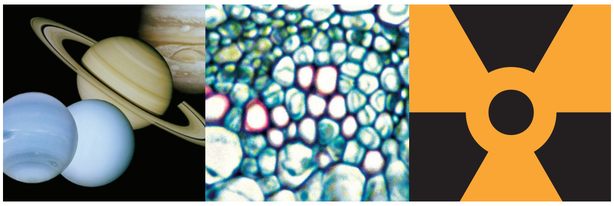
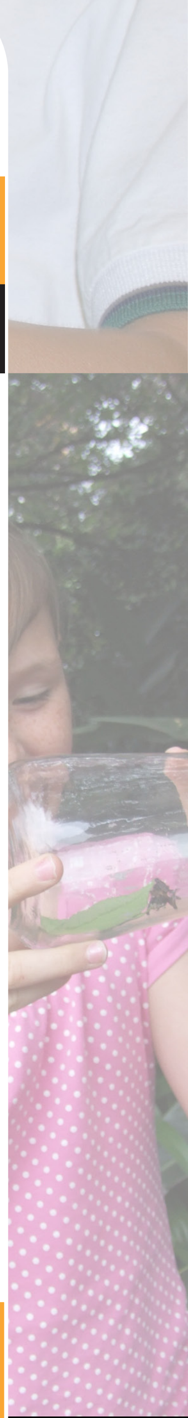


National Assessment Program – Science Literacy Year 6 Technical Report

2009



AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

NAP-SL 2009 Project Staff

Penny Hutton from Educational Assessment Australia (EAA) was the Project Director of NAP-SL 2009. Jennifer Cowing (EAA) was the Project Manager. The test development team was led by Sofia Kesidou (EAA). The Public Report was written by Penny Hutton, Sofia Kesidou, Goran Lazendic and Nathaniel Lewis. The School Release Materials were written by Penny Hutton, Jennifer Cowing and Philip Thompson (EAA).

The sampling and data analysis tasks were undertaken by Goran Lazendic, Nathaniel Lewis and Jennifer Lau from EAA and Margaret Wu and Mark Dulhunty from Educational Measurement Solutions (EMS). The Technical Report was written by Margaret Wu (EMS), Penny Hutton, Goran Lazendic and Nathaniel Lewis (EAA).

Cover Photograph
Top: courtesy of International Grammar School,
© Educational Assessment Australia

© Australian Curriculum, Assessment and Reporting Authority 2010

This work is copyright. You may download, display, print and reproduce this material in unaltered form only (retaining this notice) for your personal, non-commercial use or use within your organisation.

All other rights are reserved. Requests and inquiries concerning reproduction and rights should be addressed to:
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
2009 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 Ministerial Council for Education, Early Childhood Development and Youth Affairs, MCEECDYA) 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 in October 2006 and the third cycle was conducted in October 2009.

The National Assessment Program – Science Literacy (NAP-SL) was the first assessment program designed specifically to provide information about performance against MCEECDYA's National Goals for Schooling in the Twenty-First Century (now the Educational Goals for Young Australians). MCEECDYA has since also endorsed similar assessment programs to be conducted for Civics and Citizenship (CC) and Information and Communications Technology Literacy (ICTL). The intention is that each assessment program will be 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 against which future performance could be compared.

PMRT awarded the contract for the third cycle of science literacy testing to Educational Assessment Australia (EAA). The Benchmarking and Educational Measurement Unit (BEMU) was nominated by PMRT to liaise between the contractors and PMRT in the delivery of the project. In 2009 PMRT and BEMU were incorporated into the Australian Curriculum, Assessment and Reporting Authority (ACARA).

The Science Literacy Review Committee (SLRC), comprising members from all states, territories, sectors and specific interest groups, was a consultative group to the project.

1.2 Purposes of the Technical Report

This technical report aims to provide detailed information with regard to the conduct of the 2009 National Assessment Program – Science Literacy so that valid interpretations of the 2009 results can be made, and future cycles can be implemented with appropriate linking information from past cycles. Further, a fully documented set of the National Assessment Program – Science Literacy procedures can also provide information for researchers who are planning assessments of this kind. The methodologies used in the 2009 National Assessment Program – Science Literacy 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 domains 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 2009 results could be reported against the baseline established in 2006 are discussed in Chapter 8.

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

Appendices A – H 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 domains

The National Assessment Program – Science Literacy 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 OECD – Programme for International Student Assessment (OECD-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)

A scientific literacy assessment domain was developed for the assessment in consultation with curriculum experts from each state and territory and representatives of the Catholic and independent school sectors. This domain includes the definition of scientific literacy and outlines the development of scientific literacy across three main areas:

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

Strand B: interpreting evidence and drawing conclusions from their own or others' data, critiquing the trustworthiness of evidence and claims made by others, and communicating findings.

Strand C: using scientific understandings for describing and explaining natural phenomena, and for interpreting reports about phenomena.

A conscious effort was made to develop assessment items that related to everyday contexts.

The scientific literacy domain is detailed in Appendix A of this report. The items drew on four specific concept areas: Earth and Space (ES); Energy and Force (EF); Living Things (LT) and Matter (M). These major concept areas are found most widely in curriculum documents across all states and territories and were used by item writers to guide test development. The list of endorsed major scientific concepts and examples for each of these areas is described in Table A.2.

The intention was to ensure that all Year 6 students were familiar with the materials and experiences to be used in the National Assessment Program – Science Literacy and so avoid any systematic bias in the instruments being developed.

2.2 Test blueprint

In 2007, MCEECDYA published a Response for Tender (RFT) document. Consequently, EAA developed the following proposal for the tests:

> *It is anticipated that the 2009 final test forms will contain approximately 110 items in total (including link items from 2003 and 2006) 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 2012.*
>
> *The total number of new items to be developed is reduced by:*
>
> - *use of secure 2006 items for 2009*
> - *use of 2003 secure items used in 2006 for 2009.*
>
> *However, we recommend that 2.5 times the number of new items required for the final item pool be developed, in the expectation that some will be eliminated post-trial, and to ensure that there is a strong pool available for possible use in subsequent test cycles.*
>
> *... It is proposed that there be three types of items developed: multiple-choice items, short constructed response items (requiring one or two word responses from students); and 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.*
>
> *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 paper and pencil 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 determined 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, a rotated test booklet design was implemented. A rotational design allows a greater number of items to be assessed by using numerous 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 in the test form. Seven test forms were agreed to for the final test; eight for the trial. Table 2.1 demonstrates the rotational design used for the 2009 National Assessment Program – Science Literacy.

Table 2.1 Rotational design used in the 2009 National Assessment Program – Science Literacy

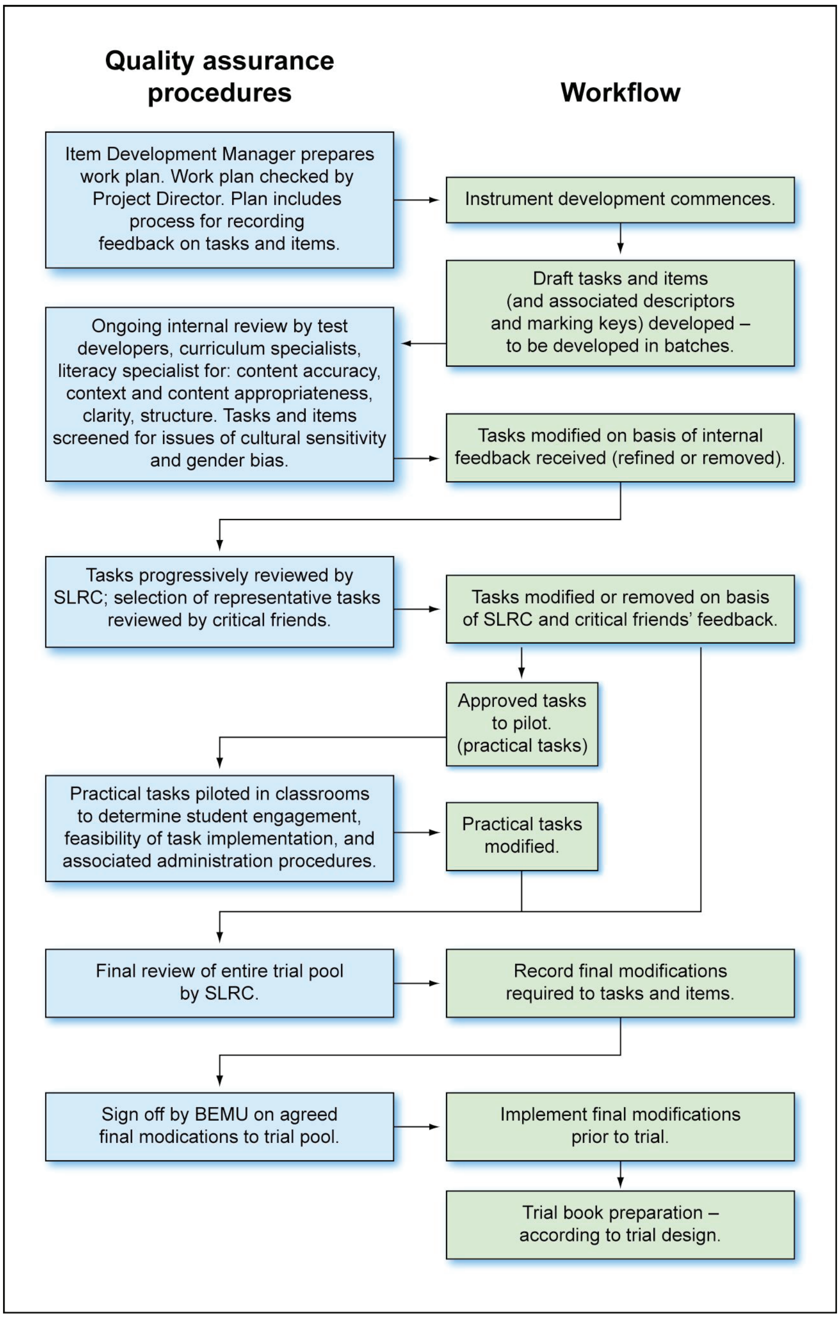
Booklet	Block 1	Block 2	Block 3
1	Cluster 1	Cluster 2	Cluster 3
2	Cluster 2	Cluster 3	Cluster 4
3	Cluster 3	Cluster 4	Cluster 5
4	Cluster 4	Cluster 5	Cluster 6
5	Cluster 5	Cluster 6	Cluster 7
6	Cluster 6	Cluster 7	Cluster 1
7	Cluster 7	Cluster 1	Cluster 2

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 work flow 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 levels and strands of the assessment domain) were developed at the same time as the items themselves, and reviewed accordingly.

EAA held review panels in-house prior to releasing materials for SLRC review. 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 Electronic Development of Items and Tests (EDIT) system, a secure item writing and reviewing application developed by EAA. Items were released progressively in batches during May, June and early August 2008. 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. Processes were also established for recording feedback on tasks and items as the review processes proceeded, and associated documentation was prepared. This documentation included unit templates and a spreadsheet tracking coverage by batch.

The EDIT review site allowed users to examine each item, provide detailed feedback and then rank it in order of priority for its inclusion in the final test. SLRC members could also enlist groups of people from their jurisdiction to review the items and submit the responses as feedback from the jurisdiction represented. All feedback was then collated and responded to by the test development team. The refined items were released to BEMU (now ACARA) for sign off prior to trial.

Table 2.2 Criteria used in SLRC item reviews

Criterion	Type
The Concept Area is appropriate.	Agree/Disagree
The Strand is appropriate.	Agree/Disagree
The Level is appropriate.	Agree/Disagree
The Descriptor is appropriate.	Agree/Disagree
The Key/Scoring Guide is appropriate. (NB: for multiple-choice questions check that there is only one possible answer)	Agree/Disagree
All distractors are plausible. (NB: This is only applicable for multiple-choice questions. For open response questions select N/A.)	Very Low/Low/High/Very High
The language demand (e.g. sentence length and structure, word familiarity, voice of verb phrase, etc) in the item is appropriate. The mathematics knowledge/skill needed to respond to the item is appropriate. (NB: Very High = appropriate, Very Low = inappropriate)	Very Low/Low/High/Very High
The question or task is clearly stated, the graphics are clear, the wording in the stem and options is clear and concise.	Very Low/Low/High/Very High
The science content is accurate in all parts of the item (including the diagrams, data tables and graphs) and the experimental design (where relevant) is sound.	Agree/Disagree
The item is not dependent on any other item in this item set.	Agree/Disagree

2.3.2 Pilot studies

Each practical task was piloted with at least two classes of students to ensure that the activities proposed and the associated administration procedures could be implemented with ease in Year 6 classroom settings. For some practical tasks, two versions of the task were piloted. A total of five schools participated in the pilot, with multiple classes in each setting. The pilot also established the degree to which the proposed tasks were engaging for students. All materials required to carry out the tasks were relatively simple in nature and were provided to schools by EAA. This was done to ensure that students were not disadvantaged by their potentially limited familiarity with specialist science equipment which is more likely to be found in secondary school laboratories.

2.3.3 Items delivered

A total of 223 items were released for review prior to trial, including 26 link items from 2006. These items were included to assess their suitability to serve as link items for the 2009 assessment. The brief for 2009 specified that the 2003 link items had to be included in the 2009 link item pool and thus they were not trialled.

The final pool of trial items developed was presented to and approved by the SLRC in August 2008.

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	43	0	43
Major concept area: EF	50	29	79
Major concept area: LT	30	15	45
Major concept area: M	42	14	56
Total	165	58	223
Strand A	26	20	46 (21%)
Strand B	64	32	96 (43%)
Strand C	75	6	81 (36%)
Total	165	58	223
Level 1	3	2	5 (2%)
Level 2	24	4	28 (13%)
Level 3	78	38	116 (52%)
Level 4	59	12	71 (32%)
Level 5	3	0	3 (1%)
Total	167	56	223
Multiple choice	81	8	89 (40%)
Short answer	55	47	102 (46%)
Extended response	29	3	32 (14%)
Total	165	58	223

EAA developed eight trial test booklets of 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 and a half clusters (cluster 8 and half of cluster 1) comprising link items drawn from the secure item pool from 2006.

2.3.4 Student Survey

An innovation for the 2009 National Assessment Program – Science Literacy was the inclusion of a Student Survey. Following discussions with BEMU a survey instrument was developed for trialling. Survey questions were written with scoring parameters and reviewed by the SLRC. Following feedback from the SLRC and BEMU, 46 questions were selected for trial. The areas covered by the survey were:

- students' perceptions of and attitudes towards science
- students' interests in science beyond the classroom
- students' experiences of science at school, including science concepts/topic areas studied.

The questions were produced on a scannable form and it was determined that the survey would be conducted following administration of the objective items and practical task.

2.4 Field trial of test items

Students from 30 selected schools across NSW, ACT, QLD and WA participated in the trial in October 2008. The SLRC representative in each of those states and territories provided support in obtaining permission to undertake research in schools. The trial schools were selected to reflect the range of educational contexts around the country, and included government, Catholic and independent schools; low and high socioeconomic drawing areas; metropolitan, regional and remote locations; large and small schools; and students from a variety of language backgrounds.

Approximately 970 students from the trial schools across the four selected states and territories participated in the trial. Each student completed one of the eight trial objective test papers and one of the four practical tasks. Within each class, teachers were asked to evenly distribute the eight objective test forms amongst students. On completion of the objective 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 one 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 procedures. A trained invigilator was sent to each trial school to deliver and collect the materials (to ensure the security of the materials) and to also observe and support the classroom teacher throughout the assessment. At the completion of each session the invigilator completed a session report form in conjunction with the classroom teacher, to provide feedback about various aspects of the trial. This feedback, in conjunction with a range of other sources of feedback, informed the selection and refinement of items for the final pool.

