as '9'. Students with no responses at all for the whole test have responses recoded to 'M'.

- For the practical task, students with no responses at all have responses recoded to 'M'. Missing responses, whether not-reached or embedded, are recoded to '9'. That is, there are no 'A' and 'B' codes. As the two practical tests have only 11 and 10 items respectively, there does not appear to be a large number of clearly 'not reached' items at the end.

- To calibrate the item parameters, response codes 'A' and '9' are treated as incorrect, whereas response codes 'B' and 'M' are treated as non-administered (i.e. as missing data).

- In contrast, to calibrate the student abilities in subsequent analyses, response code 'M' is treated as not-administered, but response codes 'A', '9' and 'B' are treated as incorrect.

7.2.2.2 *CalibrationItems.dat*

This ASCII (or text) file is used as input to IRT software to calibrate the item parameters. The codebook for the relevant data fields in the text file is given below:

Table 7.1 Codebook for *CalibrationItems.dat*

Field	Column range	Description
Booklet ID	8	Unique identifier for the student record
Item responses	10 to 121 (112 items in total)	Student responses

7.2.3 IRT analysis for calibrating item parameters

The software program used to carry out the calibration of item parameters is ConQuest. A facets model is used where the test booklet number is regarded as a facet. More specifically, the model statement used in ConQuest is:

*bookid + item + item*step*

The full syntax of ConQuest commands is in the control file *CalibrationSample.cqc*

The use of the term 'bookid' in ConQuest model statements is to ensure that the estimation of the item parameters takes into account the so-called 'booklet effect' (OECD 2012, p. 141). However, as there is only one domain in the 2012 NAP–SL (unlike PISA where there are three domains: mathematics, science and reading) and all items are calibrated together, it is not expected that there will be a significant booklet effect, as is shown later in the results of the item analysis.

Three output files are produced from ConQuest:

CalibrationSample.shw

This is a summary file, showing booklet and item parameter values, population parameter estimated and item–person maps.

CalibrationSample.itn

This file is known as the 'itanal', showing classical test statistics as well as IRT statistics for each item.

CalibrationAnchor.anc

This file is produced through an Export statement in ConQuest. It contains the values of the parameters that can be used as anchor values later when student abilities are estimated.

Once the calibrated item parameters are obtained, the transformation equations used to equate the 2012 results to the 2006 scale are then derived. Details of the equating process can be found in Chapter 8 of this report.

7.3 Estimating student Proficiency Levels and producing plausible values

In this phase, student Proficiency Levels are estimated for the full data set (*NAPSL2012_PV_2013-03-11.sav* See Appendix 7 for descriptions of variables).

The scaling model used is a one-parameter item response model with conditioning variables in the population latent regression model. See the PISA 2009 Technical Report for a description of the measurement model (OECD 2012).

The conditioning variables included are:

- School mean proficiency (average of students' weighted likelihood estimates for each school)
- State or territory
- Sector
- Gender
- Indigenous status
- Geographic location
- Language background.

To prepare the data to be used as conditioning variables, two separate steps are taken:

Step A: Produce a weighted likelihood estimate (WLE) for each student in the full data set, and compute the average WLE for each school. RUMM2020 and ConQuest were used for the estimation of WLE estimates, with item parameters anchored at values from the item calibration phase.

Step B: Dummy variables are created for State or territory, Sector, Gender, Indigenous status, Geographic location and Language background.

7.3.1 Production of plausible values

The software program ConQuest is used for the scaling of student Proficiency Levels and the generation of plausible values. Note that Case Weight is used in this analysis. Both booklet parameters and item parameters are anchored. Both embedded-missing (code '9') and not-reached items (codes 'A' and 'B') are treated as incorrect. If a test has no valid responses from a student, the responses (code 'M') are treated as not-administered. Ten plausible values are generated (instead of the usual five).

The ConQuest control file used is *NAPSL2012_Produce_2012_PV.cqc* which is shown in Appendix 8.

7.4 Estimation of statistics of interest and their standard errors

Once the plausible values are produced for each student, statistics of interest can be computed together with their standard errors. For example, the mean achievement level in science literacy for Year 6 students in Australia can be estimated, as well as jurisdiction average achievement levels. The estimates will also have associated standard errors to indicate the confidence which we have about the results.

The plausible-values methodology has been used for large-scale studies such as TIMSS, PISA and NAEP. In the 2012 NAP–SL, this methodology was also used for the estimation of statistics and standard errors. For a detailed description of the methodology, see Mislevy, Beaton, Kaplan and Sheehan (1992), and Beaton and Gonzalez (1995).

Briefly, the methodology is summarised below. The plausible values for each student show the indicative level of the student's achievement. So the estimate for a population statistic is computed using the plausible values as if they represent each student's level of achievement. For example, to compute the estimated mean of the population, take the first plausible value for each student and compute the average across students, weighted by the sampling weight (student final weight). Repeat the process with all ten plausible values, and then average the ten estimated means for the ten runs. Similarly, for the estimation of percentiles and percentages in levels, plausible values are used in the same way.

The standard errors associated with the estimated statistics are not straightforward to compute, as the sampling method is not simple random sampling but a complex two-stage sampling. Typically, for complex sampling such as the one used for the 2012 NAP–SL, replication methods such as Balanced Repeated Replicate (BRR) or Jackknife are used to compute standard errors (Rust & Rao 1996). In the 2012 NAP–SL, the Jackknife method was used. Jackknife replication weights are computed (variables RW1 to RW318 in the file *NAPSL2012_PV_2013-03-11.sav*).

The statistic of interest is computed using each of the replicate weights in turn. The variations in the estimated statistic obtained from using different replicate weights contribute to the estimate of the sampling variance for the estimated statistic. Combining this sampling variance with the variance from using the ten plausible values (the measurement error) provides an estimate of the standard error for the estimated statistic. SPSS macros were written to carry out the procedures of the estimation of statistics and their standard errors.

7.5 Transform 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 the 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 Proficiency Levels in terms of how far a score is from the mean.

As part of the equating process (refer to Chapter 8 for details), the 2012 logit scores are first translated to the 2006 scale, then transformed to the 400/100 scale. The transformation used in 2012 is given below.

Score on proficiency scale = (Logit-0.200543797)/0.954513216*100+400

where 'Logit' refers to a logit score on the 2006 scale. The conversion of 2012 logit scores to the 2006 scale is detailed in section 8.3.

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 take into account the number of students in each jurisdiction. In summary, house weights are used to set the average score of 400, not senate weights.

Chapter 8
Equating 2012 Results to 2006 Results

8.1 Setting 2006 results as the baseline

While the first cycle of the National Assessment Program – Science Literacy (NAP–SL) was conducted in 2003 (then known as PSAP) and the 2006 assessment was the second round of NAP–SL, it was decided that the 2006 assessment be used to set the scale of a mean of 400 and a standard deviation of 100 instead of the 2003 assessment. The reasons for this decision are summarised below.

(1) The 2006 assessment test design was more robust than the 2003 test design. In 2006, a balanced incomplete block (BIB) test design consisting of seven test booklets was used. In contrast, in 2003 only two test booklets were used, resulting in item-position effect for most items.

(2) There were considerably more items in 2006 than in 2003, resulting in a better coverage of the science literacy content in 2006. In 2006, 110 items were included in the final test, while only 72 items were included in the 2003 test.

(3) The 2006 assessment produced a much higher population variance in achievement than 2003 did. In logits, the 2006 population standard deviation was 0.95, while the 2003 population standard deviation was 0.78. This could be an indication that:

- the 2006 items were generally more discriminating than the 2003 items; that is, the 2006 items were higher quality items
- the 2006 sampling was more comprehensive, as remote schools were also included in the sample, while the 2003 sampling focused only on areas where students were generally well-resourced.

8.2 Equating 2012 results to 2006 results

As a consequence of the decision to use the 2006 results as the baseline, the 2012 results were equated to 2006 results. To carry out the equating, link items between the 2012 and 2006 assessments were used.

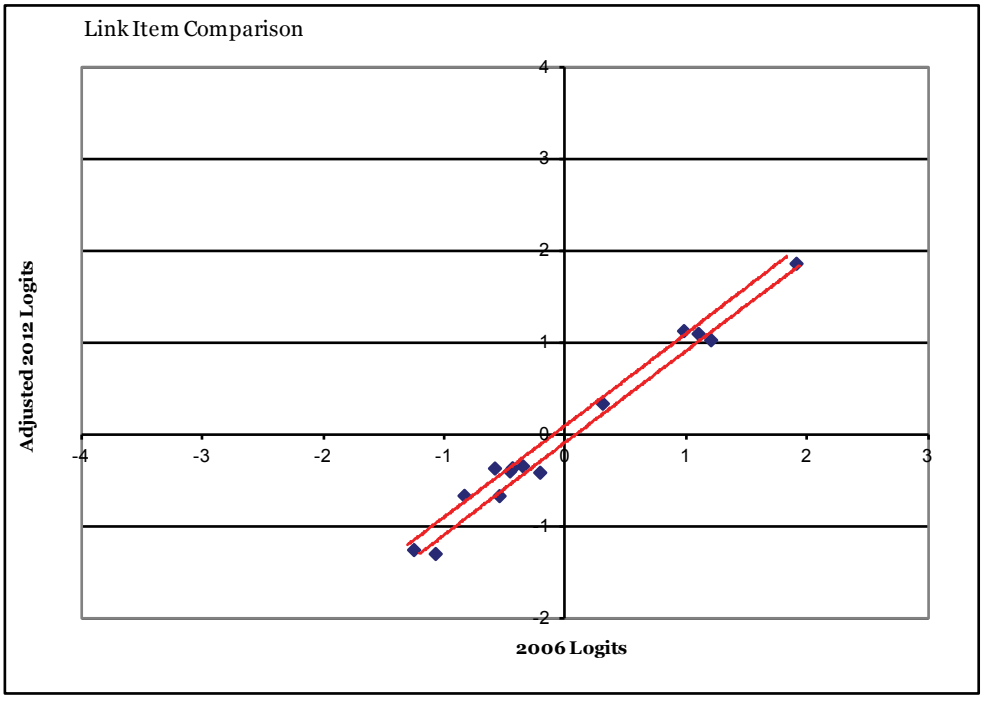
This included six link items from the 2003 assessment and 14 link items from the 2006 assessment. Care was taken to find items that performed well psychometrically and also covered the range of scientific literacy strands A, B and C.

8.2.1 Link item selection

The selection process for the final set of link items to conduct equating between the 2012 and 2006 assessments consisted of two parts. In the first part, the list of items was refined based on the comparisons of item locations in 2012 and 2006. In the second part, the final set of items was inspected by a content expert from the test construction team in order to ensure that the selected link items provided adequate coverage across the scientific literacy strands.