A team of experienced markers was engaged for a one-week period. Test developers from EAA trained the markers and remained on-site to oversee the marking process. On completion of marking of each cluster or practical task, a debrief session with the test developers was held and amendments were made to marking guides as necessary.

2.4.1 Analysis of the trial

In the first instance, the trial scores were data-entered and analysed by EAA's data analysis team using both ConQuest and RUMM2020 software. The results of the parallel analyses were consistent.

Key criteria for judging the performance of items were discrimination and measures of fit. Percentage correct was noted but only informed a decision to eliminate an item if other statistics were poor. Differential Item Functioning (DIF) for gender and Language Background Other Than English (LBOTE) were also considered.

EAA examined the item statistics and consulted with BEMU. Items with DIF were flagged but not automatically discarded. Two item sets relating to bushfires were removed for sensitivity reasons; other item sets were removed because the stimulus no longer supported a sufficient number of well-performing items. A pool of 146 well-performing items remained (which included items from only two of the four trial practical tasks).

2.4.1.1 Differential Item Functioning

By definition, Differential Item Functioning (DIF) refers to groups of students responding to an item differently, after adjusting for the groups' overall ability. For example, if a boy and a girl have the same ability, but the probability of success on an item for the girl is higher (or lower) than the probability of success for the boy, then the item exhibits 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 the National Assessment Program – Science Literacy were carried out using ConQuest by fitting a facets model where the interaction between an item and gender group is estimated. When the interaction term is significantly different from zero, at 95 per cent confidence level, an item is deemed to be showing DIF.

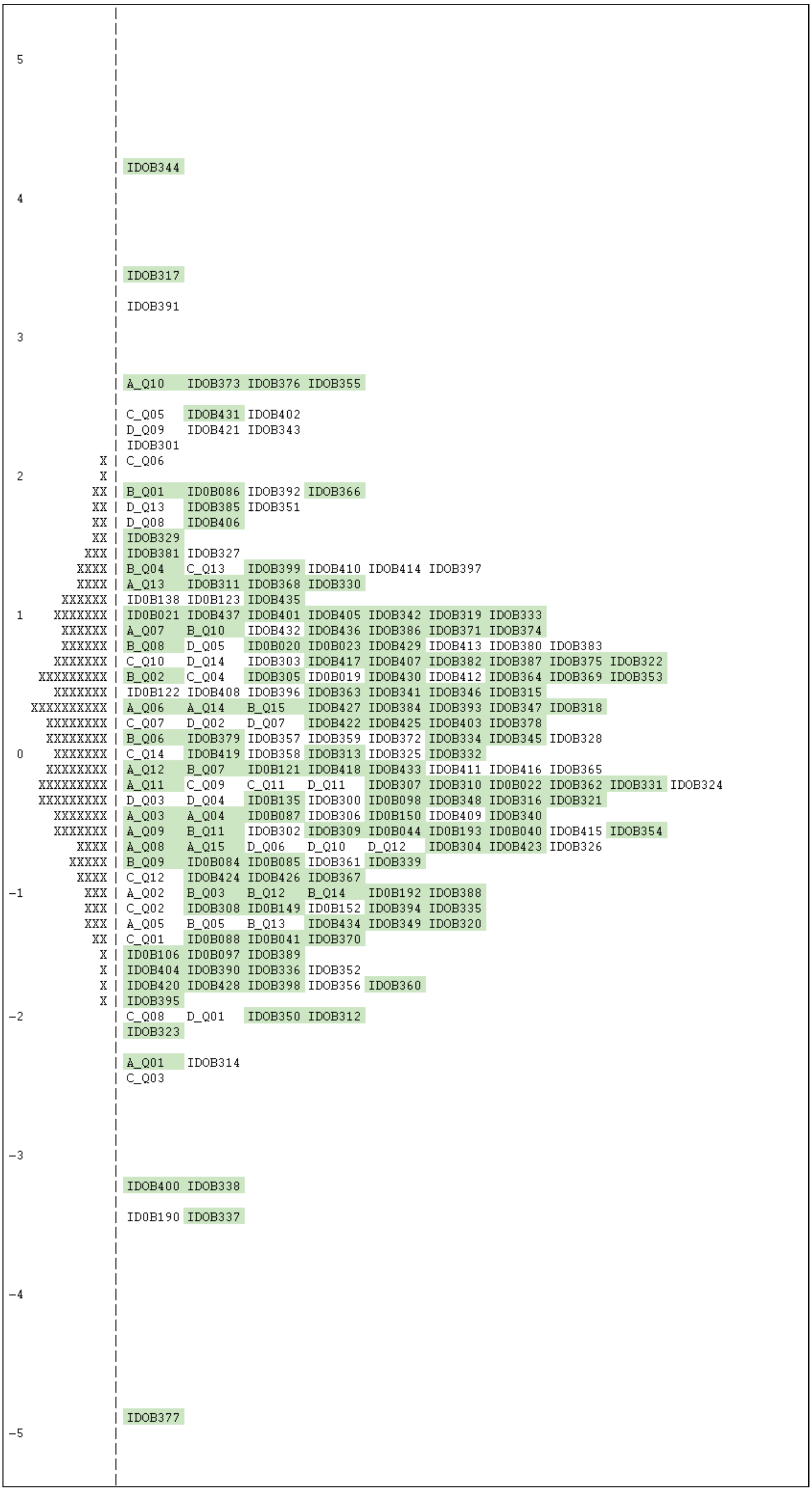
Items exhibiting DIF should not be automatically removed simply based on statistical evidence of bias. They should only be removed based on substantive reasoning. In some cases, it may well be the case that girls and boys do not perform in the same way across content areas in a subject domain, and such differential performance may be expected. Judgments should be made based on the importance of the skills tested in the specific items, and whether the inclusion of items showing DIF will bias the results in ways that are not consistent with the aims of the assessment.

The DIF findings were brought to the attention of subsequent reviewers (e.g. BEMU and the SLRC), to inform final item selection.

2.4.1.2 Item–person map

Figure 2.2 shows an item map produced from ConQuest output illustrating diagrammatically the distribution of all trialled items (indicated by item identifiers), and those comprising the 146 post-trial pool (shaded). The purpose of this diagram was to provide 'at a glance' the range of difficulty of the items and how they aligned with the ability of students in the trial pool (each 'X' represents 5.6 students). The left-hand side of the X axis represents persons, the right-hand side represents items and the Y axis represents the logistic scale. As can be seen, 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.

Figure 2.2 Item-person map for 146 post-trial items



2.4.2 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 2008. They contained a number of A4 sheets: one for each of the eight 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 objective and practical test

Items that were retained after the trial process for further consideration as possible items for the final test pool were provided to the SLRC to view on a refined version of EDIT. 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 details:

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

This pool was discussed at a meeting with the SLRC in Sydney on 21 April, 2009 and approved for use in the 2009 assessment. The SLRC recommended the deletion of one question and made suggestions for refinements in three diagrams. Where there were several consecutive items which had the same key, it also recommended changing the order of distractors.

In addition, the following changes to the Test Administrator's Manual based on invigilator feedback were discussed and agreed upon at the meeting:

- Inclusion of an instruction directing students to go back through the test and complete any items that they omitted. This was to encourage students to complete short and extended constructed response items.
- Inclusion of an instruction alerting students to the importance of information that precedes the actual questions.
- Inclusion of an instruction in the Practice Questions section that made explicit the difference between Multiple-Choice questions with only one correct answer and Tick the Boxes questions where multiple correct answers are possible.
- Inclusion of an announcement at the halfway point during the test, which would allow students to better manage their time.

EAA reviewed the SLRC feedback and further reduced the item pool. In addition, EAA developed a draft final list of preferred test items for 2009. The final pool containing 113 items was agreed as reflecting the best balance of items against the original specifications.

The final pool included nine link items from 2003 and 20 of the 26 2006 items that had been used in the trial.

The final pool of test items was presented to BEMU and approved for use in the 2009 testing.

2.5.2 Item selection for Student Survey

Students' responses to the survey questions were scanned and analysed. All items and results were presented to SLRC members in a secure document for their review. SLRC members were invited to comment on the items and provide a priority for inclusion in the final form. At the SLRC meeting on 21 April 2009, members discussed the findings and agreed on a final list of 30 survey questions that would appear as part of the main test.

2.6 Test characteristics of the final test

The actual distribution of items across the assessment domain for scientific literacy (strands and major concept areas) is shown in Table 2.4. There were 113 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	5	4	9	18
Strand B	20	3	18	41
Strand C	22	9	23	54
Total	47	16	50	113
Distribution of items by major science concept area				
Earth and Space (ES)	18	0	7	25
Energy and Force (EF)	9	5	17	31
Living Things (LT)	7	1	18	26
Matter (M)	13	10	8	31
Total	47	16	50	113

The final composition of the items (113) included in the sample test is shown by the series of tables to follow.

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	19	16	31	91
Practical	0	12	10	0	22
Total	25	31	26	31	113

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

Paper type	Strand			Total
	A	B	C	
Objective	12	27	52	91
Practical	6	14	2	22
Total	18	41	54	113

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	13	44	34	91
Practical	1	15	6	22
Total	14	59	40	113

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	44	15	32	91
Practical	3	1	18	22
Total	47	16	50	113

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	4.0 to 5.0	
Objective	1	8	7	11	13	7	17	8	4	0	4	1	1	82
Practical	1	0	3	5	2	4	4	1	1	0	1	0	0	22
Total	2	8	10	16	15	11	21	9	5	0	5	1	1	104

Note: Secure items from 2003 were not trialled in 2008 and are therefore not included in this table.

2.7 Reports to schools

Reports were developed and provided to schools that had participated in the 2009 Assessment, and were based on the reports used at trial. The reports were received in schools in December 2009. They 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 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 B. The school report includes a report for each objective booklet and a report for the practical task, *Which beak works best?*

Chapter 3
Sampling Procedures

3.1 Overview

The desired (target) population for the National Assessment Program – Science Literacy consisted of all students enrolled in Year 6 in Australian schools in 2009.

As defined in the tender specifications, the number of students sampled in each jurisdiction was to be 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 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, and consequently there would be too much administrative burden on the schools, particularly for the smaller schools.

Due to budgetary constraints, the nationwide achieved sample was to be approximately 13 000 students located within approximately 600 schools throughout Australia.

The sample design for the National Assessment Program – Science Literacy 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 an intact Year 6 class from the sampled schools selected in Stage 1.

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 2008 Year 6 enrolment sizes as supplied by MCEECDYA.

---

[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.

Generally, large scale sampling assessments of this type include provisions for excluding schools *before* sampling of schools takes place. This might be 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 and 2009, it was deemed desirable to include as many schools in the defined population as possible. Essentially this meant there were to be no school-level exclusions from the supplied sampling frame prior to sample selection. As such, the nationally defined population for the 2009 National Assessment Program – Science Literacy 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 2009 compared to 2003. Consequently, a slightly inflated sample size would be required to deal with this expected increase in non-response rate at the school level, so that the actual achieved number of schools and students in the sample was adequate.

In line with the procedures adopted in 2006, 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 2009 estimate of the number of educational institutions and students in the sampling frame for each jurisdiction, as provided by MCEECDYA.

Table 3.1 Estimated 2009 Year 6 enrolment figures as provided by MCEECDYA

State/Territory	Institutions	Students	Percentage of students
ACT	103	4501	1.7
NSW	2338	87 112	32.1
NT	150	3005	1.1
QLD	1379	56 879	21.0
SA	615	19 245	7.1
TAS	227	6756	2.5
VIC	1813	65 573	24.2
WA	883	28 017	10.3
Total	7508	271 088	100.0

Note: Some percentages may not add 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. Examples of non-participation include school remoteness, parental objection, etc. The 2009 study made provisions to document the reasons for school and student non-participation. Figure 3.1 illustrates the non-participation categories documented in the 2009 study whilst Table 3.2 details the exemption and refusal categories for non-participating schools and students.

Figure 3.1 National Assessment Program – Science Literacy 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 National Assessment Program – Science Literacy 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 +/– 0.1*s*, 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. As 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 National Assessment Program – Science Literacy specifications set the target sample at 12 000 students. The achieved precision of the statistics reported in the 2006 National

[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.

Assessment Program – Science Literacy was analysed in order to establish whether the sample size of approximately 12 000 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 +/– 0.1*s*. 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 2006

State/ Territory	Mean	SE	Confidence interval	Sample size	Desired confidence interval	Empirical design effect	Desired sample size
ACT	418	7.3	14.3	1271	10.0	6.8	2698
NSW	411	6.4	12.5	2039	10.3	7.9	3155
NT	325	17.2	33.7	740	16.1	8.5	3395
QLD	387	4.4	8.6	2016	9.7	4.2	1667
SA	392	5.1	10.0	1809	9.7	5.0	1988
TAS	406	6.1	12.1	1225	9.6	5.0	2001
VIC	408	5.2	10.2	1810	9.3	5.7	2281
WA	381	5.1	10.0	2001	9.7	5.6	2237
Total				12 911			19 422

As can be seen from Table 3.3, the analysis showed that the 2006 sample size for each state and territory was underestimated relative to the sample size that has the capacity to provide the stipulated precision. In order to rectify this problem it was proposed that the 2009 sample size be increased according to the magnitude of the design effect empirically established in 2006. However, this proposition was not approved by BEMU (now ACARA) and the brief was issued that the 2009 sample size should approximate that of 2006. Consequently the proposed target sample size for 2009 was set equal to that for 2006 and is shown in Table 3.4.

Table 3.4 Proposed 2009 sample sizes for drawing samples

State/Territory	Students	Schools
ACT	1400	59
NSW	2100	91
NT	950	49
QLD	2100	92
SA	2100	94
TAS	1400	63
VIC	2100	93
WA	2100	94
Total	14 250	635

3.5 Stratification

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 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) for a school was set to the 2008 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. 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 2006 (National Assessment Program – Science Literacy Technical Report 2006, 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 (2005) and IEA (2004). Both studies have different approaches for the treatment of small schools within the sampling frame. In the 2009 National Assessment Program – Science Literacy, OECD (2005) guidelines were utilised for classifying and stratifying small schools, whilst an adapted version of IEA's (2004) treatment of small school MOS values was used.

---

[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.

As a preliminary exercise, schools were classified into different sizes according to OECD (2005, p. 53) 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 sample size lower than the desired target sample size.

OECD (2005) 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 2005, pp. 53–57).

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 (2004) approach to ensure that selection of very small schools would not result in excessively large sampling weights (see IEA 2004, pp. 119–120, section 5.4.1).

Table 3.5 Proportions of schools by school size and jurisdiction

State/ Territory	School size	Number of schools	Percentage of schools	Number of students	Percentage of students
ACT	Large	74	71.8	4082	90.7
	Moderately small	18	17.5	352	7.8
	Very small	11	10.7	67	1.5
	Total	103	100.0	4501	100.0
NSW	Large	1390	59.5	76 900	88.3
	Moderately small	374	16.0	6999	8.0
	Very small	574	24.6	3213	3.7
	Total	2338	100.0	87 112	100.0
NT	Large	54	36.0	2212	73.6
	Moderately small	25	16.7	450	15.0
	Very small	71	47.3	343	11.4
	Total	150	100.0	3005	100.0
QLD	Large	759	55.0	50 827	89.4
	Moderately small	207	15.0	3796	6.7
	Very small	413	29.9	2256	4.0
	Total	1379	100.0	56 879	100.0
SA	Large	316	51.4	15 639	81.3
	Moderately small	132	21.5	2479	12.9
	Very small	167	27.2	1127	5.9
	Total	615	100.0	19 245	100.0
TAS	Large	121	53.3	5418	80.2
	Moderately small	53	23.3	996	14.7
	Very small	53	23.3	342	5.1
	Total	227	100.0	6756	100.0
VIC	Large	1070	59.0	56 773	86.6
	Moderately small	336	18.5	6255	9.5
	Very small	407	22.4	2545	3.9
	Total	1813	100.0	65 573	100.0
WA	Large	486	55.0	23 890	85.3
	Moderately small	142	16.1	2682	9.6
	Very small	255	28.9	1445	5.2
	Total	883	100.0	28 017	100.0

Note: Some percentages may not add 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 2004, p. 120, 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. For example, 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 OECD 2005 procedures (p. 60). 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 (2004). Classes with fewer than 20 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 2009 sample frame

Table 3.6 outlines the sample 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 C.