In the first part of the link item selection process, the 2012 location of link items was independently estimated. In order to conduct comparisons of item locations between 2006 and 2012, the 2012 locations were adjusted to have the same mean and standard deviation as observed in 2006. In the first refinement step one clear outlier in terms of overall item difficulty was removed from the set. In the second step, all items with an absolute difference between the 2006 location and the 2012 location greater than 0.3 logits were removed. After readjusting the 2012 location, the second step was repeated with logits of 0.3 cut-off changed to 0.2 logits. After this step the final item pool was identified. The final set had 14 link items consisting of three 2003 link items and 11 2006 link items. A plot of 2006 and 2012 item difficulties for the final link items, including graphical representation of 95 per cent confidence interval for the statistical difference between item locations, is given in Figure 8.1.

Figure 8.1 Calibrated item difficulties in 2006 and 2012 for the final link item set



8.2.2 Equating procedures

The 2012 data were scaled and item parameters were obtained. Using the 2006 item parameters as anchors for common items, the 2012 data were scaled and population parameters (mean and variance of the ability distributions for 2012) were produced. The mean and variance from this scaling and the mean and variance from the 2012 scaling (i.e. scaling 2012 data using 2012 item parameters) were then compared. A transformation was derived from mapping the mean and variance of the 2012 ability distribution obtained using 2012 item parameters onto the mean and variance of the 2012 ability distribution obtained using 2006 item parameters. This transformation was used to place the 2012 results directly onto the 2006 scale.

8.3 Equating transformation

The result of the equating process was the derivation of a transformation formula for the 2012 results to be placed on the 2006 scale. This equation is given below.

2012 on 2006 scale = ((2012logit-(-0.3278))/0.9524)* 0.9443+0.1620

The scale factor is very close to 1, indicating that an adjustment of the scale factor is not really necessary.

For standard errors, the transformation involved only the scale factor, as follows:

2012 standard error on 2006 scale in logit = ((2012 S.E. in logit) / 0.9524) * 0.9443

8.4 Link error

In establishing trends from 2006 to 2012, it is necessary to make judgments about the statistical significance of the difference in science literacy achievement between 2012 and 2006. An appropriate estimation of the magnitude of equating errors is important when trends are reported. An underestimate of the equating errors will often result in erroneous claims of change in achievement levels when there is no significant difference.

Equating errors come from at least two sources: the sampling of students and the sampling of items. Equating errors due to the sampling of students affect the accuracy with which the item parameters are estimated, and the magnitude of these errors diminishes when the sample size increases. However, equating errors due to the sampling of items have not often been taken into account, and the magnitude of these errors does not diminish when the sample size increases. For the estimates of population parameters (e.g. mean), the magnitude of equating errors due to the

sampling of items tends to be much larger than the magnitude of equating errors due to the sampling of students. Consequently, it is important to estimate the equating error due to the sampling of items.

Equating error (called 'link error' in PISA) is computed following the approach used in PISA 2009 (OECD 2012). Firstly calibrate the items using 2012, 2009 and 2006 data separately. If the link items behave exactly the same way in 2012, 2009 and 2006 (and they follow the Rasch model), there should only be a constant difference between 2012, 2009 and 2006 item parameters for matched items. However, in real life, items will vary across assessment cycles and some items will vary more than others.

The link error for comparison between 2009 and 2012 is 0.0308884 logits; transformed to the scientific literacy scale it is equal to a scaled score of 3.24. Similarly, the link error for comparison between 2006 and 2012 is 0.0388736 logits; transformed to the scientific literacy scale it is equal to a scaled score of 4.07.

Additional information about the computation of link error can be found in a data CD available to researchers or future contractors on application for approval to ACARA at datarequest@acara.edu.au

The link error is used only when comparisons across 2012, 2009 and 2006 results are made. For example, to test whether the mean achievement in 2012 differs from the mean achievement in 2006, the link error is added to the standard error of the difference, as illustrated in Table 8.1.

Table 8.1 Example of link error application in calculating standard error of difference

	2012 Mean on 2006 scale & S.E.	2006 Mean & S.E.	2012 Mean – 2006 Mean	Standard error of difference	Standardised difference
NSW	395 (5.06)	411 (6.38)	-16	SORT(5.06^2 + 6.38^2 + 4.07^2)	-1.76 = -16/9.10 (not significant)

Chapter 9
Scale and Proficiency Levels

For reporting purposes, student results are often summarised through the definition of a number of Proficiency Levels. That is, the proficiency scale is divided into a number of levels, with descriptions of skills attached to each level, and percentages of students at various levels are reported.

9.1 Proficiency Level cut-off points

In the 2003 assessment cycle, cut-off points along the proficiency scale were decided after consultations with science experts. In 2012, the 2003 cut-off points transformed to the 2006 scale were used (see Table 9.1). For further details, refer to the 2006 NAP–SL Technical Report, p. 73.

Table 9.1 Cut-off points for the 2012 NAP–SL

Level	2006 cut-off points (logit)	Transformed to 400/100 scale
2 and below	<–1.114	262.293
3.1	0.130	392.577
3.2	1.373	522.861
3.3	2.617	653.145
4.0	>2.617	>653.145

As for 2003, 2006 and 2009, a response probability (RP) of 0.65 is used to place items in Proficiency Levels. The RP adjustment refers to '…the probability that a student in the middle of a level would correctly answer an item of average difficulty for that level'. (OECD 2000, p. 198)

9.2 Proficiency Levels of items

Table 9.2 shows the 2012 NAP–SL items and their corresponding levels on the proficiency scale.

Table 9.2 Proficiency Levels of items

Item Label	2012 difficulty	2012 item difficulty after adjustment for RP	Operational Level	Design level	Scaled score
PAQ01	3.225	4.306	4 and above	3/4	830
PAQ02	0.902	2.003	3.3	3	589
PAQ03	-1.130	-0.011	3.1	3	378
PAQ04	-0.387	0.725	3.2	4	455
PAQ05	2.053	3.144	4 and above	3	708
PAQ06	1.179	2.278	3.3	3	618
PAQ07	1.192	2.291	3.3	3/4	619
PAQ08	1.616	2.711	4 and above	3	663
PAQ09	1.094	2.194	3.3	3	609
PAQ10	-0.965	0.152	3.2	3	395
PAQ11	-0.874	0.242	3.2	3	404
PBQ01	-2.219	-1.091	3.1	2	265
PBQ02	2.683	3.769	4 and above	3	774
PBQ03	-0.655	0.460	3.2	4	427
PBQ04	-1.249	-0.129	3.1	3	365
PBQ05	-0.038	1.071	3.2	3	491
PBQ06	3.040	4.123	4 and above	3	811
PBQ07	-2.135	-1.008	3.1	2	273
PBQ08	-2.088	-0.961	3.1	3	278
PBQ09	-0.219	0.892	3.2	3	472
PBQ10	3.501	4.580	4 and above	4	859
IDOB532	1.089	2.189	3.3	3	608
IDOB533	1.368	2.465	3.3	3	637
IDOB534	1.660	2.755	4 and above	3	668
IDOB483	0.577	1.681	3.3	2	555
IDOB484	0.240	1.347	3.2	3	520
IDOB486	-0.880	0.236	3.2	3	404
IDOB487	3.400	4.480	4 and above	3	848
IDOB489	-0.308	0.804	3.2	2	463
IDOB490	-0.291	0.820	3.2	3	465
IDOB491	-0.449	0.664	3.2	3	449
IDOB470	1.601	2.696	4 and above	3	661
IDOB471	0.192	1.299	3.2	4	515
IDOB473	0.754	1.857	3.3	3	573
IDOB097	-1.835	-0.710	3.1	2	305
IDOB098	-0.586	0.528	3.2	4	434
IDOB149	-1.216	-0.097	3.1	2	369
IDOB150	-1.285	-0.165	3.1	3	362
IDOB021	0.666	1.769	3.3	3	564
IDOB022	-0.952	0.165	3.2	4	396

Table 9.2 (Cont.) Proficiency Levels of items

Item Label	2012 difficulty	2012 item difficulty after adjustment for RP	Operational Level	Design level	Scaled score
IDOB023	0.561	1.665	3.3	5	553
IDOB177	0.514	1.619	3.3	4	549
IDOB178	0.634	1.738	3.3	3	561
IDOB084	-0.937	0.180	3.2	3	398
IDOB085	-0.879	0.237	3.2	3	404
IDOB086	1.435	2.532	3.3	3	644
IDOB087	-0.472	0.641	3.2	4	446
IDOB088	-0.904	0.213	3.2	4	401
IDOB457	-0.123	0.987	3.2	2	482
IDOB458	0.793	1.895	3.3	3	578
IDOB459	0.005	1.114	3.2	3	496
IDOB460	0.040	1.149	3.2	3	499
IDOB492	-1.979	-0.853	3.1	2	290
IDOB493	-0.085	1.025	3.2	3	486
IDOB494	0.445	1.550	3.3	4	541
IDOB496	-0.783	0.333	3.2	3	414
IDOB564	0.049	1.158	3.2	3/4	500
IDOB565	-1.779	-0.655	3.1	4	310
IDOB566	1.097	2.197	3.3	4	609
IDOB568	1.254	2.352	3.3	4	625
IDOB569	-0.301	0.811	3.2	3	464
IDOB570	1.661	2.756	4 and above	3	668
IDOB559	-1.567	-0.445	3.1	3	332
IDOB561	-0.292	0.819	3.2	3	465
IDOB562	0.899	2.000	3.3	4	589
IDOB563	0.445	1.550	3.3	3	541
IDOB451	-1.736	-0.612	3.1	3	315
IDOB452	-1.734	-0.610	3.1	3	315
IDOB453	0.043	1.152	3.2	2	500
IDOB454	0.453	1.558	3.3	3	542
IDOB551	2.049	3.140	4 and above	3	708
IDOB552	-1.012	0.106	3.1	3	390
IDOB553	-1.835	-0.710	3.1	2	305
IDOB554	1.634	2.729	4 and above	3	665
IDOB503	-1.903	-0.778	3.1	3	298
IDOB506	2.295	3.384	4 and above	3	734
IDOB041	-1.880	-0.755	3.1	3	300
IDOB044	-0.901	0.216	3.2	4	402
IDOB173	-1.219	-0.100	3.1	3	369
IDOB474	-2.589	-1.458	2 and below	4	226
IDOB475	-0.857	0.259	3.2	3	406
IDOB184	0.418	1.523	3.3	4	539
IDOB185	-1.085	0.033	3.1	2	382
IDOB186	-0.164	0.946	3.2	3	478
IDOB368	0.945	2.046	3.3	4	593