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

State/ Territory	Proposed target sample size for 2009	Number of schools by stratum					Total
		Very small	Moderately small	Large Catholic	Large govt	Large other	
ACT	1400	2	6	15	30	6	59
NSW	2100	7	9	15	52	8	91
NT	950	12	8	4	23	2	49
QLD	2100	8	8	12	55	9	92
SA	2100	9	15	14	44	12	94
TAS	1400	6	11	8	33	5	63
VIC	2100	7	11	16	50	9	93
WA	2100	10	11	13	51	9	94
Total	14 250	61	79	97	338	60	635

3.9 2009 National Assessment Program – Science Literacy sample results

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 2009 National Assessment Program – Science Literacy 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	1311	9.1	1199	9.1
NSW	2258	15.7	2092	15.9
NT	831	5.8	743	5.6
QLD	2228	15.5	2043	15.5
SA	2005	14.0	1848	14.0
TAS	1276	8.9	1167	8.9
VIC	2243	15.6	2040	15.5
WA	2208	15.4	2030	15.4
Total	14 360	100.0	13 162	100.0

Note: Numbers may not add to 100 due to rounding

The numbers of non-participation 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	100	0	7	3	1	111
NSW	158	0	2	3	3	166
NT	86	0	4	0	0	90
QLD	152	3	23	4	3	185
SA	140	0	6	2	8	156
TAS	89	4	11	4	0	108
VIC	166	1	17	5	14	203
WA	168	1	4	3	1	177
Total	1059	9	74	24	30	1196

The 2009 results were analysed in order to assess the magnitude of the misalignment between the sample size and the precision requirement of +/– 0.1*s* because the analysis of the design effect for 2006 showed that the planned sample size was not large enough to provide for the stipulated precision. As can be seen in Table 3.9, the 2009 sample size did not include a sufficient number of students in order to comply with the stipulated precision. This was also observed in 2006.

Table 3.9 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	6.0	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				13 162			20 158

Additional technical specifications can be found in Appendices E and F.

Chapter 4
Test Administration Procedures and Data Preparation

4.1 Registration of class/student lists

School Contact Officers nominated by those schools selected for the sample were informed that they were to register their students using the templates provided or, for a few jurisdictions, that this task had been done centrally. These procedures were designed so that student information could be collected, coded and then used for further analysis. Pre-registration also meant that test books could be overprinted with individual student details and their allocated practical task. 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 2009. The participating schools were sent the following materials: School Contact Officer's Manual (sent on behalf of MCEECDYA in July 2009 along with a brochure for parents explaining the assessment); Test Administrator's Manual and the assessment instruments, together with the appropriate practical materials for the particular task being undertaken. The Student Survey was also sent to all schools as part of the assessment instruments.

The assessment instruments were administered to a sample consisting of 4.86 per cent of the total Australian Year 6 student population. Tests were administered on the following dates:

- 14 October 2009 – Northern Territory, Queensland, Tasmania, Victoria
- 21 October 2009 – Australian Capital Territory, New South Wales, South Australia, Western Australia.

Students' regular class teachers administered the tests to minimise disruption to the normal class environment. Standardised administration procedures were developed and published in the Test Administrator's Manual. In all schools in which students were to complete the assessment, teachers and school administrators were provided with the Manual. Detailed instructions were also given in relation to the participation or exclusion of students with disabilities and students from non-English speaking backgrounds.

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

Teachers were required to complete a Student Participation Form, confirming details about any student who had not participated or had been excluded (see Appendix F 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 invigilators observing the administration of the Assessment in a random sample of classes in 32 (approximately 5 per cent) of the participating schools. The invigilators reported conformity with the administration procedures.

4.3 Marking procedures

The multiple-choice items had only one correct answer. The open-ended items required students to construct their own responses. The open-ended items were further categorised into those 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 or two marks depending on the quality or 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 books. 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 group leaders, the Test Development Manager and the Professional Leader). 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

Professional Leader. 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 ten per cent of the 2006 National Assessment Program – Science Literacy link items were also marked by the 2009 markers to assure 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 was performed that ensured accuracy in data capture.

Scanning software was used to capture images of all the student responses. These 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 F.

4.4.1 Data coding rules

Data coding rules for collecting student inclusion information in the SPF are explained in full on pages 9 to 11 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

Table 4.1 (Cont.) Codes used in the Student Participation Form

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 were based on procedures detailed in IEA 2004. The procedures outlined in TIMSS are designed for several different sampling scenarios. Only the procedures relevant to the National Assessment Program – Science Literacy 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, 635 schools were sampled of which there were 17 schools that did not participate in the testing (and could not be replaced). Four 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 four ineligible schools were not included in the calculation of this adjustment.[6]

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 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.

---

[6] See PISA 2003 Technical Report p. 111, TIMSS 2003 Sampling Weights and Participation Rates p. 202.

5.2.1 Class base weight

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

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

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 the National Assessment Program – Science Literacy 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:

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

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. 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[7], at the i^{th} school.

---

[7] 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:

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

5.5 Treatment of weights for schools that under-reported their number of classes

During the initial stages of preparation for calculating the sampling weights, it became apparent that some schools seem to have reported having a single Year 6 class in their submitted class lists when, according to the MCEECDYA enrolment estimates for 2009, they should have had sufficient students to form more than one class. This is the first time such under-reporting of class information has been observed in the National Assessment Program – Science Literacy. Information about the number of enrolled Year 6 students and the number of classes in each school was obtained from MCEECDYA and ACARA. The Year 6 enrolment from the 2009 MCEECDYA estimates were compared to actual enrolment figures held by ACARA, as obtained from the myschool.edu.au website, to determine the magnitude of the under-reporting of the number of students and classes. Inspection showed significant discrepancies between reported and actual enrolments for a number of schools. Consequently, a decision was made to identify schools for which the enrolment according to the class lists was smaller than one third of the enrolment as per the sampling frame. In total 32 schools were found to have under-reported their student and class figures. An adjusted class weight was calculated for these schools, based on their likely number of classes and their student enrolment data. The new class weight was set equal to the total Year 6 enrolments in the sample frame divided by the number of reported eligible students at the school. After such adjustment, final weights for these schools were calculated using the standard procedure. Further information is provided in the syntax file *NAPSL09_FinalSampleWeights.sps* (refer to Section 6.3 for details on how to obtain access to this file).

Chapter 6
Item Analysis of the Final Test

6.1 Item analyses

This chapter presents the item analyses of the 2009 National Assessment Program – Science Literacy main assessment data. Overall the items performed very well.

6.1.1 Sample size

In all, 13 162 students participated in at least one of the two components of the National Assessment Program – Science Literacy test: the paper-and-pencil test and the practical task. Table 6.1 shows the number of students by state and territory.

Table 6.1 Number of students by state and territory

State	Number of students
ACT	1199
NSW	2092
NT	743
QLD	2043
SA	1848
TAS	1167
VIC	2040
WA	2030
Total	13 162

6.1.2 Number of students by booklet

Seven test booklets with link items were rotated in each class (see Section 6.2 for 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.

Table 6.2 Number of students by test booklet

Booklet	Number of students
1	1866
2	1842
3	1856
4	1878
5	1891
6	1932
7	1897
Total	13 162

As each item appears in three test booklets, the number of students taking each item is around 5600.

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

```
                     persons      item
                 -------------------------------------
                               |81 96
                   3           |
                               |
                               |
                               |
                               |
                               |20 110
                               |
                               |
                               |1 53
                   2           |50 104
                              X|
                               |35 85 95
                              X|87
                             XX|
                              X|8 38 63 93
                             XX|2 17 79 94
                             XX|4 31 48 86
                           XXXX|84
                   1      XXXXX|88
                          XXXXX|41 76 102
                        XXXXXXX|32 43 59 109
                         XXXXXX|9 12 22 39 51
                        XXXXXXX|5 16 56 77 83
                        XXXXXXX|99 103 106
                       XXXXXXXX|26 74
                      XXXXXXXXX|23 47 55 112
Each 'X'           0 XXXXXXXXXX|27 52
represents             XXXXXXXX|10 34 82
78.1 cases            XXXXXXXXX|19 25 73 97 100
                       XXXXXXXX|14 21 70 80 90
                      XXXXXXXXX|42 46 64 75 89
                       XXXXXXXX|7 11 45 111
                       XXXXXXXX|6 29 30 40 71 113
                        XXXXXXX|18 58 65 67 69
                         XXXXXX|36 54 61 62 107
                  -1       XXXX|3
                           XXXX|24 57 66
                          XXXXX|
                            XXX|15 68 72
                             XX|33 44 92 98
                            XXX|91
                             XX|28
                              X|
                              X|60
                  -2          X|
                              X|49
                               |
                              X|108
                               |78
                               |101
                               |
                               |
                  -3           |13
                               |
                               |105
                               |
                               |
                               |37
```

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 -0.20 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 *NAPSL2009_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 113 items.

Table 6.3 Summary item statistics

Item label	Item reference number	Number of students	Percentage correct	Discrimination index	Fit mean square
A_Q01	13	6520	92.8	0.18	0.99
A_Q03	14	6520	54.8	0.24	1.09
A_Q04	15	6520	74.3	0.31	0.99
A_Q06	16	6520	37.2	0.33	0.98
A_Q07	17	6520	22.2	0.33	0.97
A_Q08	18	6520	63.2	0.29	1.03
A_Q09	19	6520	51.2	0.33	1.00
A_Q10	20	6520	15.1	0.34	0.99
A_Q12	21	6520	54.3	0.47	0.91
A_Q14	22	6520	34.7	0.26	1.05
B_Q01	1	6120	11.5	0.25	1.04
B_Q02	2	6120	22.7	0.28	0.98
B_Q03	3	6120	68.2	0.27	1.04
B_Q04	4	6120	21.0	0.42	0.99
B_Q06	5	6120	36.3	0.34	0.98
B_Q08	6	6120	61.5	0.34	0.99
B_Q09	7	6120	59.9	0.48	0.88
B_Q10	8	6119	18.6	0.31	1.07
B_Q11	9	6119	35.8	0.33	0.99
B_Q12	10	6119	48.7	0.23	1.12
B_Q14	11	6119	58.7	0.35	1.00

Table 6.3 (Cont.) Summary item statistics

Item label	Item reference number	Number of students	Percentage correct	Discrimination index	Fit mean square
B_Q15	12	6119	35.8	0.37	0.96
IDoB020	38	5594	20.4	0.21	1.05
IDoB021	39	5594	34.9	0.3	1.03
IDoB022	40	5594	60.2	0.42	0.96
IDoB023	41	5594	30.4	0.23	1.06
IDoB040	71	5681	61.7	0.44	0.94
IDoB041	44	5594	74.2	0.27	1.06
IDoB044	45	5594	58.4	0.31	1.03
IDoB084	61	5681	65.4	0.24	1.14
IDoB085	62	5681	65.9	0.41	0.98
IDoB086	63	5681	20.0	0.2	1.06
IDoB087	64	5681	55.4	0.43	0.93
IDoB088	65	5681	61.7	0.47	0.90
IDoB097	72	5681	71.9	0.44	0.96
IDoB098	73	5681	48.5	0.38	1.02
IDoB121	70	5681	51.6	0.44	0.94
IDoB135	46	5594	55.7	0.39	0.98
IDoB149	66	5681	68.4	0.44	0.94
IDoB150	67	5681	61.8	0.47	0.93
IDoB173	36	5594	65.1	0.26	1.08
IDoB174	37	5594	94.4	0.29	0.95
IDoB177	42	5594	54.8	0.39	1.04
IDoB178	43	5594	33.3	0.35	0.99
IDoB179	26	5556	41.7	0.26	1.07
IDoB180	27	5556	46.6	0.26	1.06
IDoB184	23	5556	44.9	0.36	1.10
IDoB185	24	5556	70.1	0.2	1.08
IDoB192	68	5681	73.1	0.47	0.92
IDoB193	69	5681	61.6	0.35	1.02
IDOB186	25	5556	50.6	0.22	1.11
IDOB304	54	5569	63.3	0.34	1.00
IDOB305	55	5569	41.3	0.23	1.13
IDOB308	28	5556	79.3	0.41	0.91
IDOB309	29	5556	61.7	0.38	0.98

Table 6.3 (Cont.) Summary item statistics

Item label	Item reference number	Number of students	Percentage correct	Discrimination index	Fit mean square
IDOB310	30	5556	60.7	0.33	1.01
IDOB311	31	5556	24.9	0.33	0.97
IDOB313	74	5705	42.1	0.3	1.03
IDOB315	75	5705	55.5	0.31	1.00
IDOB316	80	5705	53.6	0.34	1.02
IDOB317	81	5705	2.7	0.16	0.99
IDOB318	82	5705	47.6	0.24	1.11
IDOB319	83	5705	35.7	0.22	1.11
IDOB330	56	5569	34.0	0.26	1.07
IDOB331	57	5569	66.9	0.46	0.88
IDOB332	58	5569	60.3	0.42	0.94
IDOB333	59	5569	33.0	0.43	0.92
IDOB342	95	5687	23.6	0.42	0.96
IDOB344	96	5687	4.0	0.15	1.01
IDOB345	97	5687	48.7	0.34	1.03
IDOB353	34	5556	48.2	0.44	0.99
IDOB355	35	5556	15.4	0.33	0.95
IDOB360	98	5687	71.7	0.37	0.97
IDOB362	100	5687	50.6	0.38	1.00
IDOB363	99	5687	36.7	0.41	0.93
IDOB364	47	5569	41.4	0.52	0.94
IDOB366	48	5569	23.7	0.34	0.96
IDOB367	49	5569	83.9	0.35	0.94
IDOB368	76	5705	29.0	0.3	1.00
IDOB369	77	5705	33.6	0.36	0.97
IDOB370	78	5705	87.8	0.34	0.95
IDOB371	79	5705	22.5	0.32	0.98
IDOB373	50	5569	12.9	0.34	0.93
IDOB374	51	5569	34.0	0.39	0.96
IDOB375	52	5569	45.0	0.27	1.07
IDOB376	53	5569	11.4	0.27	0.97
IDOB381	84	5705	24.4	0.07	1.16
IDOB382	85	5705	15.6	0.34	0.95
IDOB390	60	5569	78.0	0.32	1.00

Table 6.3 (Cont.) Summary item statistics

Item label		Number of students	Percentage correct	Discrimination index	Fit mean square
IDOB395	105	5616	91.5	0.27	0.98
IDOB398	92	5687	73.8	0.29	0.99
IDOB399	93	5687	20.5	0.4	0.93
IDOB401	94	5687	22.0	0.33	0.98
IDOB403	32	5556	31.6	0.3	1.01
IDOB404	33	5556	74.4	0.39	0.95
IDOB405	86	5705	22.8	0.27	1.03
IDOB406	87	5705	16.8	0.22	1.03
IDOB417	88	5687	26.9	0.42	0.92
IDOB418	89	5687	56.0	0.38	0.98
IDOB419	90	5687	52.4	0.25	1.10
IDOB420	91	5687	75.7	0.22	1.07
IDOB422	106	5616	34.9	0.33	1.02
IDOB423	107	5616	60.9	0.43	0.93
IDOB425	112	5616	40.1	0.25	1.10
IDOB426	113	5616	58.4	0.46	0.89
IDOB428	108	5616	84.0	0.38	0.93
IDOB429	109	5616	33.2	0.3	1.01
IDOB431	110	5616	8.2	0.31	0.92
IDOB433	111	5616	55.8	0.3	1.04
IDOB434	101	5616	88.9	0.32	0.94
IDOB435	102	5616	29.3	0.33	0.99
IDOB436	103	5616	37.7	0.29	1.07
IDOB437	104	5616	13.3	0.32	0.96

6.1.3.3 Test reliability

Person separation reliability for the 2009 National Assessment Program – Science Literacy tests is 0.865, which is very acceptable[8].

[8] In comparison, the reported reliability for PISA 2003 mathematics is 0.85, and 0.89 for TIMSS 2003 Grade 8 mathematics.

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 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	-0.104	0.005
2	-0.005	0.005
3	0.032	0.005
4	0.069	0.005
5	0.006	0.006
6	0.019	0.005
7	-0.017	0.013

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 1 seems to be somewhat easier and booklet 4 appears to be more difficult than the other five booklets. However, in estimating the student Proficiency Levels, the booklet effect was taken into account. To do 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, the lowest discrimination index is 0.02 for item IDOB381 for SA and NSW. For this item, the detailed item statistics are shown in Figure 6.2. It can be seen from Figure 6.2 that both options 3 and 4 of this item strongly attracted students in SA and NSW. The item required students to identify an event which is <u>not</u> likely to be a consequence of global warming, i.e. that 'average temperatures are higher in October than in July in Adelaide'. Approximately one third of students in NSW and SA chose the incorrect option; 'There are larger and more frequent floods in southern Queensland'. This misunderstanding may be due to students recognising possible local flow on effects of global warming.

Figure 6.2 Item analysis for item IDOB381 for NSW and SA

```
NSW:    item: (IDOB381)
Cases for this item    898   Discrimination  0.02
Item Threshold(s):     0.00
Item Delta(s):        -0.00
------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1
------------------------------------------------------------------------------
   1       0.00      147      16.37   -0.07    -2.16(.031) -0.37     0.72
   2       0.00      141      15.70   -0.03    -0.97(.332) -0.33     0.73
   3       1.00      226      25.17    0.02     0.62(.533) -0.17     0.82
   4       0.00      304      33.85    0.16     4.78(.000) -0.13     0.66
   7       0.00       14       1.56   -0.00    -0.13(.900) -0.28     0.64
   9       0.00       17       1.89   -0.07    -2.08(.038) -0.60     0.76
   A       0.00       11       1.22   -0.04    -1.33(.185) -0.54     0.64
   B       0.00       38       4.23   -0.15    -4.58(.000) -0.84     0.82

SA:     item: (IDOB381)
Cases for this item    798   Discrimination  0.02
Item Threshold(s):     0.00
Item Delta(s):        -0.00
------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1
------------------------------------------------------------------------------
   1       0.00       95      11.90   -0.07    -2.05(.040) -0.43     0.60
   2       0.00      166      20.80   -0.06    -1.81(.071) -0.44     0.74
   3       1.00      213      26.69    0.02     0.57(.569) -0.23     0.67
   4       0.00      256      32.08    0.18     5.11(.000) -0.13     0.63
   7       0.00        8       1.00   -0.01    -0.20(.842) -0.15     0.64
   9       0.00       12       1.50   -0.04    -1.20(.229) -0.46     0.54
   A       0.00       10       1.25   -0.05    -1.31(.190) -0.58     0.67
   B       0.00       38       4.76   -0.15    -4.23(.000) -0.85     0.71
==============================================================================
```

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. A few items fall outside of the confidence interval (when comparing the state and territory location to the whole sample location). For further details please refer to the spreadsheet *NAPSL2009_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

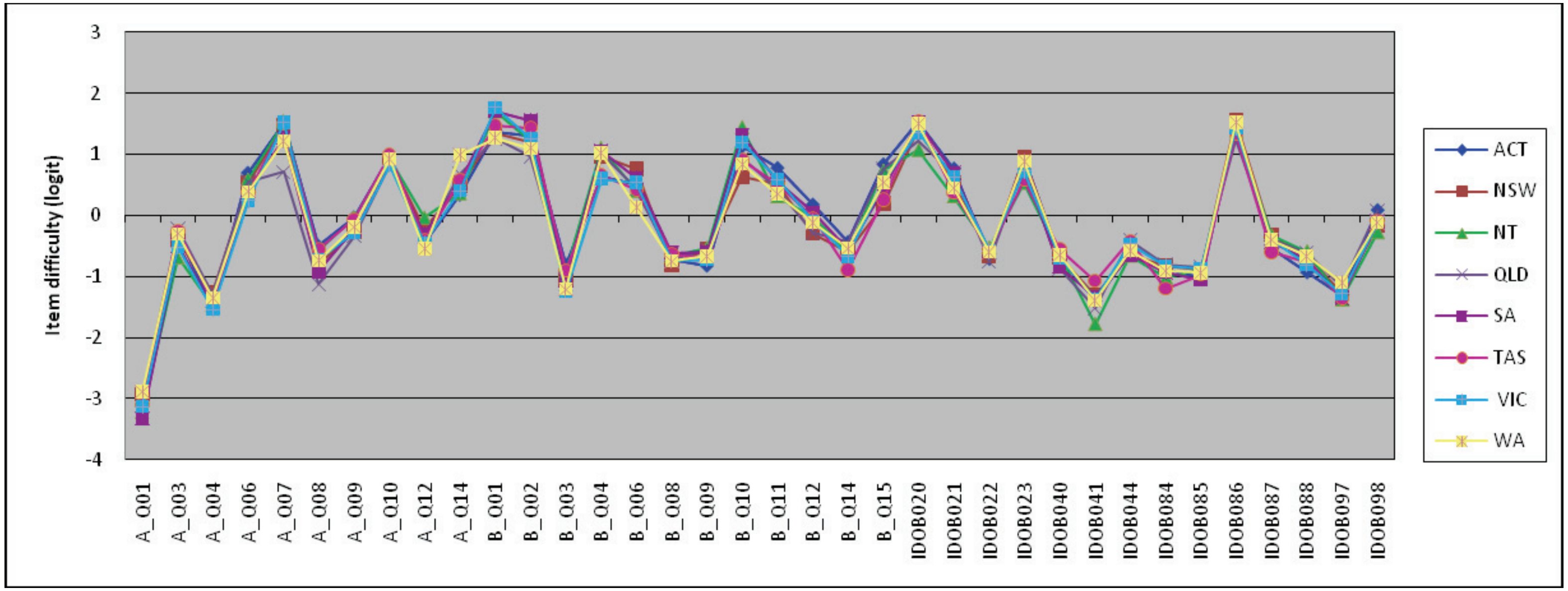


Figure 6.3 (Cont.) Comparison of item difficulty parameters across states and territories

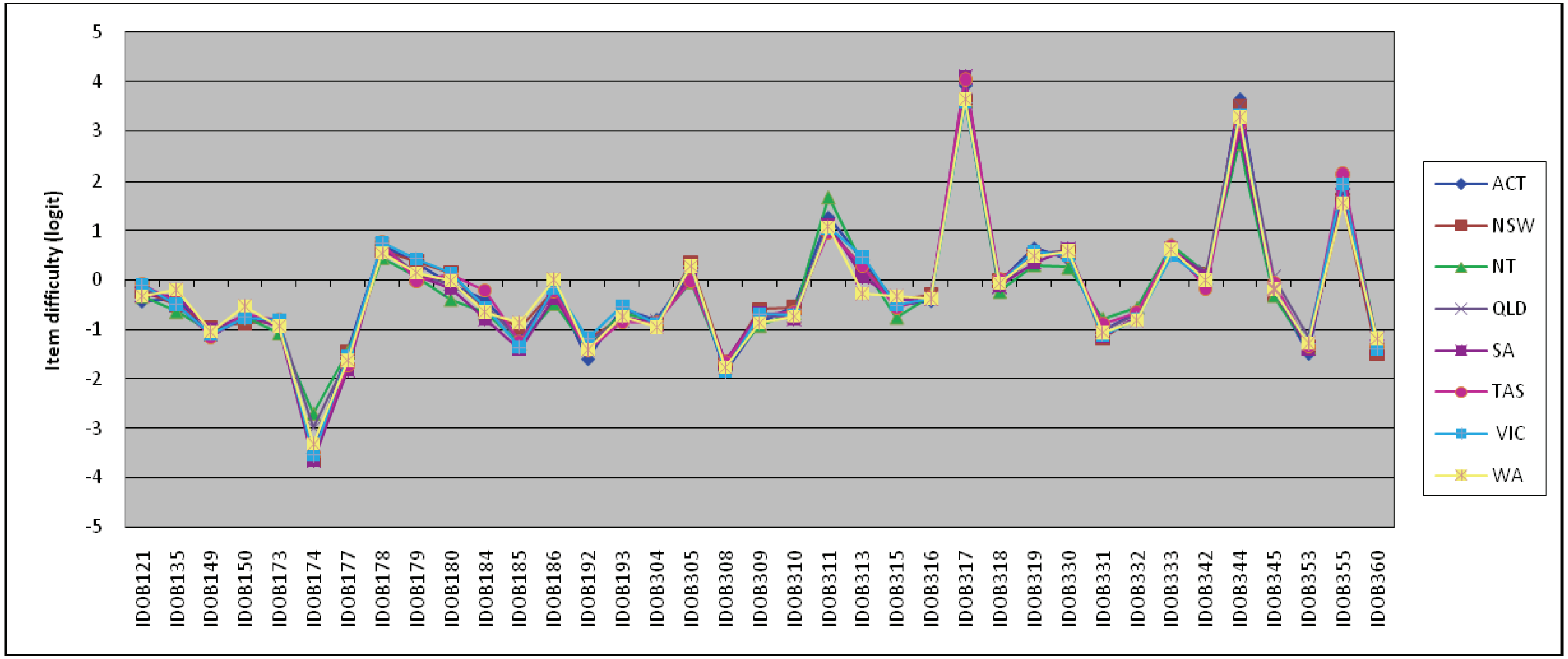


Figure 6.3 (Cont.) Comparison of item difficulty parameters across states and territories