Table 9.2 (Cont.) Proficiency Levels of items

Item Label	2012 difficulty	2012 item difficulty after adjustment for RP	Operational Level	Design level	Scaled score
IDOB370	-2.874	-1.740	2 and below	3	197
IDOB371	0.136	1.244	3.2	4	509
IDOB461	-2.034	-0.908	3.1	2	284
IDOB462	-0.055	1.054	3.2	4	489
IDOB517	0.090	1.198	3.2	3	505
IDOB518	-2.221	-1.093	3.1	2	264
IDOB519	-0.647	0.468	3.2	3	428
IDOB521	-0.394	0.718	3.2	4	454
IDOB522	0.096	1.204	3.2	3	505
IDOB529	0.283	1.390	3.3	3	525
IDOB530	1.735	2.829	4 and above	3	675
IDOB531	2.299	3.388	4 and above	3	734
IDOB444	0.935	2.036	3.3	4	592
IDOB446	2.763	3.848	4 and above	3/4	782
IDOB417	0.450	1.555	3.3	3	542
IDOB418	-0.450	0.663	3.2	4	448
IDOB419	-0.510	0.603	3.2	4	442
IDOB360	-1.891	-0.766	3.1	3	299
IDOB363	0.300	1.406	3.3	3	526
IDOB362	-0.417	0.696	3.2	3	452
IDOB308	-1.709	-0.585	3.1	3	318
IDOB309	-0.909	0.208	3.2	4	401
IDOB310	-0.748	0.367	3.2	3	417
IDOB313	0.599	1.703	3.3	3	557
IDOB315	-0.649	0.466	3.2	3	428
IDOB405	1.187	2.286	3.3	3	618
IDOB406	1.718	2.812	4 and above	4	674

References

Ball, S., Rae, I., & Tognolini, J. (2000). Options for the assessment and reporting of primary students in the key learning area of science to be used for the reporting of nationally comparable outcomes of schooling within the context of the National Goals for Schooling in the Twenty-First Century: National Education Performance Monitoring Taskforce.

Beaton, A.E., & Gonzalez, E. (1995). NAEP primer. Chestnut Hill, MA: Boston College.

Biggs, J., & Collis, K.F. (1982). Evaluating the quality of learning: The SOLO taxonomy. New York: Academic Press.

Goodrum, D., Hackling, M.W., & Rennie, L. (2001). The status and quality of teaching and learning of science in Australian schools. Canberra: Department of Education, Training and Youth Affairs.

IEA (2009). TIMSS 2007 Technical Report. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Laugksch, R.C. (2000). Scientific literacy: A conceptual overview. Science Education, 84(1), 71–94.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 131–154.

OECD (1999). Measuring student knowledge and skills: A new framework for assessment. OECD: Paris.

OECD (2000). PISA 2000 Technical Report. OECD: Paris.

OECD (2012). PISA 2009 Technical Report. OECD: Paris.

Rust, K.F., & Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research, 5, 283–310.

Wu, M.L. (2005). The role of plausible values in large-scale surveys. Studies in Educational Evaluation, 31, 114–128.

Appendix 1
National Year 6 Primary Science Assessment Domain

Assessment strands: Scientific literacy

The national review of the status and quality of teaching and learning of science in Australian schools (Goodrum, Hackling & Rennie 2001) argued that the broad purpose of science in the compulsory years of schooling is to develop scientific literacy for all students.

Scientific literacy is a high priority for all citizens, helping them to:

- be interested in and understand the world around them
- engage in discourses of and about science
- be sceptical and questioning of claims made by others about scientific matters
- be able to identify questions, investigate and draw evidence-based conclusions
- make informed decisions about the environment and their own health and wellbeing.

Scientific literacy is important because it contributes to the economic and social wellbeing of the nation and improved decision-making at public and personal levels (Laugksch 2000).

The Programme for International Student Assessment (PISA) focuses on aspects of preparedness for adult life in terms of functional knowledge and skills that allow citizens to participate actively in society. It is argued that scientifically literate people are 'able to use scientific knowledge and processes not just to understand the natural world but also to participate in decisions that affect it' (OECD 1999, p. 13).

The OECD–PISA defined scientific literacy as:

> *... the capacity to use scientific knowledge, to identify questions (investigate)*[1] *and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity.*
>
> (OECD 1999, p. 60)

1 Because of the constraints of large-scale testing, PISA was not able to include performance tasks such as conducting investigations. Consequently, its definition of scientific literacy omitted reference to investigating. The word 'investigate' was inserted into the definition for the purposes of NAP–SL, as the sample testing methodology allowed for assessments of students' ability to conduct investigations.

This definition has been adopted for the National Assessment Program – Science Literacy (NAP–SL) in accordance with the Ball et al. 2000 report recommendation.

Scientific literacy: Progress Map

A scientific literacy Progress Map was developed based on the construct of scientific literacy and an analysis of state and territory curriculum and assessment frameworks. The Progress Map describes the development of scientific literacy across three strands of knowledge which are inclusive of Ball et al.'s concepts and processes and the elements of the OECD–PISA definition.

The five elements of scientific literacy, including concepts and processes used in PISA 2000 (OECD 1999), include:

1. demonstrating understanding of scientific concepts
2. recognising scientifically investigable questions
3. identifying evidence needed in a scientific investigation
4. drawing or evaluating conclusions
5. communicating valid conclusions.

These elements have been clustered into three more holistic strands which are described below. The second and third elements and conducting investigations to collect data are encompassed in Strand A; the fourth and fifth elements and conducting investigations to collect and interpret data are included in Strand B; and the first element is included in Strand C.

Strand A: Formulating or identifying investigable questions and hypotheses, planning investigations and collecting evidence.

This process strand includes posing questions or hypotheses for investigation or recognising scientifically investigable questions; planning investigations by identifying variables and devising procedures where variables are controlled; gathering evidence through measurement and observation; and making records of data in the form of descriptions, drawings, tables and graphs using a range of information and communication technologies.

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.

This process strand includes identifying, describing and explaining the patterns and relationships between variables in scientific data; drawing conclusions that are evidence-based and related to the questions or hypotheses posed; critiquing the trustworthiness of evidence and claims made by others; and communicating findings using a range of scientific genres and information and communications technologies.

Strand C: Using science understandings for describing and explaining natural phenomena, and for interpreting reports about phenomena.

This conceptual strand includes demonstrating conceptual understandings by being able to describe, explain and make sense of natural phenomena; understand and interpret reports (e.g. TV documentaries, newspaper or magazine articles or conversations) related to scientific matters; and make decisions about scientific matters in students' own lives which may involve some consideration of social, environmental and economic costs and benefits.

Scientific literacy has been described here in three strands to facilitate the interpretation of student responses to assessment tasks. However, authentic tasks should require students to apply concepts and processes together to address problems set in real-world contexts. These tasks may involve ethical decision-making about scientific matters in students' own lives and some consideration of social, environmental and economic costs and benefits.

The scientific literacy Progress Map (see Table A1.1) describes progression in six levels from 1 to 6 in terms of three aspects:

- increasing complexity, from explanations that involve one aspect to several aspects, through to relationships between aspects of a phenomenon
- progression from explanations that refer to and are limited to directly experienced phenomena (concrete) to explanations that go beyond what can be observed directly and involve abstract scientific concepts (abstract)
- progression from descriptions of 'what' happened in terms of objects and events, to explanations of 'how' it happened in terms of processes, to explanations of 'why' it happened in terms of science concepts.

Strand C has been abstracted and makes no reference to particular science concepts or contexts. As the progression in this strand is based on increasing complexity and abstraction, links have been made to the Structure of Observed Learning Outcomes (SOLO) taxonomy (Biggs & Collis 1982).

The taxonomy was written to describe levels of student responses to assessment tasks. The basic SOLO categories include:

prestructural	no logical response
unistructural	refers to only one aspect
multistructural	refers to several independent aspects
relational	can generalise (describe relationships between aspects) within the given or experienced context
extended abstract	can generalise to situations not experienced.

The three main categories of unistructural, multistructural and relational can also be applied, as cycles of learning, to the four modes of representation:

sensorimotor	the world is understood and represented through motor activity
iconic	the world is represented as internal images
concrete	writing and other symbols are used to represent and describe the experienced world
formal	the world is represented and explained using abstract conceptual systems.

The conceptual strand, Strand C, of the Progress Map therefore makes links to the SOLO categories of concrete unistructural (level 1), concrete multistructural (level 2), concrete relational (level 3), abstract unistructural (level 4), abstract multistructural (level 5) and abstract relational (level 6).

The SOLO levels of performance should not be confused with Piagetian stages of cognitive development. Biggs and Collis (1982, p. 22) explain that the relationship between Piagetian stages and SOLO levels 'is exactly analogous to that between ability and attainment' and that level of performance depends on quality of instruction, motivation to perform, prior knowledge and familiarity with the context. Consequently, performance for a given individual is highly variable and often sub-optimal.

NAP–SL focuses on levels 2, 3 and 4 of the scientific literacy Progress Map, the levels of scientific literacy attained by students in Year 6.

The agreed Proficiency Levels serve to further elaborate the Progress Map. Level 3 is described as 3.1, 3.2, and 3.3. A 'proficient' standard is a challenging level of performance, with students needing to demonstrate more than minimal or elementary skills.