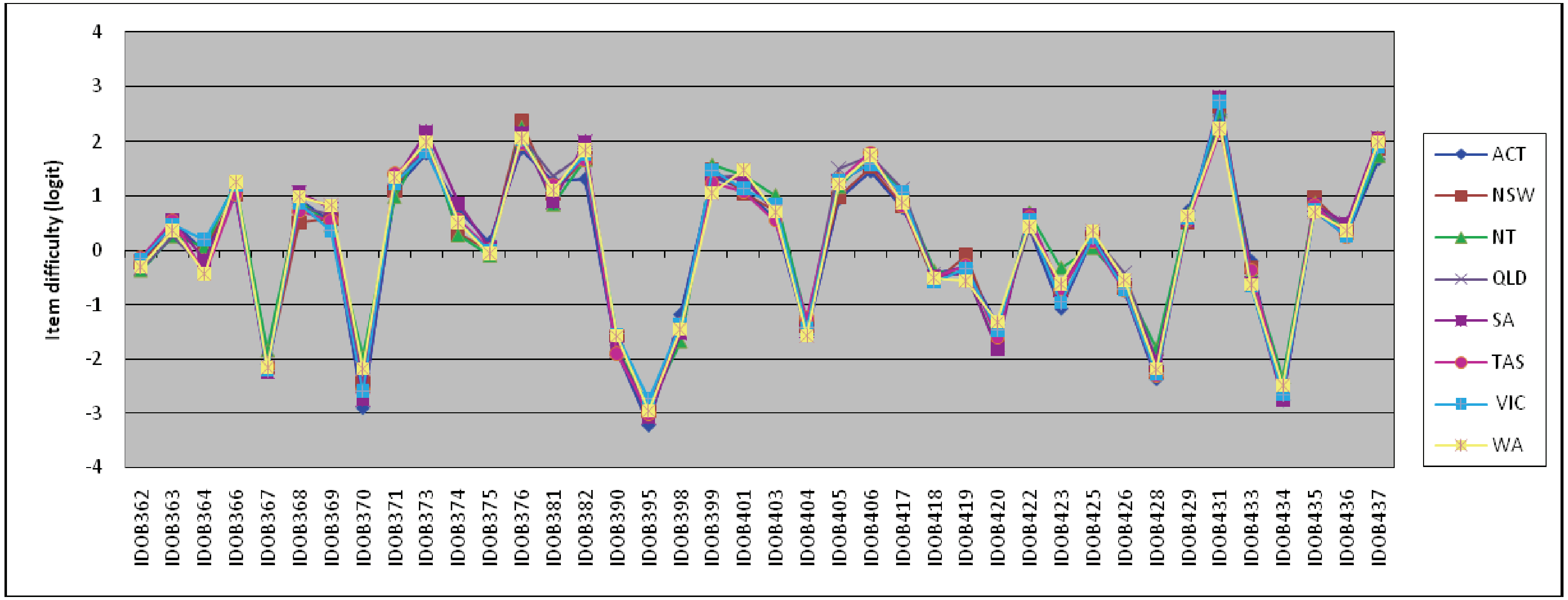


Figure 6.4 Discrimination index by state and territory

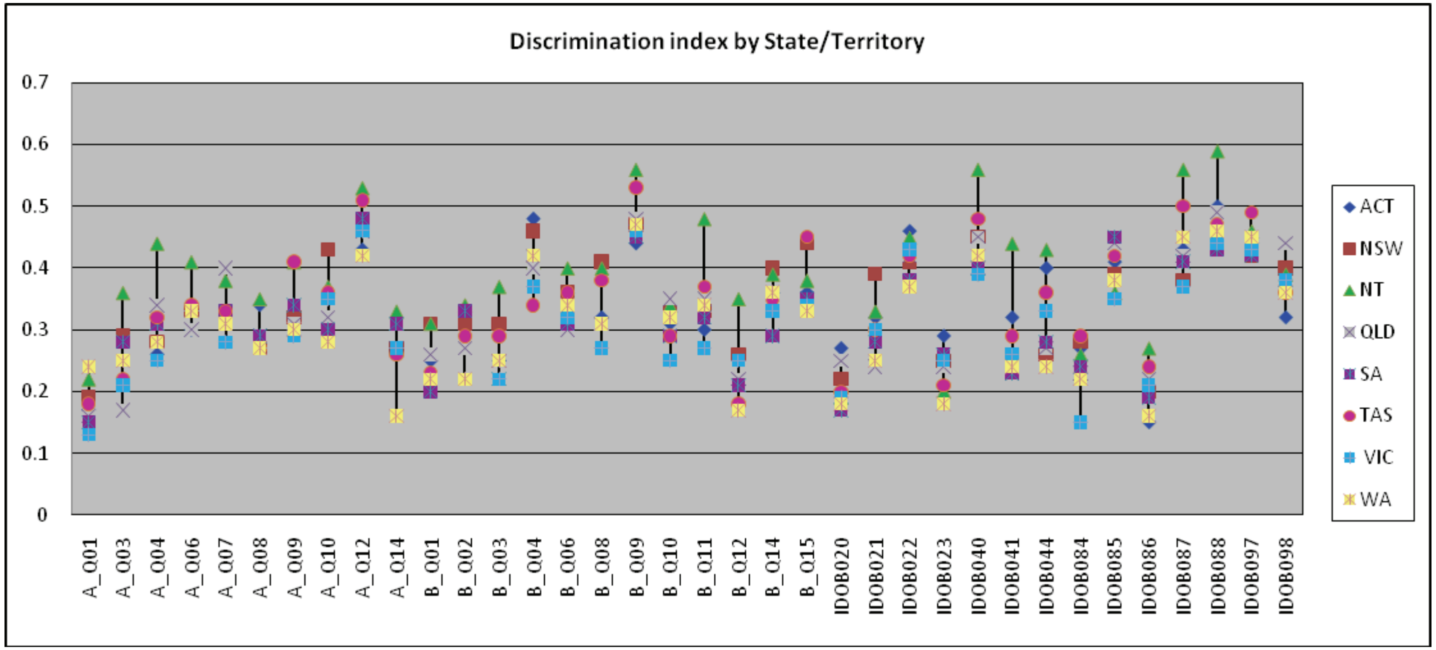


Figure 6.4 (Cont.) Discrimination index by state and territory

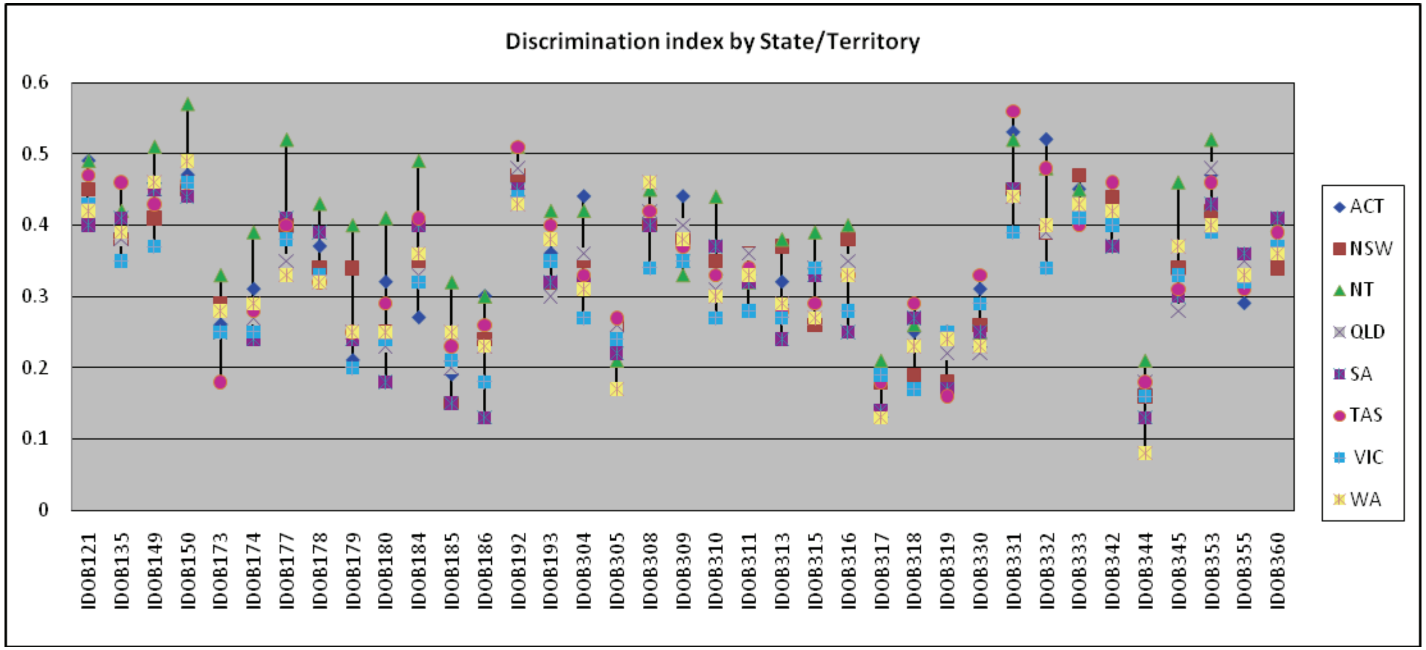
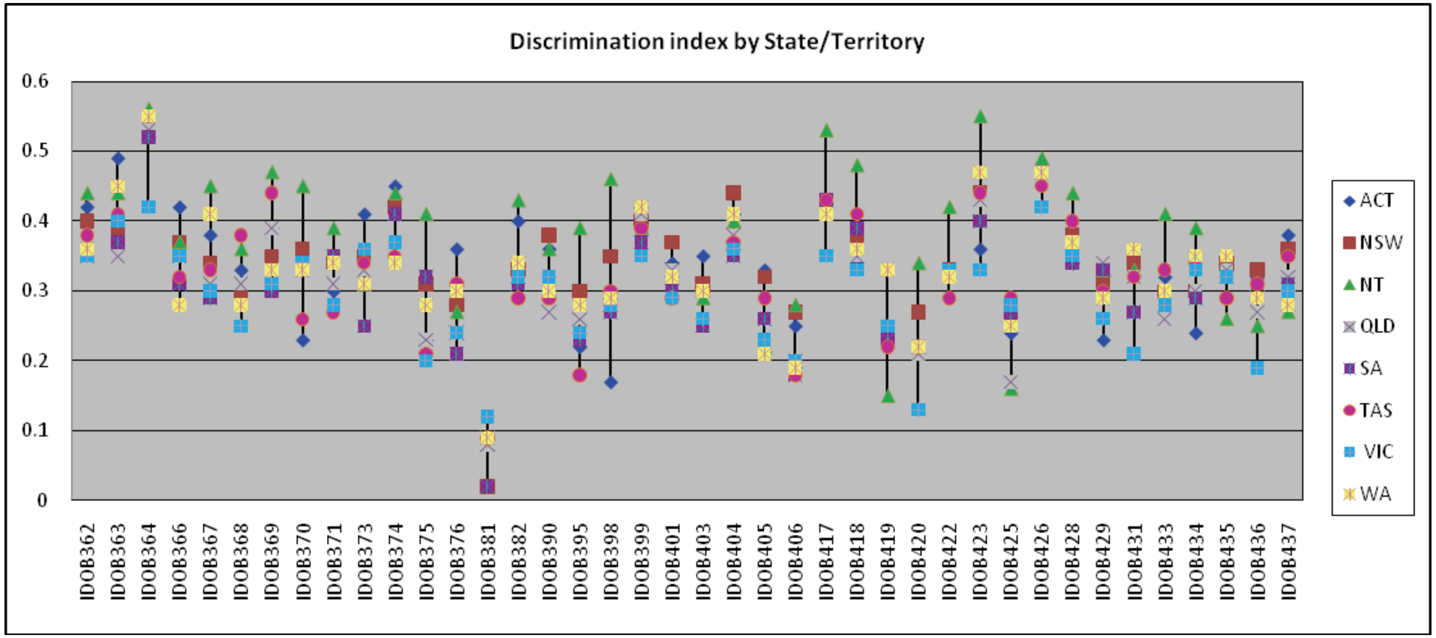


Figure 6.4 (Cont.) Discrimination index by state and territory



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 IDOB305, item IDOB318 and IDOB084, and girls performed better on item A_Q10, item IDOB192 and item IDOB174. 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 IDOB310 showed no difference 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
IDOB305	0.624	-0.206	0.830	A_Q10	2.223	2.792	-0.569
IDOB318	0.183	-0.388	0.571	IDOB192	-1.697	-1.129	-0.568
IDOB084	-0.617	-1.163	0.546	IDOB174	-3.809	-3.255	-0.554
IDOB375	0.301	-0.197	0.498	B_Q02	1.068	1.525	-0.457
IDOB121	-0.073	-0.469	0.396	IDOB401	1.035	1.483	-0.448
IDOB317	3.957	3.567	0.390	IDOB369	0.418	0.862	-0.444
IDOB390	-1.770	-2.159	0.389	B_Q06	0.352	0.721	-0.369
IDOB374	0.790	0.413	0.377	A_Q07	1.154	1.516	-0.362
IDOB304	-0.696	-1.060	0.364	B_Q03	-1.185	-0.827	-0.358
IDOB363	0.556	0.199	0.357	A_Q01	-3.170	-2.819	-0.351
IDOB362	-0.151	-0.508	0.357	IDOB423	-0.931	-0.606	-0.325
IDOB193	-0.596	-0.920	0.324	IDOB370	-2.716	-2.426	-0.290
IDOB353	0.101	-0.214	0.315	A_Q04	-1.473	-1.189	-0.284
IDOB020	1.609	1.304	0.305	A_Q09	-0.300	-0.019	-0.281
IDOB395	-3.057	-3.328	0.271	IDOB040	-0.924	-0.644	-0.280
IDOB085	-0.787	-1.051	0.264	B_Q12	-0.188	0.080	-0.268
IDOB344	3.442	3.181	0.261	B_Q10	1.285	1.530	-0.245
B_Q04	1.368	1.116	0.252	IDOB333	0.454	0.699	-0.245
IDOB149	-0.972	-1.205	0.233	IDOB308	-1.815	-1.573	-0.242
IDOB425	0.318	0.090	0.228	B_Q01	2.007	2.246	-0.239
IDOB345	-0.100	-0.300	0.200	IDOB428	-2.486	-2.266	-0.220
IDOB184	0.230	0.032	0.198	IDOB309	-0.791	-0.571	-0.220
IDOB319	0.569	0.379	0.190	A_Q06	0.405	0.606	-0.201
IDOB023	0.935	0.746	0.189	IDOB437	1.842	2.032	-0.190
IDOB435	0.931	0.745	0.186	IDOB376	2.052	2.233	-0.181

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

Boys performed better			
Code	**Girls**	**Boys**	**Diff**
IDOB433	-0.491	-0.668	0.177
IDOB173	-0.758	-0.933	0.175
B_Q14	-0.448	-0.616	0.168
IDOB179	0.355	0.193	0.162
IDOB405	1.250	1.097	0.153
IDOB404	-1.335	-1.486	0.151
IDOB331	-1.041	-1.171	0.130
IDOB431	2.569	2.440	0.129
IDOB315	-0.358	-0.468	0.110
IDOB021	0.670	0.561	0.109
IDOB398	-1.405	-1.511	0.106
A_Q14	0.682	0.578	0.104
IDOB087	-0.348	-0.449	0.101
IDOB180	0.092	-0.007	0.099
IDOB044	-0.526	-0.619	0.093
IDOB041	-1.379	-1.472	0.093
IDOB366	1.225	1.141	0.084
B_Q15	0.592	0.517	0.075
IDOB381	1.136	1.062	0.074
IDOB426	-0.726	-0.794	0.068
IDOB088	-0.691	-0.754	0.063
IDOB420	-1.544	-1.606	0.062
IDOB330	0.585	0.534	0.051
IDOB436	0.420	0.372	0.048
IDOB332	-0.767	-0.814	0.047
IDOB403	0.795	0.750	0.045
IDOB316	-0.350	-0.382	0.032
IDOB418	-0.473	-0.480	0.007
IDOB373	2.005	1.999	0.006

Girls performed better			
Code	**Girls**	**Boys**	**Diff**
IDOB422	0.431	0.610	-0.179
IDOB342	1.679	1.857	-0.178
IDOB185	-1.192	-1.016	-0.176
IDOB086	1.374	1.550	-0.176
B_Q09	-0.673	-0.499	-0.174
B_Q11	0.476	0.634	-0.158
A_Q08	-0.807	-0.649	-0.158
IDOB311	1.074	1.232	-0.158
IDOB097	-1.463	-1.321	-0.142
IDOB022	-0.683	-0.549	-0.134
IDOB419	-0.381	-0.254	-0.127
IDOB367	-2.183	-2.057	-0.126
IDOB434	-2.731	-2.608	-0.123
IDOB364	0.143	0.249	-0.106
IDOB313	0.170	0.269	-0.099
IDOB150	-0.783	-0.690	-0.093
IDOB429	0.536	0.628	-0.092
IDOB177	-0.472	-0.383	-0.089
IDOB399	1.319	1.396	-0.077
IDOB371	1.215	1.283	-0.068
A_Q03	-0.355	-0.293	-0.062
IDOB368	0.848	0.901	-0.053
IDOB178	0.653	0.706	-0.053
IDOB406	1.576	1.615	-0.039
IDOB135	-0.462	-0.429	-0.033
IDOB098	-0.160	-0.128	-0.032
IDOB417	0.957	0.982	-0.025
B_Q08	-0.673	-0.651	-0.022
IDOB355	1.796	1.817	-0.021
IDOB360	-1.481	-1.463	-0.018
IDOB186	-0.148	-0.134	-0.014
IDOB382	1.699	1.712	-0.013
A_Q12	-0.302	-0.299	-0.003

6.1.8 Impact of item type on student performance

In contrast to 2006, when the proportions of students omitting responses to extended-response items were approximately double those omitting responses to multiple-choice items, in 2009 the difference in the omit rate overall was relatively smaller with the exception of the Northern Territory, as can be seen in Table 6.6.

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	4.79	7.74	6.38
	Males	4.13	8.49	7.29
NSW	Females	4.04	6.58	5.17
	Males	4.61	7.86	6.38
NT	Females	7.75	14.79	11.61
	Males	8.75	14.93	11.53
QLD	Females	4.99	7.56	5.83
	Males	5.95	9.28	7.51
SA	Females	4.38	8.52	6.82
	Males	3.14	7.80	6.65
TAS	Females	3.35	6.79	5.06
	Males	4.45	9.09	7.17
VIC	Females	3.94	6.44	5.20
	Males	3.67	7.17	5.73
WA	Females	4.45	7.96	5.76
	Males	3.71	8.45	6.54
Total	Females	4.51	7.78	6.07
	Males	4.50	8.62	6.96

It is interesting to note that the omit rates appear to be somewhat higher for short-answer items compared to the omit rates for extended-response items.

6.2 Test design

6.2.1 Sample test design

Each 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
A_Q01	Practical	AQ01			Prac 1	Which beak works best?
A_Q03	Practical	AQ02			Prac 2	Which beak works best?
A_Q04	Practical	AQ03			Prac 3	Which beak works best?
A_Q06	Practical	AQ04			Prac 4	Which beak works best?
A_Q07	Practical	AQ05			Prac 5	Which beak works best?
A_Q08	Practical	AQ06			Prac 6	Which beak works best?
A_Q09	Practical	AQ07			Prac 7	Which beak works best?
A_Q10	Practical	AQ08			Prac 8	Which beak works best?
A_Q12	Practical	AQ09			Prac 9	Which beak works best?
A_Q14	Practical	AQ10			Prac 10	Which beak works best?
B_Q01	Practical	BQ01				
B_Q02	Practical	BQ02				
B_Q03	Practical	BQ03				
B_Q04	Practical	BQ04				
B_Q06	Practical	BQ05				
B_Q08	Practical	BQ06				
B_Q09	Practical	BQ07				
B_Q10	Practical	BQ08				
B_Q11	Practical	BQ09				
B_Q12	Practical	BQ10				
B_Q14	Practical	BQ11				
B_Q15	Practical	BQ12				
IDOB020	Objective	B1Q15	B2Q03	B7Q29		
IDOB021	Objective	B1Q16	B2Q04	B7Q30		
IDOB022	Objective	B1Q17	B2Q05	B7Q31		
IDOB023	Objective	B1Q18	B2Q06	B7Q32		
IDOB040	Objective	B1Q10	B6Q37	B7Q24		
IDOB041	Objective	B1Q21	B2Q09	B7Q35		
IDOB044	Objective	B1Q22	B2Q10	B7Q36		
IDOB084	Objective	B1Q01	B6Q28	B7Q15		
IDOB085	Objective	B1Q02	B6Q29	B7Q16		
IDOB086	Objective	B1Q03	B6Q30	B7Q17		
IDOB087	Objective	B1Q04a	B6Q31a	B7Q18a		
IDOB088	Objective	B1Q04b	B6Q31b	B7Q18b		
IDOB097	Objective	B1Q11	B6Q38	B7Q25		
IDOB098	Objective	B1Q12	B6Q39	B7Q26		
IDOB121	Objective	B1Q09	B6Q36	B7Q23		
IDOB135	Objective	B1Q23	B2Q11	B7Q37		
IDOB149	Objective	B1Q05	B6Q32	B7Q19		
IDOB150	Objective	B1Q06	B6Q33	B7Q20		
IDOB173	Objective	B1Q13	B2Q01	B7Q27		
IDOB174	Objective	B1Q14	B2Q02	B7Q28		
IDOB177	Objective	B1Q19	B2Q07	B7Q33		
IDOB178	Objective	B1Q20	B2Q08	B7Q34		
IDOB179	Objective	B1Q27	B2Q15	B3Q04		
IDOB180	Objective	B1Q28	B2Q16	B3Q05		