Table A1.1 Scientific Literacy Progress Map

Level	Strands of scientific literacy		
	Strand A Formulating or identifying investigable questions and hypotheses, planning investigations and collecting evidence. Process strand: experimental design and data gathering.	**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. Process strand: interpreting experimental data.	**Strand C** Using understandings for describing and explaining natural phenomena, and for interpreting reports about phenomena. Conceptual strand: applies conceptual understanding.
6	Uses scientific knowledge to formulate questions, hypotheses and predictions and to identify the variables to be changed, measured and controlled. Trials and modifies techniques to enhance reliability of data collection.	Selects graph type and scales that display the data effectively. Conclusions are consistent with the data, explain the patterns and relationships in terms of scientific concepts and principles, and relate to the question, hypothesis or prediction. Critiques the trustworthiness of reported data (e.g. adequate control of variables, sample or consistency of measurements, assumptions made in formulating the methodology), and consistency between data and claims.	Explains complex interactions, systems or relationships using several abstract scientific concepts or principles and the relationships between them. SOLO taxonomy: Abstract relational
5	Formulates scientific questions or hypotheses for testing and plans experiments in which most variables are controlled. Selects equipment that is appropriate and trials measurement procedure to improve techniques and ensure safety. When provided with an experimental design involving multiple independent variables, can identify the questions being investigated.	Conclusions explain the patterns in the data using science concepts, and are consistent with the data. Makes specific suggestions for improving/extending the existing methodology (e.g. controlling an additional variable, changing an aspect of measurement technique). Interprets/compares data from two or more sources. Critiques reports of investigations noting any major flaw in design or inconsistencies in data.	Explains phenomena, or interprets reports about phenomena, using several abstract scientific concepts. SOLO taxonomy: Abstract multistructural
4	Formulates scientific questions, identifies the variable to be changed, the variable to be measured and in addition identifies at least one variable to be controlled. Uses repeated trials or replicates. Collects and records data involving two or more variables.	Calculates averages from repeat trials or replicates, plots line graphs where appropriate. Interprets data from line graph or bar graph. Conclusions summarise and explain the patterns in the science data. Able to make general suggestions for improving an investigation (e.g. make more measurements).	Explains interactions, processes or effects that have been experienced or reported, in terms of a non-observable property or abstract science concept. SOLO taxonomy: Abstract unistructural

Table A1.1 (Cont.) Scientific Literacy Progress Map

Level	Strands of scientific literacy		
	Strand A Formulating or identifying investigable questions and hypotheses, planning investigations and collecting evidence. Process strand: experimental design and data gathering.	**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. Process strand: interpreting experimental data.	**Strand C** Using understandings for describing and explaining natural phenomena, and for interpreting reports about phenomena. Conceptual strand: applies conceptual understanding.
3	Formulates simple scientific questions for testing and makes predictions. Demonstrates awareness of the need for fair testing and appreciates scientific meaning of 'fair testing'. Identifies variable to be changed and/or measured but does not indicate variables to be controlled. Makes simple standard measurements. Records data as tables, diagrams or descriptions.	Displays data as tables or constructs bar graphs when given the variables for each axis. Identifies and summarises patterns in science data in the form of a rule. Recognises the need for improvement to the method. Applies the rule by extrapolating and predicting.	Describes the relationships between individual events (including cause and effect relationships) that have been experienced or reported. Can generalise and apply the rule by predicting future events. SOLO taxonomy: Concrete relational
2	Given a question in a familiar context, identifies that one variable/factor is to be changed (but does not necessarily use the term 'variable' to describe the changed variable). Demonstrates intuitive level of awareness of fair testing. Observes and describes or makes non-standard measurements and limited records of data.	Makes comparisons between objects or events observed. Compares aspects of data in a simple supplied table of results. Can complete simple tables and bar graphs given table column headings or prepared graph axes.	Describes changes to, differences between or properties of objects or events that have been experienced or reported. SOLO taxonomy: Concrete multistructural
1	Responds to the teacher's questions and suggestions, manipulates materials and observes what happens.	Shares observations; tells, acts out or draws what happened. Focuses on one aspect of the data.	Describes (or recognises) one aspect or property of an individual object or event that has been experienced or reported. SOLO taxonomy: Concrete unistructural

Major scientific concepts in NAP–SL

A table of the major scientific concepts found most widely in the various state and territory curriculum documents has been developed to accompany the scientific literacy Progress Map (see Table A1.2).

These major concepts are broad statements of scientific understandings that Year 6 students would be expected to demonstrate. They provided item writers with a specific context in which to assess scientific literacy. An illustrative list of examples for each of the major concepts provides elaboration of these broad conceptual statements and, in conjunction with the scientific literacy Progress Map which describes the typical developmental stages for scientific literacy, was used as a guide for the development of assessment items.

It should be noted that, because the NAP–SL test instruments are constructed within the constraints of test length, it is not feasible to include all the listed concepts in instruments constructed for a single testing cycle.

Table A1.2 Major scientific concepts in NAP–SL

Major scientific concepts	Examples
Earth and Space	
Earth, sky and people: Our lives depend on air, water and materials from the ground; the ways we live depend on landscape, weather and climate.	Features of weather, soil and sky, and effects on me.
	People use resources from the Earth; need to use them wisely.
	Sustainability.
The changing Earth: The Earth is composed of materials that are altered by forces within and upon its surface.	Changes in weather, weather data, seasons, soil landscape and sky (e.g. Moon phases, weathering and erosion, movement of the Sun and shadows, bush fires, land clearing).
	Climate change.
Our place in space: The Earth and life on Earth are part of an immense system called the universe.	Rotation of the Earth and night/day, spatial relationships between Sun, Earth and Moon.
	Planets of our solar system and their characteristics.
	Space exploration and new developments.
Energy and Force	
Energy and us: Energy is vital to our existence and our quality of life as individuals and as a society.	Uses of energy, patterns of energy use and variations with time of day and season.
	Energy sources, renewable and non-renewable.
Transferring energy: Interaction and change involve energy transfers; control of energy transfer enables particular changes to be achieved.	Sources, transfers, carriers and receivers of energy, energy and change.
	Types of energy, energy of motion – toys and other simple machines – light, sound.
Energy sources and receivers: Observed change in an object or system is indicated by the form and amount of energy transferred to or from it.	Forces as pushes and pulls, magnetic attraction and repulsion.
Living Things	
Living together: Organisms in a particular environment are interdependent.	Living vs non-living.
	Plant vs animal and major groups.
	Dependence on the environment: Survival needs – food, space and shelter.
	Interactions between organisms and interdependence (e.g. simple food chains).
Structure and function: Living things can be understood in terms of functional units and systems.	Major structures and systems and their functions.
	Healthy lifestyle, diet and exercise.
Biodiversity, change and continuity: Life on Earth has a history of change and disruption, yet continues generation to generation.	Change over lifetime, reproduction and lifecycles.
	Adaptation to physical environment.
Matter	
Materials and their uses: The properties of materials determine their uses; properties can be modified.	Materials have different properties and uses.
	Processing materials to make useful things produces waste, use of alternative materials to better care for the environment.
	Waste reduction – recycling.
	Nanotechnology.
Structure and properties: The substructure of materials determines their behaviour and properties.	The properties of materials can be explained in terms of their visible substructure, such as fibres.
	Materials can change their state and properties.
Reactions and change: Patterns of interaction of materials enable us to understand and control those interactions.	Solids, liquids and gases.

Appendix 2
Sample School Reports

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

UNSW Global THE UNIVERSITY OF NEW SOUTH WALES SYDNEY • AUSTRALIA

7 December 2012

Principal Name
Sample School
Sample Lane
Sampleville SAMPLE 9998

Dear Principal Name

Re: 2012 National Assessment Program – Science Literacy (NAP-SL)

On behalf of the Australian Curriculum, Assessment and Reporting Authority (ACARA), I wish to thank you, your staff and the Year 6 students for participating in the 2012 National Assessment Program – Science Literacy in October this year.

We appreciate the effort your staff made to ensure that the assessment was administered consistently, completed and returned to us.

Enclosed with this letter is a report for your school outlining the performance of participating Year 6 students. There are two sets of results for each student: one for the objective (pencil and paper) test and one for the practical task.

There are seven A4 report sections – one for each of the seven test booklets used in the assessment. The results for each student for the objective (pencil and paper) test are located on the A4 report sheet(s) corresponding to the objective test booklet they completed. The students' results for the practical task are located on the A3 report sheet(s). All participating students at your school performed the same practical task.

We have included an information sheet to help you interpret your school's report. Please provide a copy of this information sheet to anyone at your school reviewing the report.

Please pass on our thanks to the staff and students involved in the 2012 National Assessment Program – Science Literacy.

Yours sincerely

Dr Sofia Kesidou
Project Director
Educational Assessment Australia
UNSW Global

2012 National Assessment Program – Science Literacy

Interpreting the Student Reports

Each Year 6 student completed one of the seven different objective (pencil and paper) test booklets and one of two practical tasks. The student reports provide information about each student's achievement on the particular objective test and practical task that s/he completed. Each question tested appeared in three of the seven test booklets in a different position. So although each test booklet was different there were commonalities between the booklets. Each test booklet comprised a different number of questions and only one third of the questions were in common with another booklet. Therefore, the total score achieved by any one student can only be compared to other students who completed the same booklet.

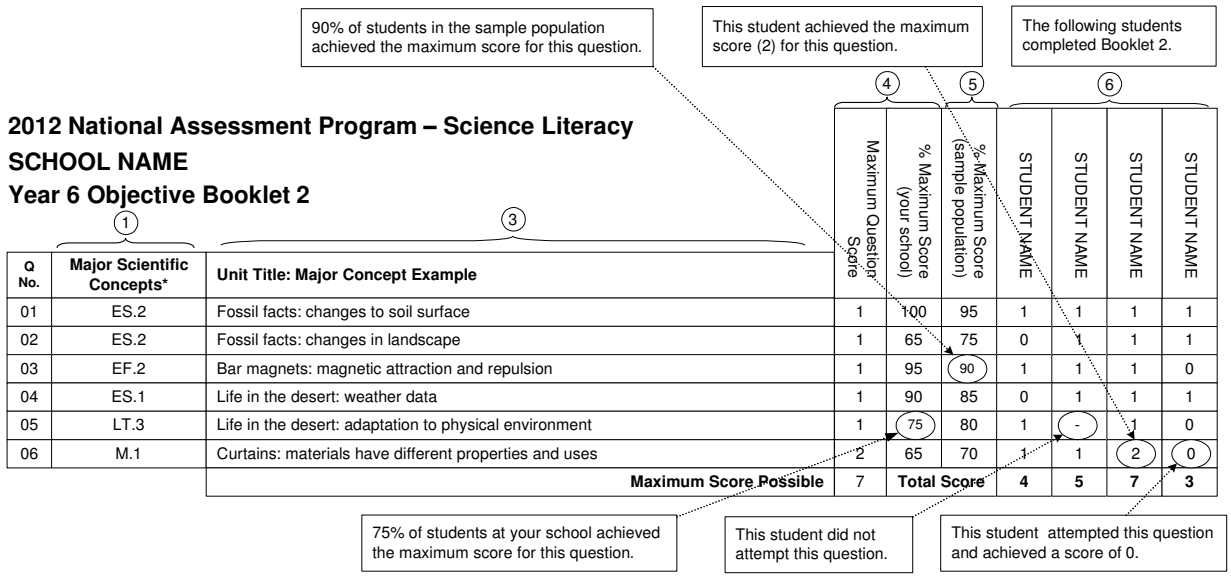
The objective test report and the practical task report include the following information:

1. the relevant major scientific concept addressed by each question (please refer to the key at the end of the A3 practical task report for more information)
2. a description of the skill tested by the question – practical task report only
3. a description of the question context and major concept example – objective booklets report only
4. the maximum possible score for each question and the percentage of students in the school (across multiple booklets) who achieved that score
5. the percentage of students in the sample population who achieved the maximum score on each question (the sample population includes approximately 5% of the Year 6 national population)
6. the name of each student who completed the test for the corresponding test booklet, his/her achievement on each question and overall score on the test.

These reports can be used to:

- compare your students' achievement on each question against the sample population (by comparing the two columns showing the % of students attaining the maximum score)
- compare student achievement within the seven booklets and practical task by looking at the maximum possible score and the total for each student for each test
- identify areas in the curriculum that may need to be covered in more detail by examining the performance of students in each science concept area.

Below is part of a sample report form with some key information explained.



2012 National Assessment Program – Science Literacy

SAMPLE SCHOOL

Year 6 Objective Booklet 4

Q No.	Major Scientific Concepts*	Unit Title: Major Concept Example	Maximum Question Score	% Maximum Score (your school)	% Maximum Score (sample population)	66524 STUDENT	66531 STUDENT
01	M.3	Making jelly: materials can change their state and properties	1	10	34	0	0
02	M.3	Making jelly: materials can change their state and properties	1	20	53	0	0
03	M.3	Making jelly: materials can change their state and properties	1	70	55	1	1
04	LT.1	Pond life: dependence on the environment	1	100	80	1	1
05	LT.1	Pond life: interactions between organisms and interdependence	1	30	38	0	1
06	LT.1	Pond life: interactions between organisms and interdependence	1	30	52	0	0
07	ES.1	Soil salinity: features of soil	1	80	76	1	0
08	ES.1	Soil salinity: features of soil	1	40	61	0	0
09	ES.1	Soil salinity: features of soil	1	80	57	1	0
10	M.1	Testing paper towels: materials have different properties and uses	1	10	32	0	0
11	M.1	Testing paper towels: materials have different properties and uses	1	30	56	0	1
12	M.1	Floating and sinking: materials have different properties and uses	1	10	23	0	0
13	M.1	Floating and sinking: materials have different properties and uses	1	0	14	0	0
14	M.3	Mixing liquids: materials can change their state and properties	1	70	81	1	1
15	M.3	Mixing liquids: materials can change their state and properties	1	30	44	0	1
16	LT.1	Parasites: interactions between organisms	1	20	41	-	1
17	LT.1	Parasites: plant vs animal and major groups	1	70	84	0	-
18	LT.3	Parasites: adaptation to physical environment	1	20	58	0	1
19	LT.3	Parasites: adaptation to physical environment	1	50	52	0	1
20	LT.1	Parasites: interactions between organisms	1	60	41	0	1
21	ES.2	Changing rocks: weathering and erosion	1	40	38	0	1
22	ES.2	Changing rocks: weathering and erosion	1	0	14	0	0
23	ES.2	Changing rocks: weathering and erosion	1	0	9	0	0
24	EF.2	Solar energy: energy and change	1	50	26	1	1
25	EF.2	Solar energy: energy and change	2	0	1	0	0
26	LT.1	Growing Rocket lettuce: survival needs - food, space and shelter	1	18	24	1	0
27	LT.1	Growing Rocket lettuce: survival needs - food, space and shelter	1	9	19	0	0
28	LT.3	Growing Rocket lettuce: change over lifetime, reproductions and life cycles	1	9	16	0	0
29	M.3	Evaporating liquids: materials can change their state and properties	1	9	32	0	1
30	M.3	Evaporating liquids: materials can change their state and properties	1	45	38	1	0
31	M.3	Evaporating liquids: materials can change their state and properties	1	55	62	1	1
32	M.3	Evaporating liquids: materials can change their state and properties	1	0	4	0	0
33	M.3	Evaporating liquids: materials can change their state and properties	1	36	51	0	1
34	M.3	Evaporating liquids: materials can change their state and properties	1	27	48	0	0
35	M.3	Evaporating liquids: materials can change their state and properties	1	27	53	1	1
36	EF.3	Coloured cans: energy sources and receivers	1	9	15	0	0
37	EF.3	Coloured cans: energy sources and receivers	1	18	38	1	0
38	EF.3	Coloured cans: energy sources and receivers	1	9	28	0	0
		Maximum Score Possible	**39**	**Total Score**		**11**	**16**

* Please refer to the key on the last page of this report for an explanation of the major scientific concepts.

2012 National Assessment Program – Science Literacy

SAMPLE SCHOOL

Year 6 Practical Task: Reaction time

Major Scientific Concept: LT.2*

Major Concept Example: Major structures and systems and their functions

Q No.	Descriptor	Maximum Question Score	% Maximum Score (your school)	% Maximum Score (sample population)	66519 STUDENT	66520 STUDENT	66521 STUDENT	66522 STUDENT	66523 STUDENT	66524 STUDENT	66525 STUDENT	66526 STUDENT	66527 STUDENT	66528 STUDENT	66529 STUDENT	66530 STUDENT	66531 STUDENT	66532 STUDENT	66533 STUDENT	66534 STUDENT	66535 STUDENT	66536 STUDENT	66537 STUDENT
01	Compares data to identify the smallest value recorded	1	82	84	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	0	1	1
02	Draws a conclusion that summarises the pattern in the data	1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
03	Identifies an advantage of calculating average values from repeated trials	1	41	59	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0
04	Applies knowledge of the nervous system to explain an experienced phenomenon	1	68	71	1	1	0	0	1	0	1	1	1	0	1	1	1	1	1	1	0	0	1
05	Selects an appropriate change to the method that would improve data collection	1	27	44	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0
06	Identifies a suggestion that would result in an improvement to the method and justifies choice	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
07	Reads a conversion chart to determine reaction time	1	64	84	1	-	0	0	1	0	0	1	1	0	1	0	1	1	1	1	1	0	1
08	Interprets a column graph to identify the number of categories that match a specified criterion	1	50	83	1	-	0	0	1	1	0	0	1	0	1	0	1	1	1	1	0	0	0
09	Provides a justification for disagreeing with a statement that summarises the data in a column graph	1	27	50	0	-	0	0	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0
10	Interprets a column graph to identify the category most likely to contain unreliable data	1	5	4	0	-	0	0	0	0	-	0	0	-	0	0	0	0	0	-	0	-	1
	Maximum Score Possible	**10**	**Total Score**		**5**	**3**	**2**	**1**	**5**	**2**	**1**	**3**	**7**	**2**	**5**	**2**	**6**	**4**	**6**	**3**	**2**	**1**	**4**

A Science Literacy Progress Map has been developed based upon the construct of science literacy and on an analysis of State and Territory curriculum and assessment frameworks. A table of the major scientific concepts (listed below) found most widely in the various State and Territory documents has been developed to accompany the Science Literacy Progress Map. These major scientific concepts are broad statements of scientific understandings that Year 6 students would be expected to demonstrate. For further details please visit: www.nap.edu.au/nap-sample-assessments/napsa-assessment-frameworks.html

* KEY: Major Scientific Concepts

Concept Area:

ES = Earth and Space

Major Scientific Concepts

ES.1 = Earth, sky and people: Our lives depend on air, water and materials from the ground; the ways we live depend on landscape, weather and climate.

ES.2 = The changing Earth: The Earth is composed of materials that are altered by forces within and upon its surface.

ES.3 = Our place in space: The Earth and life on Earth are part of an immense system called the universe.

Concept Area:

EF = Energy and Force

Major Scientific Concepts

EF.1 = Energy and us: Energy is vital to our existence and our quality of life as individuals and as a society.

EF.2 = Transferring energy: Interaction and change involve energy transfers; control of energy transfer enables particular changes to be achieved.

EF.3 = Energy sources and receivers: Observed change in an object or system is indicated by the form and amount of energy transferred to or from it.

Concept Area:

LT = Living Things

Major Scientific Concepts

LT.1 = Living together: Organisms in a particular environment are interdependent.

LT.2 = Structure and function: Living things can be understood in terms of functional units and systems.

LT.3 = Biodiversity, change and continuity: Life on Earth has a history of change and disruption, yet continues generation to generation.

Concept Area:

M = Matter

Major Scientific Concepts

M.1 = Materials and their uses: The properties of materials determine their uses; properties can be modified.

M.2 = Structure and properties: The substructure of materials determines their behaviour and properties.

M.3 = Reactions and change: Patterns of interaction of materials enable us to understand and control those interactions.

Appendix 3
Characteristics of the 2012 Sample

It was desirable to have sampling errors of similar magnitude between jurisdictions. Whilst equal sample sizes were initially assigned to each jurisdiction, the sample sizes were reduced for the ACT, NT and TAS given their relatively smaller populations. The procedures used to draw the 2012 sample of schools were nearly identical to those used in the 2006 and 2009 assessments. Table A3.1 shows the number of sampled schools and students. For example, it can be seen that the percentage of the students sampled from ACT, NT and TAS is smaller compared with other jurisdictions.

Table A3.1 Number of sampled schools and students in each jurisdiction

State/ Territory	Number of sampled schools[1]	Number of sampled students	Percentage of total population of students sampled
ACT	54	1305	8.9
NSW	92	2246	15.3
NT	50	959	6.5
QLD	92	2207	15.0
SA	94	2082	14.2
TAS	64	1420	9.7
VIC	93	2112	14.4
WA	94	2344	16.0
Total	633	14675	100.0

In this and the following tables, percentages have been rounded and may not add up to 100.

1 The number of sampled schools in Table A3.1 differs slightly from those presented in Table A3.4 in some jurisdictions. This difference is due to the rounding of estimates provided (to end up with whole school numbers) and the adjustment of the measure of size for very large schools (so that very large schools are not selected more than once) when drawing the sample. Not all the sampled schools have participated. Of these 633 schools, 16 schools did not participate in the testing (and could not be replaced).

Table A3.2 shows the proportion of students in each sector by jurisdiction for both the selected sample and the population according to the sample frame. The table shows that the difference between the selected sample and the population is generally less than 2 per cent. This indicates that the proportion of students in the selected sample closely matches the population when comparing sector by sector within a jurisdiction.