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

Item label	Paper	Block 1	Block 2	Block 3	SRM question number	Unit title
IDOB184	Objective	B1Q24	B2Q12	B3Q01		
IDOB185	Objective	B1Q25	B2Q13	B3Q02		
IDOB192	Objective	B1Q07	B6Q34	B7Q21		
IDOB193	Objective	B1Q08	B6Q35	B7Q22		
IDOB186	Objective	B1Q26	B2Q14	B3Q03		
IDOB304	Objective	B2Q32	B3Q21	B4Q08	5	Lifting weights
IDOB305	Objective	B2Q33	B3Q22	B4Q09	6	Lifting weights
IDOB308	Objective	B1Q29	B2Q17	B3Q06		
IDOB309	Objective	B1Q30	B2Q18	B3Q07		
IDOB310	Objective	B1Q31	B2Q19	B3Q08		
IDOB311	Objective	B1Q32	B2Q20	B3Q09		
IDOB313	Objective	B5Q27	B6Q14	B7Q01		
IDOB315	Objective	B5Q28	B6Q15	B7Q02		
IDOB316	Objective	B5Q33	B6Q20	B7Q07	19	Phases of the Moon
IDOB317	Objective	B5Q34	B6Q21	B7Q08	20	Phases of the Moon
IDOB318	Objective	B5Q35	B6Q22	B7Q09	21	Phases of the Moon
IDOB319	Objective	B5Q36	B6Q23	B7Q10	22	Phases of the Moon
IDOB330	Objective	B2Q34	B3Q23	B4Q10	7	Separating mixtures
IDOB331	Objective	B2Q35	B3Q24	B4Q11	8	Separating mixtures
IDOB332	Objective	B2Q36	B3Q25	B4Q12	9	Separating mixtures
IDOB333	Objective	B2Q37	B3Q26	B4Q13	10	Separating mixtures
IDOB342	Objective	B4Q35	B5Q21	B6Q08	14	Heating and cooling
IDOB344	Objective	B4Q36	B5Q22	B6Q09	15	Heating and cooling
IDOB345	Objective	B4Q37	B5Q23	B6Q10	16	Heating and cooling
IDOB353	Objective	B1Q35	B2Q23	B3Q12	23	Using and saving energy
IDOB355	Objective	B1Q36	B2Q24	B3Q13	24	Using and saving energy
IDOB360	Objective	B4Q38	B5Q24	B6Q11		
IDOB362	Objective	B4Q40	B5Q26	B6Q13		
IDOB363	Objective	B4Q39	B5Q25	B6Q12		
IDOB364	Objective	B2Q25	B3Q14	B4Q01	27	Collecting ants
IDOB366	Objective	B2Q26	B3Q15	B4Q02	28	Collecting ants
IDOB367	Objective	B2Q27	B3Q16	B4Q03	29	Collecting ants
IDOB368	Objective	B5Q29	B6Q16	B7Q03		
IDOB369	Objective	B5Q30	B6Q17	B7Q04		
IDOB370	Objective	B5Q31	B6Q18	B7Q05		
IDOB371	Objective	B5Q32	B6Q19	B7Q06		
IDOB373	Objective	B2Q28	B3Q17	B4Q04	30	Tomato plants
IDOB374	Objective	B2Q29	B3Q18	B4Q05	31	Tomato plants
IDOB375	Objective	B2Q30	B3Q19	B4Q06	32	Tomato plants
IDOB376	Objective	B2Q31	B3Q20	B4Q07	33	Tomato plants
IDOB381	Objective	B5Q37	B6Q24	B7Q11	25	Climate change
IDOB382	Objective	B5Q38	B6Q25	B7Q12	26	Climate change
IDOB390	Objective	B2Q38	B3Q27	B4Q14	1	Native and introduced animals
IDOB395	Objective	B3Q32	B4Q19	B5Q05	2	Energy efficient light bulbs
IDOB398	Objective	B4Q32	B5Q18	B6Q05	11	Cola fountain
IDOB399	Objective	B4Q33	B5Q19	B6Q06	12	Cola fountain
IDOB401	Objective	B4Q34	B5Q20	B6Q07	13	Cola fountain
IDOB403	Objective	B1Q33	B2Q21	B3Q10		
IDOB404	Objective	B1Q34	B2Q22	B3Q11		

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

Item label	Paper	Block 1	Block 2	Block 3	SRM question number	Unit title
IDOB405	Objective	B5Q39	B6Q26	B7Q13		
IDOB406	Objective	B5Q40	B6Q27	B7Q14		
IDOB417	Objective	B4Q28	B5Q14	B6Q01		
IDOB418	Objective	B4Q29	B5Q15	B6Q02		
IDOB419	Objective	B4Q30	B5Q16	B6Q03		
IDOB420	Objective	B4Q31	B5Q17	B6Q04		
IDOB422	Objective	B3Q33	B4Q20	B5Q06	3	Water resources
IDOB423	Objective	B3Q34	B4Q21	B5Q07	4	Water resources
IDOB425	Objective	B3Q39	B4Q26	B5Q12	17	Greenhouse gas emissions
IDOB426	Objective	B3Q40	B4Q27	B5Q13	18	Greenhouse gas emissions
IDOB428	Objective	B3Q35	B4Q22	B5Q08	34	Burning foods
IDOB429	Objective	B3Q36	B4Q23	B5Q09	35	Burning foods
IDOB431	Objective	B3Q37	B4Q24	B5Q10	36	Burning foods
IDOB433	Objective	B3Q38	B4Q25	B5Q11	37	Burning foods
IDOB434	Objective	B3Q28	B4Q15	B5Q01		
IDOB435	Objective	B3Q29	B4Q16	B5Q02		
IDOB436	Objective	B3Q30	B4Q17	B5Q03		
IDOB437	Objective	B3Q31	B4Q18	B5Q04		

Note: Unit titles are shown for items which appear in the 2009 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 MCEECDYA Secretariat at *enquiries@mceecdya.edu.au*. Relevant data files are listed throughout the 2009 Technical 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, 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 2009 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 in which equal numbers of respondents from each jurisdiction were included. See section 7.2 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 743 responses were included in the calibration sample. For each of the other jurisdictions, a random sample of 743 responses was selected. Consequently, the calibration sample consisted of 5944 (=743×8) 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 Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA) Secretariat at *enquiries@mceecdya.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. IDOB186) 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 10 and 12 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	7	Unique identifier for the student record
Item responses	8 to 120 (113 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 2005, p. 198). However, as there is only one domain in the 2009 National Assessment Program – Science Literacy (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 2009 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 (*NAPSL2009_PV_2010-04-13.sav*. See Appendix G 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 PISA 2003 Technical Report for a description of the model (OECD 2005).

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 *NAPSL2009_Produce_2009_PV.cqc*, which is shown in Appendix H.

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 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 2009 National Assessment Program – Science Literacy, 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 2009 National Assessment Program – Science Literacy, replication methods such as Balanced Repeated Replicate (BRR) or Jackknife are used to compute standard errors (Rust & Rao 1996). In the 2009 National Assessment Program – Science Literacy, the Jackknife method was used. Jackknife replication weights are computed (variables RW1 to RW318 in the file *NAPSL2009_PV_2010-04-13.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 (measurement error) provides an estimate of the standard error for the estimated statistic. SPSS macros were written to carry 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 National Assessment Program – Science Literacy 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 away a score is from the mean.

As part of the equating process (refer to Chapter 8 for details), the 2009 logit scores are first translated to the 2006 scale, then transformed to the 400/100 scale. The transformation used in 2009 is given below.

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

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 2009 Results to 2006 Results

8.1 Setting 2006 results as the baseline

While the first cycle of the National Assessment Program – Science Literacy was conducted in 2003 (then known as PSAP), and the 2006 assessment was the second round of the National Assessment Program – Science Literacy, 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 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 2009 results to 2006 results

As a consequence of the decision to use 2006 results as the baseline, 2009 results were equated to 2006 results. To carry out the equating, link items between the 2009 and 2006 tests were used.

8.2.1 Link items

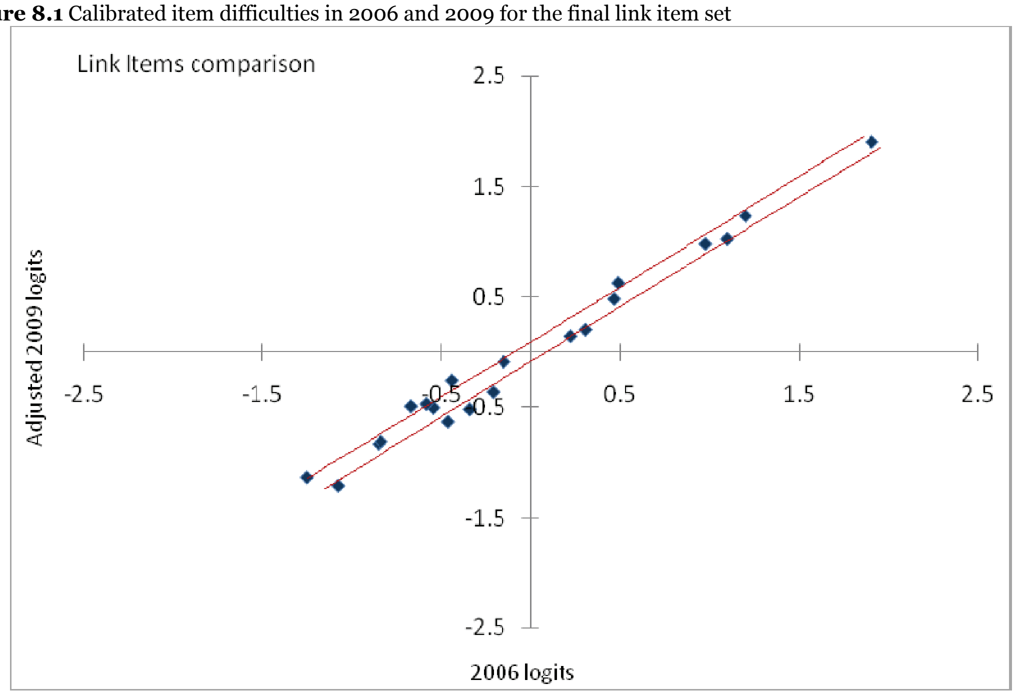
The adopted equating methodology frame proposed that around 25 items from the 2006 test be embedded in the 2009 test as potential link items in addition to nine items from the 2003 test that were used to conduct equating between 2003 and 2006. Care was taken to find items that performed well statistically and also covered the range of scientific literacy strands A, B and C and the science concept areas.

8.2.2 Link item selection

The selection process for the final set of link items to conduct equating between 2006 and 2009 consisted of two parts. In the first part, the list of items was refined based on the comparisons of item locations in 2009 and 2006. In the second part, the final set of items was inspected by a content expert from the test construction team. The purpose of the latter was to ensure that the selected link items provided adequate coverage across the scientific literacy strands.

In the first part, the 2009 location of link items was independently estimated. In order to conduct comparisons of item locations between 2006 and 2009, the 2009 locations were adjusted to have the same mean and standard deviation as observed in 2006. In the first refinement step two clear outliers in terms of overall item difficulty were removed from the set. In the second step, all items with a difference between the 2006 location and the 2009 location greater than 0.3 logits were removed. After readjusting the 2009 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 20 link items consisting of seven 2003 link items and thirteen 2006 link items. A plot of 2006 and 2009 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 2009 for the final link item set



8.2.3 Equating procedures

The 2009 data were scaled and item parameters obtained. Using the 2009 item parameters as anchors for common items, the 2006 data were scaled and population parameters (mean and variance of the ability distribution for 2006) were produced. The mean and variance from this new scaling and the mean and variance of ability distribution from the 2006 scaling (using 2006 item parameters) were then compared. A transformation was derived from mapping the mean and variance of the 2006 ability distribution obtained using 2009 item parameters onto the mean and variance of the 2006 ability distribution obtained using 2006 item parameters. This transformation was used to place the 2009 results onto the 2006 scale.

8.3 Equating transformation

The result of the equating process was the derivation of a transformation formula for the 2009 results to be placed on the 2006 scale. This equation is given below.

2009 result on 2006 scale = ((2009 logit – (-0.1004) / 0. 9476) * 0.9545) + 0.2005

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:

2009 standard error on 2006 scale in logit = ((2009 S.E. in logit) / 0. 9476) * 0.9545

8.4 Link error

In establishing trends from 2006 to 2009, it is necessary to make judgments about the statistical significance of the difference in science achievement between 2009 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 2003 (OECD 2005). Firstly calibrate the items using 2009 and 2006 data separately. If the link items behave exactly the same way in 2009 and 2006 (and they follow the Rasch model), there should only be a constant difference between 2009 and 2006 item parameters for matched items. However, in real life, items will vary from 2009 to 2006 and some items will vary more than others.

The link error is 0.035165 logits; transformed to the scientific literacy scale it is equal to a scaled score of 3.68.

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 MCEECDYA Secretariat at *enquiries@mceecdya.edu.au*.

The link error is used only when comparisons between 2009 and 2006 results are made. For example, to test whether the mean achievement in 2009 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

	2009 Mean on 2006 scale & S.E.	**2006 Mean & S.E.**	**2009 Mean – 2006 Mean**	**Standard error of difference**	**Standardised difference**
NSW	396 (6.17)	411 (6.38)	-15	$\sqrt{6.17^2 + 6.38^2 + 3.68^2}$	-1.56 = -15 /9.61 (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 Link error

In 2006, cut-points along the proficiency scale were decided after consultations with science experts. In 2009, the same cut-points were used, as shown in Table 9.1.

Table 9.1 Cut-points for the 2009 National Assessment Program – Science Literacy

Level	2006 cut-points[9]	Transformed to 400/100 scale[10]
2 and below	<−1.113 89	262.2932
3.1	0.129 692	392.5772
3.2	1.373 269	522.8611
3.3	2.616 846	653.145
4.0	>2.616 846	>653.145

As for 2003 and 2006, 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] The transformation used is (2009 logit−−0.100467)/0.947624*0.954513216+0.200543797.

[10] The transformation used is scaled score=(2009 logit−0.200543797)/0.954513216*100+400.

9.2 Proficiency Levels of items

Table 9.2 shows the 2009 National Assessment Program – Science Literacy items and their corresponding levels on the proficiency scale.