Table A3.2 Comparison of selected sample and population sector proportions across jurisdictions

State/ Territory	Sector	Population			Selected sample			Difference (population –sample) proportions
		Schools	Students	Sector proportions	Schools	Students	Sector proportions	
ACT	Cath	25	1317	29.2%	15	404	31.0%	-1.7%
	Govt	59	2544	56.5%	32	737	56.5%	0.0%
	Other	13	642	14.3%	7	164	12.6%	1.7%
	Total	97	4503	100.0%	54	1305	100.0%	0.0%
NSW	Cath	442	17808	20.3%	17	463	20.6%	-0.3%
	Govt	1633	59610	68.0%	64	1549	69.0%	-1.0%
	Other	289	10291	11.7%	11	234	10.4%	1.3%
	Total	2364	87709	100.0%	92	2246	100.0%	0.0%
NT	Cath	12	369	11.5%	5	114	11.9%	-0.3%
	Govt	131	2499	78.1%	40	732	76.3%	1.8%
	Other	15	330	10.3%	5	113	11.8%	-1.5%
	Total	158	3198	100.0%	50	959	100.0%	0.0%
QLD	Cath	224	10181	17.2%	15	448	20.3%	-3.1%
	Govt	1001	41371	69.8%	66	1503	68.1%	1.7%
	Other	167	7679	13.0%	11	256	11.6%	1.4%
	Total	1392	59231	100.0%	92	2207	100.0%	0.0%
SA	Cath	83	3632	18.8%	17	411	19.7%	-0.9%
	Govt	434	12345	63.9%	61	1320	63.4%	0.5%
	Other	88	3349	17.3%	16	351	16.9%	0.5%
	Total	605	19326	100.0%	94	2082	100.0%	0.0%
TAS	Cath	32	1110	16.7%	10	240	16.9%	-0.2%
	Govt	153	4838	72.8%	46	1018	71.7%	1.1%
	Other	36	695	10.5%	8	162	11.4%	-0.9%
	Total	221	6643	100.0%	64	1420	100.0%	0.0%
VIC	Cath	396	14262	22.7%	21	490	23.2%	-0.5%
	Govt	1151	40717	64.7%	60	1354	64.1%	0.6%
	Other	219	7937	12.6%	12	268	12.7%	-0.1%
	Total	1766	62916	100.0%	93	2112	100.0%	0.0%
WA	Cath	132	4989	16.9%	17	451	19.2%	-2.3%
	Govt	624	20163	68.4%	64	1582	67.5%	0.9%
	Other	127	4334	14.7%	13	311	13.3%	1.4%
	Total	883	29486	100.0%	94	2344	100.0%	0.0%
AUST	Cath	1346	53668	19.7%	117	3021	20.6%	-0.9%
	Govt	5186	184087	67.4%	433	9795	66.7%	0.7%
	Other	954	35257	12.9%	83	1859	12.7%	0.2%
	Total	7486	273012	100.0%	633	14675	100.0%	0.0%

Schools were also classified according to their enrolment size. Small schools (i.e. *moderately small* and *very small* schools) were under-sampled and large schools were slightly over-sampled. This approach was adopted to ensure that an adequate number of students would be assessed, while still ensuring very small schools would be represented without vastly increasing the overall number of schools sampled. *Very small* schools were under-sampled to a larger degree than *moderately small* schools. Table A3.3 shows the number of schools according to school size for the population and the selected sample. Table A3.3 also shows the percentage of students in the population compared to the selected sample according to school size. When considered in terms of the number of students, the under-sampling of small schools is not as noticeable. For example, 4.3 per cent of the population attend a *very small* school which is very similar to the 3.1 per cent of students from *very small* schools included in the selected sample.

Table A3.3 Comparison of population and selected sample proportions according to school size

School size	Population			Selected sample		
	Schools	Students	Proportion of students by school size	Schools	Students	Proportion of students by school size
Large	4298	238819	87.5	502	12914	88.0
Moderately small	1209	22250	8.2	72	1313	8.9
Very small	1979	11872	4.3	59	448	3.1
Total	7486	272941	100.0	633	14675	100.0

Appendix 4
Technical Notes on Sampling

Stratification details

For each jurisdiction, schools were separated into three separate strata according to their size: very small; moderately small; and large. The target proportion of students and number of schools selected within each of the strata were determined using the PISA treatment of small schools (OECD 2012, pp. 68–69). Essentially, the aim was to balance selecting an adequate sample without substantially increasing the number of sampled schools.

Large schools within each jurisdiction were further separated according to their school sector. The target numbers of large schools were proportionally allocated amongst the school sectors for each jurisdiction. Very small and moderately small strata were sorted according to school sector, then by the remaining implicit stratification variables – NAPLAN quintile, Geographic location and Measure of Size (MOS). This strategy meant that the sampling frame was divided into 40 explicit strata overall. That is, there were 24 strata containing large schools (8 jurisdictions × 3 sectors); eight moderately small school strata (1 per jurisdiction); and eight very small school strata (1 per jurisdiction).

The stratification for small schools was slightly more complex than for large schools. Small schools were ordered by Sector, NAPLAN quintile, Geographic location and then MOS. The sort order was alternated so that 'like schools' were always nearby.

Each stratum was sorted first by sector. Within each sector, schools were further sorted by NAPLAN quintile. This sort order was alternated between ascending to descending between sectors (i.e. Sector 1 had NAPLAN quintile sorted ascending, Sector 2 had NAPLAN quintile sorted descending, Sector 3 had NAPLAN quintile sorted ascending). Similarly, within each NAPLAN quintile category, schools were further sorted by Geographic location. This sort order was alternated between ascending to descending between sectors (i.e. NAPLAN quintile 1 had Geographic location sorted ascending, NAPLAN quintile 2 had Geographic location sorted descending, NAPLAN quintile 3 had Geographic location sorted ascending etc.) The sort order for MOS was then alternated from low to high, then high to low, each time a new Sector/ NAPLAN quintile / Geographic location classification was encountered. Table A4.1 illustrates the sort-order procedures that were employed for small Catholic schools.

Table A4.1 The sort ordering procedures employed for small Catholic schools

Sector	NAPLAN quintile	Geographic location	ENR sort order
1	1	1	A
1	1	2	D
1	1	3	A
1	2	3	D
1	2	2	A
1	2	1	D
1	3	1	A
1	3	2	D
1	3	3	A
1	4	3	D
1	4	2	A
1	4	1	D
1	5	1	A
1	5	2	D
1	5	3	A

After small schools were stratified, the MOS for each school in the stratum was set equal to the average ENR of all schools within that particular stratum. This was equivalent to selecting a simple random sample of small schools. Such a strategy meant that very small schools would not be assigned excessively large sampling weights.

Random start and sampling interval values

The sampling interval ([stratum enrolment size]/[planned number of schools]) is rounded to the nearest integer. Table A4.2 shows the starting values used to draw the sample for each explicit stratum.

Table A4.2 Stratum variables for sample selection

Stratum	MOS	Number of schools	Interval	Random start
ACT_Large_C	1198	15	80	12
ACT_Large_G	2400	30	80	65
ACT_Large_I	592	7	85	38
ACT_ModSmall	268	4	67	61
ACT_VerySmall	45	1	45	12
NSW_Large_C	15616	15	1041	194
NSW_Large_G	52816	52	1016	482
NSW_Large_I	8930	9	992	758
NSW_ModSmall	7001	9	778	144
NSW_VerySmall	3346	7	478	445
NT_Large_C	326	4	82	38
NT_Large_G	1826	23	79	2
NT_Large_I	203	3	68	2
NT_ModSmall	454	8	57	10
NT_VerySmall	389	12	32	17
QLD_Large_C	9337	14	667	636
QLD_Large_G	37320	54	691	342
QLD_Large_I	7034	10	703	97
QLD_ModSmall	3161	6	527	132
QLD_VerySmall	2379	8	297	218
SA_Large_C	3201	14	229	168
SA_Large_G	9779	44	222	107
SA_Large_I	2943	13	226	183
SA_ModSmall	2464	15	164	147
SA_VerySmall	939	8	117	34
TAS_Large_C	926	8	116	13
TAS_Large_G	4029	35	115	76
TAS_Large_I	498	4	125	102
TAS_ModSmall	747	9	83	13
TAS_VerySmall	443	7	63	35
VIC_Large_C	12145	17	714	331
VIC_Large_G	35059	48	730	209
VIC_Large_I	7025	10	703	423
VIC_ModSmall	6131	11	557	217
VIC_VerySmall	2556	7	365	145
WA_Large_C	4517	13	347	4
WA_Large_G	17574	51	345	268
WA_Large_I	3756	11	341	180
WA_ModSmall	2163	9	240	49
WA_VerySmall	1476	9	164	32