Table 9.2 Proficiency Levels of items

Item	2009 difficulty	2009 item difficulty after adjustment for RP	Operational Level	Design level	Scaled score
A_Q01	-2.874	-1.970	2 and below	1	173
A_Q03	-0.315	0.608	3.2	3	443
A_Q04	-1.304	-0.388	3.1	3	338
A_Q06	0.474	1.403	3.3	3	526
A_Q07	1.386	2.321	3.3	3	622
A_Q08	-0.666	0.254	3.2	3	406
A_Q09	-0.129	0.795	3.2	3	462
A_Q10	2.432	3.375	4 and above	3/4	733
A_Q12	-0.273	0.650	3.2	3	447
A_Q14	0.618	1.548	3.3	4	541
B_Q01	2.156	3.097	4 and above	3	710
B_Q02	1.266	2.200	3.3	3/4	620
B_Q03	-0.951	-0.033	3.1	3	397
B_Q04	1.245	2.179	3.3	4	618
B_Q06	0.478	1.407	3.3	4	541
B_Q08	-0.639	0.282	3.2	3	428
B_Q09	-0.587	0.334	3.2	3	433
B_Q10	1.422	2.358	3.3	3	636
B_Q11	0.54	1.469	3.3	3	547
B_Q12	-0.04	0.885	3.2	3	488
B_Q14	-0.539	0.382	3.2	4	438
B_Q15	0.531	1.460	3.3	3	546
IDOB020	1.468	2.404	3.3	3	640
IDOB021	0.599	1.529	3.3	3	553
IDOB022	-0.664	0.256	3.2	4	426
IDOB023	0.837	1.768	3.3	5	577
IDOB040	-0.749	0.171	3.2	3	417
IDOB041	-1.464	-0.549	3.1	3	345
IDOB044	-0.566	0.355	3.2	4	436
IDOB084	-0.918	0.001	3.1	3	400
IDOB085	-0.93	-0.011	3.1	3	399
IDOB086	1.467	2.403	3.3	3	640
IDOB087	-0.393	0.529	3.2	4	453
IDOB088	-0.766	0.154	3.2	4	415
IDOB097	-1.392	-0.477	3.1	2	352
IDOB098	-0.189	0.735	3.2	4	473
IDOB121	-0.227	0.697	3.2	2	470
IDOB135	-0.464	0.458	3.2	2	446
IDOB149	-1.087	-0.170	3.1	2	383
IDOB150	-0.812	0.107	3.1	3	411

Table 9.2 (Cont.) Proficiency Levels of items

Item	2009 difficulty	2009 item difficulty after adjustment for RP	Operational Level	Design level	Scaled score
IDOB173	-0.798	0.121	3.1	3	412
IDOB174	-3.393	-2.492	2 and below	2	151
IDOB177	-0.406	0.516	3.2	4	452
IDOB178	0.641	1.571	3.3	3	557
IDOB179	0.265	1.192	3.2	3	519
IDOB180	-0.014	0.911	3.2	3	491
IDOB184	0.13	1.056	3.2	4	506
IDOB185	-1.109	-0.192	3.1	2	381
IDOB186	-0.133	0.791	3.2	3	479
IDOB192	-1.474	-0.559	3.1	3	344
IDOB193	-0.785	0.135	3.2	4	413
IDOB304	-0.843	0.076	3.1	3	387
IDOB305	0.123	1.049	3.2	3	489
IDOB308	-1.693	-0.780	3.1	3	322
IDOB309	-0.746	0.174	3.2	4	417
IDOB310	-0.665	0.255	3.2	3	426
IDOB311	1.197	2.131	3.3	3	613
IDOB313	0.273	1.200	3.2	3	520
IDOB315	-0.497	0.425	3.2	3	442
IDOB316	-0.399	0.523	3.2	3	434
IDOB317	3.798	4.751	4 and above	4	877
IDOB318	-0.111	0.813	3.2	4	464
IDOB319	0.425	1.353	3.2	3	521
IDOB330	0.52	1.449	3.3	4	531
IDOB331	-1.136	-0.219	3.1	3	356
IDOB332	-0.822	0.097	3.1	3	389
IDOB333	0.588	1.518	3.3	4	538
IDOB342	1.95	2.889	4 and above	3/4	682
IDOB344	3.195	4.144	4 and above	4	813
IDOB345	-0.201	0.723	3.2	3	455
IDOB353	-0.113	0.811	3.2	4	464
IDOB355	1.868	2.807	4 and above	3	673
IDOB360	-1.411	-0.496	3.1	3	350
IDOB362	-0.356	0.567	3.2	3	457
IDOB363	0.348	1.276	3.2	3	528
IDOB364	0.24	1.167	3.2	3/4	501
IDOB366	1.232	2.166	3.3	4	606
IDOB367	-2.031	-1.120	2 and below	3	262
IDOB368	0.888	1.820	3.3	4	582
IDOB369	0.656	1.586	3.3	4	559
IDOB370	-2.626	-1.720	2 and below	3	228
IDOB371	1.252	2.186	3.3	4	619
IDOB373	1.992	2.932	4 and above	4	686
IDOB374	0.602	1.532	3.3	4	539
IDOB375	0.039	0.965	3.2	4	480

Table 9.2 (Cont.) Proficiency Levels of items

Item	2009 difficulty	2009 item difficulty after adjustment for RP	Operational Level	Design level	Scaled score
IDOB376	2.116	3.057	4 and above	4	699
IDOB381	1.107	2.040	3.3	3	593
IDOB382	1.71	2.648	4 and above	4	656
IDOB390	-1.897	-0.986	3.1	3	276
IDOB395	-3.125	-2.222	2 and below	3	146
IDOB398	-1.459	-0.544	3.1	2	322
IDOB399	1.287	2.222	3.3	4	612
IDOB401	1.24	2.174	3.3	3	607
IDOB403	0.786	1.717	3.3	3	572
IDOB404	-1.443	-0.528	3.1	2	347
IDOB405	1.127	2.060	3.3	3	595
IDOB406	1.64	2.577	3.3	4	658
IDOB417	0.982	1.914	3.3	3	591
IDOB418	-0.459	0.463	3.2	4	446
IDOB419	-0.312	0.611	3.2	4	461
IDOB420	-1.634	-0.721	3.1	4	328
IDOB422	0.621	1.551	3.3	3	541
IDOB423	-0.768	0.152	3.2	2	395
IDOB425	0.174	1.101	3.2	2	494
IDOB426	-0.832	0.087	3.1	3	388
IDOB428	-2.478	-1.571	2 and below	2	214
IDOB429	0.599	1.529	3.3	4	539
IDOB431	2.635	3.579	4 and above	4	754
IDOB433	-0.529	0.392	3.2	2	420

References

Ball, S. et al. (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. and Gonzalez, E. (1995) NAEP primer. Chestnut Hill, MA, Boston College: Boston.

Biggs, J. and 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 (2004) TIMSS 2003 Technical Report. Chestnut Hill, MA: 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. and Sheehan, K.M. (1992) Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, pp. 133–161.

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

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

OECD (2005) PISA 2003 Technical Report. Paris: OECD.

Rust, K.F. and Rao, J.N.K. (1996) Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, Vol. 5, Hodder Arnold, London, pp. 283–310.

Wu, M.L. (2005) The role of plausible values in large-scale surveys. Postlethwaite (ed.). Special Issue of *Studies in Educational Evaluation* (SEE) in memory of R.M. Wolf. 31 (2005) pp. 114–128.

Appendix A
National Year 6 Primary Science Assessment Domain

A.1 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 the 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 as it contributes to the economic and social wellbeing of the nation and improved decision making at public and personal levels (Laugksch 2000).

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)*[11] *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 National Assessment Program – Science Literacy in accord with the Ball, Rae and Tognolini (2000) report recommendation.

---

[11] 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 the National Assessment Program – Science Literacy, as the sample testing methodology to be used allowed for assessments of students' ability to conduct investigations.

A.2 Scientific literacy: Progress Map

A scientific literacy progress map was developed based on the construct of scientific literacy and on an analysis of state and territory curriculum and assessment frameworks. The Progress Map describes the development of science 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 (OECD–PISA 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 have been 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 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 communications technologies.

Strand B: Interpreting evidence and drawing conclusions from their 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 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, and then 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); and
- progression from descriptions of 'what' happened in terms of the objects and events, to explanations of 'how' it happened in terms of processes, to explanations of 'why' it happened in terms of science concepts.

The process strands (strands A and B) are most closely representative of the learning outcomes described in the Statements of Learning.

The conceptual strand (strand C) has been abstracted across conceptual strands and makes no reference to particular concepts or contexts. As the progression in the conceptual domain 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; and
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; and
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.

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

Table A.1 Scientific Literacy Progress Map – July 2004 version from DEST Science Education Assessment Resource (SEAR) project

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 their 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

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 the National Assessment Program – Science Literacy

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 map (see Table A.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 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 National Assessment Program – Science Literacy test instruments were constructed within the constraints of test length, it will not be feasible to include all the listed concepts in instruments constructed for a specific testing cycle.

Table A.2 Major scientific concepts in the 2009 National Assessment Program – Science Literacy

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, reproductions 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.
Reactions and change: Patterns of interaction of materials enable us to understand and control those interactions.	Materials can change their state and properties. Solids, liquids and gases.

Appendix B
Sample School Reports

MCEECDYA
Ministerial Council for Education,
Early Childhood Development and Youth Affairs

THE UNIVERSITY OF NEW SOUTH WALES
EDUCATIONAL ASSESSMENT AUSTRALIA

4 December 2009

PrincipalName
Sample School
1 Sample Lane
Sampleville SAMPLE 9998

Dear PrincipalName

Re: 2009 National Assessment Program – Science Literacy

On behalf of Educational Assessment Australia I wish to thank you, your staff and the Year 6 students for participating in the Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA) 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 are the reports for the participating Year 6 students at your school. There are two reports for each student: one for the pencil and paper (objective) test and one for the practical task.

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

We have included an information sheet to help interpret these reports. Please provide a copy of this information sheet to anyone requesting these results.

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

Yours sincerely

P. Hutton

Penny Hutton
Project Director

MCEECDYA
Ministerial Council for Education,
Early Childhood Development and Youth Affairs

THE UNIVERSITY OF NEW SOUTH WALES

EDUCATIONAL ASSESSMENT AUSTRALIA

2009 National Assessment Program – Science Literacy

Interpreting the Student Reports

Each Year 6 student completed one of the seven different pencil and paper (objective) 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 item 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 completing the same booklet.

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

1. the relevant major science concept area 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 examples – objective test booklets only
4. the maximum possible score for each item and the percentage of students in the school (across multiple booklets) who achieved that score
5. the percentage of students in the national sample population who achieved the maximum score on each item (the sample population contains approximately 5% of the total Year 6 national population)
6. the name of each student who completed the test for the corresponding test booklet, his/her achievement on each item and overall score on the test.

These reports can be used to:

1. compare your students' achievement on each item against the sample population (by comparing the two columns showing the % of students attaining the maximum score)
2. 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
3. identify areas in the curriculum that may need to be covered in more detail by examining the performance of students in each major science concept area.

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

90% of students in the sample population achieved the maximum score for this item.

This student achieved the maximum score (2) for this item.

The following students completed Booklet 3.

2009 National Assessment Program - Science Literacy
School Name
Year 6 Objective Booklet 3

Q No.	Major Science Concept Area*	Unit Title: Science Context	Item max score	% maximum score (your school)	% maximum score (sample population)	Student Name	Student Name	Student Name	Student Name
1	M.1	Curtains: materials have different properties and uses	2	95	100	1	2	1	0
2	M.2	Curtains: properties related to substructure (fibres)	1	90	95	0	1	1	1
3	M.2	Curtains: properties related to substructure (fibres)	1	85	90	1	1	1	0
4	LT.3	Cave diggers: adaptations for feeding (animals)	1	80	85	0	-	1	1
5	LT.3	Cave diggers: adaptations for feeding (animals)	1	75	80	1	1	1	0
		Maximum Score Possible	6	Total Score		3	5	5	2

75% of students at your school achieved the maximum score for this item.

This student did not attempt this item.

This student attempted this item and achieved a score of 0.

2009 National Assessment Program - Science Literacy

Sample School

Year 6 Objective Booklet 4

Q No.	Major Science Concept Area*	Unit Title: Science Context	Item max score	% maximum score (your school)	% maximum score (sample population)	66524 STUDENT	66531 STUDENT
1	LT.1	Collecting ants: dependence on the environment	2	38	26	2	0
2	EF.2	Collecting ants: forces as pushes and pulls	1	38	45	1	0
3	LT.3	Collecting ants: adaptation to physical environment	1	62	50	0	0
4	LT.1	Tomato plants: dependence on the environment	1	75	50	0	0
5	LT.1	Tomato plants: dependence on the environment	1	62	54	1	0
6	LT.1	Tomato plants: dependence on the environment	1	38	46	0	0
7	LT.1	Tomato plants: dependence on the environment	1	50	51	1	1
8	EF.2	Lifting weights: forces as pushes and pulls	1	75	51	1	1
9	EF.2	Lifting weights: forces as pushes and pulls	1	62	45	1	1
10	M.3	Separating mixtures: patterns of interaction of materials enable us to understand and control those interactions	1	62	43	1	-
11	M.3	Separating mixtures: patterns of interaction of materials enable us to understand and control those interactions	1	25	44	0	0
12	M.3	Separating mixtures: patterns of interaction of materials enable us to understand and control those interactions	1	50	50	1	0
13	M.3	Separating mixtures: patterns of interaction of materials enable us to understand and control those interactions	1	25	50	0	0
14	LT.1	Native and introduced animals: interactions between organisms and interdependence	1	62	46	1	0
15	M.1	Swimming costumes: materials have different properties and uses	1	29	49	0	0
16	M.1	Swimming costumes: materials have different properties and uses	1	86	50	1	1
17	M.1	Swimming costumes: materials have different properties and uses	1	57	49	0	0
18	M.1	Swimming costumes: materials have different properties and uses	1	29	49	1	0
19	EF.2	Energy efficient light bulbs: transferring energy	1	86	47	1	0
20	ES.1	Water resources: people use resources from the Earth; need to use them wisely	1	43	50	1	0
21	ES.1	Water resources: people use resources from the Earth; need to use them wisely	1	43	50	1	0
22	EF.2	Burning foods: transferring energy	1	86	43	1	0
23	EF.2	Burning foods: transferring energy	1	57	50	1	0
24	EF.2	Burning foods: transferring energy	1	57	51	1	0
25	M.3	Burning foods: reactions and change	1	57	51	0	0
26	ES.2	Greenhouse gas emissions: the changing Earth	1	57	40	0	0
27	ES.2	Greenhouse gas emissions: the changing Earth	1	43	39	0	0
28	M.3	Making jelly: materials can change their state and properties	1	14	51	0	1
29	M.3	Making jelly: materials can change their state and properties	1	57	50	1	0
30	M.3	Making jelly: materials can change their state and properties	1	86	50	1	-
31	M.3	Making jelly: materials can change their state and properties	1	57	48	0	0
32	M.3	Cola fountain: reactions and change	1	43	51	0	1
33	M.3	Cola fountain: reactions and change	1	71	51	0	1
34	M.3	Cola fountain: reactions and change	1	71	50	0	0
35	EF.2	Heating and cooling: transferring energy	2	29	33	2	1
36	EF.2	Heating and cooling: transferring energy	1	57	50	0	1
37	EF.2	Heating and cooling: transferring energy	1	71	43	1	0
38	LT.1	Pond life: dependence on the environment	1	43	41	0	0
39	LT.1	Pond life: interactions between organisms and interdependence	1	43	40	0	0
40	LT.1	Pond life: interactions between organisms and interdependence	1	29	39	0	0
		Maximum Score Possible	42	Total Score		23	9

* Refer to the key on the last page of this report document

2009 National Assessment Program - Science Literacy

Sample School

Year 6 Practical Task: Which beak works best?