Appendix 5
Programming Notes on Sampling

E.1 SPSS syntax for sample selection

```
*==========================================
*==========================================
        NAP-SL 2012 SAMPLE PROCEDURE
*==========================================
*==========================================.
*SPSS version 20.

*==========================================
        PPS SAMPLE MACRO
*==========================================.
*-------------------------------------------------------------------------------
 This macro will select sample schools for a particular stratum

 The following arguments are required:
 ~~~enrsize is equal to average enrolment size for modsmall and verysmall strata
     otherwise, set enrsize equal to 999 for large school strata
 ~~~strata is the name of the current stratum
 ~~~randm is a random number
 ~~~const is the sampling interval

*-------------------------------------------------------------------------------.

DEFINE !SAMPLE (enrsize = !DEFAULT(999) !TOKENS(1)
                / strata = !TOKENS(1)
                / randm = !TOKENS(1)
                / const = !TOKENS(1)).

DATASET CLOSE ALL.
GET FILE='SampleFrame.sav'.
*-----------------------------------------------------------------.
*   EXPLICIT STRATIFICATION.
*-----------------------------------------------------------------.
select if (RTRIM(Stratum)=!strata).
exe.
*-----------------------------------------------------------------.
*   IMPLICIT STRATIFICATION.
*-----------------------------------------------------------------.
*all implicit stratification variables need to be numeric ordinal categories.
```

```
*sequential numbering is not required.

*~~~assign SectorId as implicit stratification variable.
COMPUTE imp_0 = SectorId.
SORT CASES BY imp_0 (A).
RANK VARIABLES = imp_0 (A) /RANK /PRINT=YES /TIES=CONDENSE.

*~~~add NAPLANDATA as implicit stratification variable.
DO IF (MOD(Rimp_0,2) > 0).
  compute imp_1 = (Rimp_0*100) + (NAPLANDATA*-1).
ELSE.
  compute imp_1 = (Rimp_0*100) + NAPLANDATA.
END IF.
SORT CASES BY imp_1(A).
RANK VARIABLES = imp_1(A) /RANK /PRINT=YES /TIES=CONDENSE.

*~~~add GeoId as implicit stratification variable.
DO IF (MOD(Rimp_1,2) > 0).
  compute imp_2 = (Rimp_1*100) + (GeoId*-1).
ELSE.
  compute imp_2 = (Rimp_1*100) + GeoId.
END IF.
SORT CASES BY imp_2(A).
RANK VARIABLES = imp_2(A) /RANK /PRINT=YES /TIES=CONDENSE.

*~~~add gr06 as implicit stratification variable.
DO IF (MOD(Rimp_2,2) > 0).
  compute imp_3 = (Rimp_2*1000) + (gr06*-1).
ELSE.
  compute imp_3 = (Rimp_2*1000) + gr06.
END IF.
SORT CASES BY imp_3 (A).
RANK VARIABLES = imp_3 (A) /RANK /PRINT=YES /TIES=CONDENSE.

SORT CASES BY imp_0 (A) imp_1 (A) imp_2 (A) imp_3 (A).
*file is now implicitly stratified.
*----------------------------------------------------------------.
*   MEASURE OF SIZE (MOS) ADJUSTMENTS.
*----------------------------------------------------------------.
*=======SMALL SCHOOLS=======.
!IF (!enrsize = 999)!THEN.
  *do nothing.
!ELSE.
 * for small schools set MOS equal to avg enr size for the explicit stratum.
        compute tmpgr06 = gr06.
        compute gr06 = !enrsize.
!IFEND.

*=======SET VERY LARGE SCHOOLS EQUAL TO THE SAMPLING
INTERVAL=======.
```

```
if (gr06>!const) gr06 = !const.
exe.

*-----------------------------------------------------------------.
*   SELECT SCHOOLS WITH PROBABILITY PROPORTIONAL TO SIZE (PPS)
*-----------------------------------------------------------------.
compute ranstart = !randm.
compute interval = !const.
compute case = $casenum.
exe.
if ($casenum = 1) ticket1 = 1.
if ($casenum = 1) ticket2 = gr06.
if ($casenum > 1) ticket1 = lag(ticket2) + 1.
if ($casenum > 1) ticket2 = lag(ticket2) + gr06.
if ($casenum = 1) selector = ranstart.
if ($casenum > 1) selector = lag(selector).
string select (a3).
compute select = '___'.
if (ticket1 <= selector and selector <= ticket2) select = 'YES'.
if (select = 'YES') selector = selector + interval.
*HANDLE FOR LARGE SCHOOLS.
if (select = 'YES' and selector < ticket2) select = 'SOS'.
exe.

if ($casenum = 1) wintickt=ranstart.
if ($casenum > 1) wintickt=lag(selector).
exe.

*=======SELECT REPLACEMENT SCHOOLS=======.
DO IF ((lag(select)='YES' or lag(select)='SOS') and select = '___').
        compute select = 'R_1'.
        compute replaceid = lag(schoolid).
END IF.
DO IF ((lag(select,2)='YES' or lag(select,2)='SOS') and select = '___' and
lag($casenum,2)=1).
        compute select = 'R_2'.
        compute replaceid = lag(schoolid,2).
END IF.
SORT CASES BY case (D) .
DO IF ((lag(select)='YES' or lag(select)='SOS') and select = '___').
        compute select = 'R_2'.
        compute replaceid = lag(schoolid).
END IF.
DO IF ((lag(select,2)='YES' or lag(select,2)='SOS') and select = '___' and
lag($casenum,2)=1).
        compute select = 'R_1'.
        compute replaceid = lag(schoolid,2).
END IF.
SORT CASES BY case (A) .
if (select = 'YES' or select = 'SOS') replaceid = schoolid.
```

```
exe.
SAVE OUTFILE=!QUOTE(!CONCAT('All_',!UNQUOTE(!strata) , '.sav')).

*=======KEEP SAMPLED AND REPLACEMENT SCHOOLS=======.
set width = 120.
set length = 1000.
title Schools Selected from the Specified Stratum !strata.
select if (select='YES' or select='SOS').
list var=OfficialSchoolName stratum gr06 ticket1 ticket2 wintickt select / format =
numbered.
title.
SAVE OUTFILE=!QUOTE(!CONCAT('Sample_',!UNQUOTE(!strata) , '.sav')).

!ENDDEFINE.

*=========================================
        DRAW SAMPLE
*=========================================.

*set working file directory.
CD 'C:\EMS\NAPSL12\SampleSchools'.

!SAMPLE strata='ACT_Large_C' const=80 randm=12 enrsize=999.
!SAMPLE strata='ACT_Large_G' const=80 randm=65 enrsize=999.
!SAMPLE strata='ACT_Large_I' const=85 randm=38 enrsize=999.
!SAMPLE strata='ACT_ModSmall' const=67 randm=61 enrsize=19.
!SAMPLE strata='ACT_VerySmall' const=45 randm=12 enrsize=8.
!SAMPLE strata='NSW_Large_C' const=1041 randm=194 enrsize=999.
!SAMPLE strata='NSW_Large_G' const=1016 randm=482 enrsize=999.
!SAMPLE strata='NSW_Large_I' const=992 randm=758 enrsize=999.
!SAMPLE strata='NSW_ModSmall' const=778 randm=144 enrsize=18.
!SAMPLE strata='NSW_VerySmall' const=478 randm=445 enrsize=6.
!SAMPLE strata='NT_Large_C' const=82 randm=38 enrsize=999.
!SAMPLE strata='NT_Large_G' const=79 randm=2 enrsize=999.
!SAMPLE strata='NT_Large_I' const=68 randm=2 enrsize=999.
!SAMPLE strata='NT_ModSmall' const=57 randm=10 enrsize=17.
!SAMPLE strata='NT_VerySmall' const=32 randm=17 enrsize=5.
!SAMPLE strata='QLD_Large_C' const=667 randm=636 enrsize=999.
!SAMPLE strata='QLD_Large_G' const=691 randm=342 enrsize=999.
!SAMPLE strata='QLD_Large_I' const=703 randm=97 enrsize=999.
!SAMPLE strata='QLD_ModSmall' const=527 randm=132 enrsize=18.
!SAMPLE strata='QLD_VerySmall' const=297 randm=218 enrsize=6.
!SAMPLE strata='SA_Large_C' const=229 randm=168 enrsize=999.
!SAMPLE strata='SA_Large_G' const=222 randm=107 enrsize=999.
!SAMPLE strata='SA_Large_I' const=226 randm=183 enrsize=999.
!SAMPLE strata='SA_ModSmall' const=164 randm=147 enrsize=19.
!SAMPLE strata='SA_VerySmall' const=117 randm=34 enrsize=6.
!SAMPLE strata='TAS_Large_C' const=116 randm=13 enrsize=999.
!SAMPLE strata='TAS_Large_G' const=115 randm=76 enrsize=999.
!SAMPLE strata='TAS_Large_I' const=125 randm=102 enrsize=999.
```

```
!SAMPLE strata='TAS_ModSmall' const=83 randm=13 enrsize=19.
!SAMPLE strata='TAS_VerySmall' const=63 randm=35 enrsize=7.
!SAMPLE strata='VIC_Large_C' const=714 randm=331 enrsize=999.
!SAMPLE strata='VIC_Large_G' const=730 randm=209 enrsize=999.
!SAMPLE strata='VIC_Large_I' const=703 randm=423 enrsize=999.
!SAMPLE strata='VIC_ModSmall' const=557 randm=217 enrsize=19.
!SAMPLE strata='VIC_VerySmall' const=365 randm=145 enrsize=6.
!SAMPLE strata='WA_Large_C' const=347 randm=4 enrsize=999.
!SAMPLE strata='WA_Large_G' const=345 randm=268 enrsize=999.
!SAMPLE strata='WA_Large_I' const=341 randm=180 enrsize=999.
!SAMPLE strata='WA_ModSmall' const=240 randm=49 enrsize=18.
!SAMPLE strata='WA_VerySmall' const=164 randm=32 enrsize=6.

*=========================================
        ALL SCHOOLS IN SINGLE FILE WITH RESULTS
*=========================================.
ADD FILES
 /FILE='All_ACT_Large_C.sav'
 /FILE='All_ACT_Large_G.sav'
 /FILE='All_ACT_Large_I.sav'
 /FILE='All_ACT_ModSmall.sav'
 /FILE='All_ACT_VerySmall.sav'
 /FILE='All_NSW_Large_C.sav'
 /FILE='All_NSW_Large_G.sav'
 /FILE='All_NSW_Large_I.sav'
 /FILE='All_NSW_ModSmall.sav'
 /FILE='All_NSW_VerySmall.sav'
 /FILE='All_NT_Large_C.sav'
 /FILE='All_NT_Large_G.sav'
 /FILE='All_NT_Large_I.sav'
 /FILE='All_NT_ModSmall.sav'
 /FILE='All_NT_VerySmall.sav'
 /FILE='All_QLD_Large_C.sav'
 /FILE='All_QLD_Large_G.sav'
 /FILE='All_QLD_Large_I.sav'
 /FILE='All_QLD_ModSmall.sav'
 /FILE='All_QLD_VerySmall.sav'
 /FILE='All_SA_Large_C.sav'
 /FILE='All_SA_Large_G.sav'
 /FILE='All_SA_Large_I.sav'
 /FILE='All_SA_ModSmall.sav'
 /FILE='All_SA_VerySmall.sav'
 /FILE='All_TAS_Large_C.sav'
 /FILE='All_TAS_Large_G.sav'
 /FILE='All_TAS_Large_I.sav'
 /FILE='All_TAS_ModSmall.sav'
 /FILE='All_TAS_VerySmall.sav'
 /FILE='All_VIC_Large_C.sav'
 /FILE='All_VIC_Large_G.sav'
 /FILE='All_VIC_Large_I.sav'
```

```
  /FILE='All_VIC_ModSmall.sav'
  /FILE='All_VIC_VerySmall.sav'
  /FILE='All_WA_Large_C.sav'
  /FILE='All_WA_Large_G.sav'
  /FILE='All_WA_Large_I.sav'
  /FILE='All_WA_ModSmall.sav'
  /FILE='All_WA_VerySmall.sav'.
EXECUTE.
SAVE OUTFILE='AllSchools.sav'.

*==========================================
          LIST OF SAMPLE SCHOOLS
*==========================================.
ADD FILES
  /FILE='Sample_ACT_Large_C.sav'
  /FILE='Sample_ACT_Large_G.sav'
  /FILE='Sample_ACT_Large_I.sav'
  /FILE='Sample_ACT_ModSmall.sav'
  /FILE='Sample_ACT_VerySmall.sav'
  /FILE='Sample_NSW_Large_C.sav'
  /FILE='Sample_NSW_Large_G.sav'
  /FILE='Sample_NSW_Large_I.sav'
  /FILE='Sample_NSW_ModSmall.sav'
  /FILE='Sample_NSW_VerySmall.sav'
  /FILE='Sample_NT_Large_C.sav'
  /FILE='Sample_NT_Large_G.sav'
  /FILE='Sample_NT_Large_I.sav'
  /FILE='Sample_NT_ModSmall.sav'
  /FILE='Sample_NT_VerySmall.sav'
  /FILE='Sample_QLD_Large_C.sav'
  /FILE='Sample_QLD_Large_G.sav'
  /FILE='Sample_QLD_Large_I.sav'
  /FILE='Sample_QLD_ModSmall.sav'
  /FILE='Sample_QLD_VerySmall.sav'
  /FILE='Sample_SA_Large_C.sav'
  /FILE='Sample_SA_Large_G.sav'
  /FILE='Sample_SA_Large_I.sav'
  /FILE='Sample_SA_ModSmall.sav'
  /FILE='Sample_SA_VerySmall.sav'
  /FILE='Sample_TAS_Large_C.sav'
  /FILE='Sample_TAS_Large_G.sav'
  /FILE='Sample_TAS_Large_I.sav'
  /FILE='Sample_TAS_ModSmall.sav'
  /FILE='Sample_TAS_VerySmall.sav'
  /FILE='Sample_VIC_Large_C.sav'
  /FILE='Sample_VIC_Large_G.sav'
  /FILE='Sample_VIC_Large_I.sav'
  /FILE='Sample_VIC_ModSmall.sav'
  /FILE='Sample_VIC_VerySmall.sav'
  /FILE='Sample_WA_Large_C.sav'
```

```
  /FILE='Sample_WA_Large_G.sav'
  /FILE='Sample_WA_Large_I.sav'
  /FILE='Sample_WA_ModSmall.sav'
  /FILE='Sample_WA_VerySmall.sav'.
EXECUTE.
SAVE OUTFILE='SampleSchools2012.sav'.
```

Appendix 6
Student Participation Form

NAP-SL STUDENT PARTICIPATION FORM (SPF)

The Student Participation Form (SPF) lists students registered to take part in the National Assessment Program – Science Literacy. Please complete Part A – Sampling Information (below) and Part B – Student Participation (overleaf). Please refer to pages 8-10 of the Test Administrator's Manual for further details of how to complete this form.

School Name:	Sample School
State/Territory:	Sample State
School ID:	4
Class(es) involved:	Year 6
Class practical task:	Reaction time

PART A – SAMPLING INFORMATION

(A) No. of Students in Year 6	(B) No. of Classes in Year 6	(C) Estimated Sample Size	(D) Enrolled Sample Size
24	1	24	_____

Please sign below to acknowledge that you have checked the Test Booklets and Student Participation Form and that all is complete and in order.
Don't forget to take a photocopy of both sides of this form and keep a copy for your records. Return the original with the test booklets.

School Contact Officer: Name: _____ Signature: _____

Test Administrator: Name: _____ Signature: _____

This table relates to PART B – STUDENT PARTICIPATION: See full explanation on pages 8-10 of the Test Administrator's Manual		
INDIGENOUS CODES (Column 5)	SPECIAL EDUCATION NEEDS (SEN) CODES (Column 7)	NON-INCLUSION CODES (Columns 9 and 11)
1 = Aboriginal but not Torres Strait Islander Origin	0 = No special education needs	10 = Absent
2 = Torres Strait Islander but not Aboriginal Origin	1 = Functional disability	11 = Not included; functional disability
3 = Both Aboriginal and Torres Strait Islander Origin	2 = Intellectual disability	12 = Not included; intellectual disability
4 = Neither Aboriginal nor Torres Strait Islander Origin	3 = Limited test language proficiency	13 = Not included; limited test language proficiency
9 = Not stated/Unknown		14 = Student or parent refusal

PART B – STUDENT PARTICIPATION (Completed by the School Contact Officer and Test Administrator)

(1) Student ID	(2) Student Name	(3) Booklet No. (1-7)	(4) Sex F or M	(5) Indigenous Code (see overleaf)	(6) Birth Date (DD-MM-YY)	(7) SEN Code (see overleaf)	(8) Objective Test Not attempted = 0 Attempted = 1	(9) Non-Inclusion Code (see overleaf)	(10) Practical Task Not attempted = 0 Attempted = 1	(11) Non-Inclusion Code (see overleaf)
858415	Student 858415	4	F	4	20/05/01					
858423	Student 858423	5	M	4	09/08/00					
858431	Student 858431	6	M	9	16/02/01					
858357	Student 858357	7	M	4	25/04/01					
858365	Student 858365	1	M	4	14/11/00					
858373	Student 858373	2	M	4	07/11/00					
858266	Student 858266	3	F	9	23/05/01					
858449	Student 858449	4	M	4	12/03/01					
858456	Student 858456	5	M	4	22/01/01					
858381	Student 858381	6	M	4	19/03/01					
858274	Student 858274	7	M	4	07/07/00					
858282	Student 858282	1	M	4	09/01/01					
858290	Student 858290	2	M	4	02/06/00					
858464	Student 858464	3	M	4	07/11/00					
858308	Student 858308	4	F	4	25/05/01					
858316	Student 858316	5	F	4	25/01/01					
858498	Student 858498	6	F	3	28/02/01					
858399	Student 858399	7	M	9	03/04/01					
858324	Student 858324	1	M	4	27/06/01					
858332	Student 858332	2	M	4	28/10/00					
858407	Student 858407	3	M	4	30/10/00					
858472	Student 858472	4	M	4	21/08/00					
858480	Student 858480	5	M	4	22/01/00					
858340	Student 858340	6	F	1	14/12/00					
858506		7								
858514		1								
858522		2								

Appendix 7
Variables in File

Table A7.1 File Name: NAPSL2012_PV_2013-03-11.sav

Variable names	Description
BarcodeID	Student Barcode
Booklet	Booklet Number (Objective items)
PAQ01 to PBQ10	Practical Tasks Items
ID0B532 to IDOB406	Objective Items
Geolocation	Geolocation Code
State1	State regression variable 1
State2	State regression variable 2
State3	State regression variable 3
State4	State regression variable 4
State5	State regression variable 5
State6	State regression variable 6
State7	State regression variable 7
Gender1	Gender regression variable 1
Gender2	Gender regression variable 2
ATSI	ATSI recode for stratification
ATSI1	ATSI regression variable 1
ATSI2	ATSI regression variable 2
Sector1	Sector regression variable 1
Sector2	Sector regression variable 2
Geolocation1	Geolocation regression variable 1
Geolocation2	Geolocation regression variable 2
LBOTE1	LBOTE regression variable 1
LBOTE2	LBOTE regression variable 2
Prac	Practical Task identifier
ObjectiveTest	Did not sit code for Objective Test
PracticalTask	Did not sit code for Practical Task
SchoolID	School ID
State	State
Stratum	Stratum Identifier
Participant	Participant flag for weight
NonParticipant	Non-Participant flag
NonInclusionCode	Non inclusion code
FinalStudentWeight	Final Student Weight
FinalClassWeight	Final Class Weight
FinalSchoolWeight	Final School Weight
FinalWeight	Final Weight
SampleZone	Sampling Zone
PairNum	Sampling Zone Pair identifier
DblWgtPairNum	Sampling Zone Pair weight flag
RW0 - RW318	Replicate weight 0 to 318
RUMMWLE	Weighted Likelihood estimate from RUMM2020

Table A7.1 (Cont.) File Name: NAPSL2012_PV_2013-03-11.sav

Variable names	Description
SchWleRUMM	School Mean WLE from RUMM2020
Free_PV1 - PV10	2012 Plausible Value calibrated free (1 to 10)
EAP	EAP value
EAP_SE	EAP SE Value
PV1 - PV10	2012 PV1 -10 (on 2006 scale)
Level1 - Level 10	2012 Level for PV1 - 10
YearLevel	School Year Level
Gender	Student Gender
DOB	Student Date of Birth
CountryBirth	Country of Birth Code
AtsiID	ATSI Code
SECodeP1ID	Parent 1 School Education code
SECodeP2ID	Parent 2 School Education code
NSECodeP1ID	Parent 1 non-School Education code
NSECodeP2ID	Parent 2 non-School Education code
OccupationP1	Parent 1 Occupation Code
OccupationP2	Parent 2 Occupation Code
LboteSID	Language Background of Student
LbotеP1ID	Language Background of Parent 1
LbotеP2ID	Language Background of Parent 2
SENCode	Special Education Needs code
ObjNonInclusionCode	Objective Test non-inclusion code
PracNonInclusionCode	Practical Task non-inclusion code
Sector	School Sector

Appendix 8
ConQuest Control File for Producing Plausible Values

Table A8.1 File Name: NAPSL2012_Produce_2012_PV.cqc

```
reset;

data NAPSL2012_PV_Check_2013_03_14.dat;

format pid 1-6, responses 10-121, booklet 8

   FinalWeight 123-130

   schoolmeanWLE 132-139

   stateNSW 141

   stateNT  142

   stateQLD 143

   stateSA  144

   stateTAS 145

   stateVIC 146

   stateWA  147

   sectorG  149

   sectorI  150

   geoProvincial 152

   geoRemote    153

   gender2  155

   gender9  156

   atsi1    158

   atsi9    159

   lbote1   161

   lbote9   162;

label << NAPSL2012.lab;
```

```
set constraint=none;

set n_plausible=10;

caseweight FinalWeight;



key 1111111111111231134111311111111211321143411121111141121213111131114321121111121111214214214134111311411 ! 1;

key 2xxxxx2xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx2xxxxxxxxxxxxxx2xxxxxxxxxxxxxxxxxxxxxx2xxxxxxxxxxxxxxxxx2xxxxxxxxxxxxx ! 2;

codes 0,1,2,3,4,5,7,9,A,B;

import anchor_param << CalibrationSample.anc;

model booklet + item + item*step;

regression

   schoolmeanWLE

   stateNSW

   stateNT

   stateQLD

   stateSA

   stateTAS

   stateVIC

   stateWA

   sectorG

   sectorI

   geoProvincial

   geoRemote

   gender2

   gender9

   atsi1

   atsi9

   lbote1

   lbote9   ;
```

```
set warnings = no;

estimate ! iterations=2000, fit=no, nodes=200;

show !estimate=latent >> NAPSL2012_PV_Check_2013_03_14.shw;

itanal >> NAPSL2012_PV_Check_2013_03_14.itn;

show cases !estimate=latent >> NAPSL2012_PV_Check_2013_03_14.pls;

show cases !estimate=wle >> NAPSL2012_PV_Check_2013_03_14.WLE;
```