Science Context: Adaptation to physical environment

Major Science Concept Area: LT.3*

Q No.	Item Descriptor	Item max 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
1	Identifies the number of beads in a simple table of results	1	47	50	1	1	1	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0	1
2	Identifies similarities between a model beak and the beak type it represents	1	58	49	1	0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	1	1	1
3	Identifies differences between a model beak and the beak type it represents	1	63	51	1	1	1	0	0	1	1	1	0	0	0	0	1	0	1	1	1	1	1
4	Concludes whether the best type of beak to collect floating weed is a sieve based on collected data	1	84	50	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0
5	In graphs provides labels for axes	1	63	51	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	0	1	1
6	Constructs scale for vertical axis with an appropriate range of values and with intervals of equal measure	1	37	51	0	0	0	1	1	0	1	0	0	0	1	0	1	0	0	0	0	1	1
7	Plots data accurately on column (bar) graph	1	42	50	0	1	1	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	1
8	Draws conclusion about the relationship between shape of bird beak and type of food a bird eats	2	21	29	2	0	0	0	1	0	0	0	0	0	0	2	1	2	0	1	1	2	1
9	Explains why using a clock would be a better timing method than counting	1	37	51	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	1	1	1	0
10	Identifies reason for using repeated trials	1	32	32	0	-	-	1	0	0	-	1	0	0	0	1	1	1	0	-	-	1	-
	Maximum Score Possible	11	Total Score		6	5	5	4	6	4	6	5	3	3	4	6	8	6	5	5	5	8	7

A science literacy progress map has been developed based on the construct of science literacy and on an analysis of State and Territory curriculum and assessment frameworks. A table of the major science concept areas (listed below) found most widely in the various State and Territory documents has been developed to accompany the Science Literacy Progress Map. These major concept areas are broad statements of scientific understandings that Year 6 students would be expected to demonstrate. For further details please visit http://www.mceecdya.edu.au/mceecdya/nap_science,12181.html

*KEY: Major Science Concept Area

Concept Area:
ES = Earth and Space

Major Science Concept Area

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 Science Concept Area

EF.1 = Energy and us: Energy is vital to our existence and o ur 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 Science Concept Area

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 Science Concept Area

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 C
Characteristics of the 2009 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 2009 sample of schools were nearly identical to those used in 2006, with the exception of the indicator of enrolment for WA. Specifically, the sample drawn in 2009 for WA used the 2008 enrolments for Year 5 as opposed to the 2008 Year 6 enrolments. The 2008 Year 6 cohort in WA commenced school when a change in school starting age was introduced. The change in school starting age resulted in the 2008 Year 6 cohort having a significantly reduced cohort size compared to subsequent cohorts. As such, the 2008 Year 5 enrolments were deemed a more suitable estimate of enrolment size for the 2009 Year 6 students. Table C.1 shows the number of sampled schools and students after drawing the sample. For example, it can be seen that approximately 10 per cent of the students in the sample were from Tasmania.

Table C.1 Number of sampled schools and estimated number of students in each jurisdiction

State/ Territory	Number of sampled schools[12]	Number of sampled students[13]	Percentage of total population of students sampled
ACT	56	1324	9
NSW	92	2094	15
NT	50	955	7
QLD	92	2098	15
SA	95	2108	15
TAS	63	1390	10
VIC	93	2128	15
WA	94	2083	15
Total	635	14 180	100

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

[12] The number of sampled schools in Table C.1 differs slightly from those presented in Table 3.4 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 635 schools, 17 schools did not participate in the testing (and could not be replaced).

[13] These are the numbers of students enrolled according to the sampling frame. These differ slightly from the numbers shown in Table 3.7, where the number of students refers to those enrolled at *the time of testing*.

Table C.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 1 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 C.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	1284	28.5%	15	361	27.3%	1.2%
	Govt	65	2660	59.1%	34	808	61.0%	-1.9%
	Other	13	557	12.4%	7	155	11.7%	0.7%
	Total	103	4501	100.0%	56	1324	100.0%	0.0%
NSW	Cath	441	17 530	20.1%	18	414	19.8%	0.3%
	Govt	1635	59 805	68.7%	63	1447	69.1%	-0.4%
	Other	262	9777	11.2%	11	233	11.1%	0.1%
	Total	2338	87 112	100.0%	92	2094	100.0%	0.0%
NT	Cath	12	375	12.5%	5	125	13.1%	-0.6%
	Govt	125	2378	79.1%	40	735	77.0%	2.1%
	Other	13	252	8.4%	5	95	9.9%	-1.5%
	Total	150	3005	100.0%	50	955	100.0%	0.0%
QLD	Cath	218	9185	16.1%	15	348	16.6%	-0.5%
	Govt	1005	40 682	71.5%	66	1493	71.2%	0.3%
	Other	156	7012	12.3%	11	257	12.2%	0.1%
	Total	1379	56 879	100.0%	92	2098	100.0%	0.0%
SA	Cath	82	3557	18.5%	17	404	19.2%	-0.7%
	Govt	453	12 606	65.5%	63	1359	64.5%	1.0%
	Other	80	3082	16.0%	15	345	16.4%	-0.4%
	Total	615	19 245	100.0%	95	2108	100.0%	0.0%
TAS	Cath	32	1124	16.6%	10	238	17.1%	-0.5%
	Govt	168	4924	72.9%	46	1006	72.4%	0.5%
	Other	27	708	10.5%	7	146	10.5%	0.0%
	Total	227	6756	100.0%	63	1390	100.0%	0.0%
VIC	Cath	393	14 058	21.4%	20	464	21.8%	-0.4%
	Govt	1245	44 075	67.2%	62	1413	66.4%	0.8%
	Other	175	7440	11.3%	11	251	11.8%	-0.5%
	Total	1813	65 573	100.0%	93	2128	100.0%	0.0%
WA	Cath	130	4850	17.3%	16	367	17.6%	-0.3%
	Govt	633	19 677	70.2%	66	1468	70.5%	-0.3%
	Other	120	3490	12.5%	12	248	11.9%	0.6%
	Total	883	28 017	100.0%	94	2083	100.0%	0.0%
Total	Cath	1333	51 963	19.2%	116	2721	19.2%	0.0%
	Govt	5329	186 807	68.9%	440	9729	68.6%	0.3%
	Other	846	32 318	11.9%	79	1730	12.2%	-0.3%
	Total	7508	271 088	100.0%	635	14 180	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 C.3 shows the number of schools according to school size for the population and the selected sample. Table C.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, approximately 4 per cent of the population attend a *very small* school which is very similar to the 3 per cent of students from *very small* schools included in the selected sample.

Table C.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	4270	235 741	87.0%	493	12 325	86.9%
Moderately small	1287	24 009	8.9%	81	1498	10.6%
Very small	1951	11 338	4.2%	61	357	2.5%
Total	7508	271 088	100.0%	635	14 180	100.0%

Appendix D
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 OECD (2005) treatment of small schools (pp. 53–56). 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 – 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, Geographic location and then MOS. The *sort order* was alternated so that 'like schools' were always nearby.

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

Table D.1 The sort ordering procedures employed for small schools

Sector	Geographic location	ENR sort order
1	1	A
1	2	D
1	3	A
2	3	D
2	2	A
2	1	D
3	1	A
3	2	D
3	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 D.2 shows the starting values used to draw the sample for each explicit stratum.

Table D.2 Stratum variables for sample selection

Stratum	MOS	Number of schools	Interval	Random start
ACT_Large_C	1163	19	78	46
ACT_Large_G	2427	48	81	74
ACT_Large_O	492	7	82	20
ACT_ModSmall	352	18	59	8
ACT_VerySmall	67	11	34	17
NSW_Large_C	15 520	294	1035	288
NSW_Large_G	52 920	948	1018	860
NSW_Large_O	8460	148	1058	178
NSW_ModSmall	6999	374	778	28
NSW_VerySmall	3213	574	459	336
NT_Large_C	313	8	78	1
NT_Large_G	1745	42	76	35
NT_Large_O	154	4	77	36
NT_ModSmall	450	25	56	55
NT_VerySmall	343	71	29	19
QLD_Large_C	8242	147	687	43
QLD_Large_G	36 397	521	662	448
QLD_Large_O	6188	91	688	47
QLD_ModSmall	3796	207	475	152
QLD_VerySmall	2256	413	282	15
SA_Large_C	3157	55	226	151
SA_Large_G	9773	209	222	54
SA_Large_O	2709	52	226	31
SA_ModSmall	2479	132	165	6
SA_VerySmall	1127	167	125	108
TAS_Large_C	945	22	118	68
TAS_Large_G	3938	89	119	90
TAS_Large_O	535	10	107	5
TAS_ModSmall	996	53	91	37
TAS_VerySmall	342	53	57	24

Table D.2 (Cont.) Stratum variables for sample selection

Stratum	MOS	Number of schools	Interval	Random start
VIC_Large_C	11 975	250	748	208
VIC_Large_G	38 119	711	762	584
VIC_Large_O	6679	109	742	600
VIC_ModSmall	6255	336	569	343
VIC_VerySmall	2545	407	364	228
WA_Large_C	4316	91	332	40
WA_Large_G	16 638	342	326	197
WA_Large_O	2936	53	326	36
WA_ModSmall	2682	142	244	41
WA_VerySmall	1445	255	145	100

Appendix E
Programming Notes on Sampling

E.1 SPSS syntax for sample selection

```
*===========================================
*===========================================
       NAP-SL 2009 SAMPLE PROCEDURE
*===========================================
*===========================================.
*relative file paths used.
*SPSS version 14.



*===========================================
       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)).

GET FILE='SampleFrame.sav'.

*=======SELECT STRATUM=======.
select if (RTRIM(Stratum)=!strata).
exe.
SORT CASES BY StateId (A) SectorId (A) GeoId (A) gr06 (A) .

*=======IDENTIFY SUBGROUPS=======.
if ($casenum = 1) stratumsort = 1.
do if (sectorid = lag(sectorid) and geoid = lag(geoid)).
       compute stratumsort = lag(stratumsort).
else.
       compute stratumsort = lag(stratumsort) + 1.
```

```
end if.
exe.

*=======STRATIFY SUBGROUPS=======.
!IF (!enrsize = 999)!THEN.
        *=======LARGE SCHOOL SORT=======.
        title 'Large school sort'.
        do if (MOD(stratumsort,2) > 0).
                compute sort2 = stratumsort * 1000 + gr06.
        else.
                compute sort2 = stratumsort * 1000 - gr06.
        end if.
        exe.
!ELSE.
        *=======SMALL SCHOOL SORT=======.
        title 'Small school sort'.
        do if (MOD(sectorid,2) > 0).
                compute sort1 = sectorid * 100 + stratumsort.
        else.
                compute sort1 = sectorid * 100 - stratumsort.
        end if.
        RANK
          VARIABLES=sort1  (A) /RANK /PRINT=YES
          /TIES=CONDENSE .
        do if (MOD(Rsort1,2) > 0).
                compute sort2 = Rsort1 * 1000 + gr06.
        else.
                compute sort2 = Rsort1 * 1000 - gr06.
        end if.
        exe.
        compute tmpgr06 = gr06.
        compute gr06 = !enrsize.
!IFEND.
SORT CASES BY Sort2 (A).

*=======SET VERY LARGE SCHOOLS EQUAL TO THE SAMPLING INTERVAL=======.
if (gr06>!const) gr06 = !const.
exe.

*=======RANDOMLY SELECT SCHOOLS WITH 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=inst_name stratum gr06 ticket1 ticket2 wintickt select / format = numbered.
title.
SAVE OUTFILE=!QUOTE(!CONCAT('Sample_',!UNQUOTE(!strata) , '.sav')).

!ENDDEFINE.

*============================================
        DRAW SAMPLE
*============================================.
!SAMPLE strata='ACT_Large_C' const=78 randm=46 enrsize=999.
!SAMPLE strata='ACT_Large_G' const=81 randm=74 enrsize=999.
!SAMPLE strata='ACT_Large_O' const=82 randm=20 enrsize=999.
!SAMPLE strata='ACT_ModSmall' const=59 randm=8 enrsize=20.
!SAMPLE strata='ACT_VerySmall' const=34 randm=17 enrsize=6.
!SAMPLE strata='NSW_Large_C' const=1035 randm=288 enrsize=999.
!SAMPLE strata='NSW_Large_G' const=1018 randm=860 enrsize=999.
!SAMPLE strata='NSW_Large_O' const=1058 randm=178 enrsize=999.
!SAMPLE strata='NSW_ModSmall' const=778 randm=28 enrsize=19.
!SAMPLE strata='NSW_VerySmall' const=459 randm=336 enrsize=6.
!SAMPLE strata='NT_Large_C' const=78 randm=1 enrsize=999.
!SAMPLE strata='NT_Large_G' const=76 randm=35 enrsize=999.
!SAMPLE strata='NT_Large_O' const=77 randm=36 enrsize=999.
!SAMPLE strata='NT_ModSmall' const=56 randm=55 enrsize=18.
!SAMPLE strata='NT_VerySmall' const=29 randm=19 enrsize=5.
!SAMPLE strata='QLD_Large_C' const=687 randm=43 enrsize=999.
!SAMPLE strata='QLD_Large_G' const=662 randm=448 enrsize=999.
!SAMPLE strata='QLD_Large_O' const=688 randm=47 enrsize=999.
!SAMPLE strata='QLD_ModSmall' const=475 randm=152 enrsize=18.
!SAMPLE strata='QLD_VerySmall' const=282 randm=15 enrsize=5.
!SAMPLE strata='SA_Large_C' const=226 randm=151 enrsize=999.
!SAMPLE strata='SA_Large_G' const=222 randm=54 enrsize=999.
!SAMPLE strata='SA_Large_O' const=226 randm=31 enrsize=999.
!SAMPLE strata='SA_ModSmall' const=165 randm=6 enrsize=19.
!SAMPLE strata='SA_VerySmall' const=125 randm=108 enrsize=7.
!SAMPLE strata='TAS_Large_C' const=118 randm=68 enrsize=999.
!SAMPLE strata='TAS_Large_G' const=119 randm=90 enrsize=999.
!SAMPLE strata='TAS_Large_O' const=107 randm=5 enrsize=999.
!SAMPLE strata='TAS_ModSmall' const=91 randm=37 enrsize=19.
!SAMPLE strata='TAS_VerySmall' const=57 randm=24 enrsize=6.
!SAMPLE strata='VIC_Large_C' const=748 randm=208 enrsize=999.
!SAMPLE strata='VIC_Large_G' const=762 randm=584 enrsize=999.
!SAMPLE strata='VIC_Large_O' const=742 randm=600 enrsize=999.
!SAMPLE strata='VIC_ModSmall' const=569 randm=343 enrsize=19.
!SAMPLE strata='VIC_VerySmall' const=364 randm=228 enrsize=6.
!SAMPLE strata='WA_Large_C' const=332 randm=40 enrsize=999.
!SAMPLE strata='WA_Large_G' const=326 randm=197 enrsize=999.
```

```
!SAMPLE strata='WA_Large_O' const=326 randm=36 enrsize=999.
!SAMPLE strata='WA_ModSmall' const=244 randm=41 enrsize=19.
!SAMPLE strata='WA_VerySmall' const=145 randm=100 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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.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_O.sav'
  /FILE='Sample_WA_ModSmall.sav'
  /FILE='Sample_WA_VerySmall.sav'.
EXECUTE.
SAVE OUTFILE='SampleSchools2009.sav'.
```

Appendix F
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 I nformation (below) and Part B - Student Participation (overleaf). Please refer to pages 9 - 11 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:	Which beak works best?

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: ______________________

SPECIAL EDUCATION NEEDS (SEN) CODES (Column 7)	NON-INCLUSION CODES (Columns 9 and 11)	INDIGENOUS CODES (Column 5)
0 = No special education needs	10 = Absent	1 = Aboriginal but not Torres Strait Islander Origin
1 = Functional disability	11 = Not included; functional disability	2 = Torres Strait Islander but not Aboriginal Origin
2 = Intellectual disability	12 = Not included; intellectual disability	3 = Both Aboriginal and Torres Strait Islander Origin
3 = Limited test language proficiency	13 = Not included; limited test language proficiency	4 = Neither Aboriginal nor Torres Strait Islander Origin
	14 = Student or parent refusal	9 = Not stated/Unknown

See full explanation on pages 9-11 of the Test Administrator's Manual

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 Didn't complete = 0 Completed = 1	(9) Non-Inclusion Code (see overleaf)	(10) Practical Task Didn't complete = 0 Completed = 1	(11) Non-Inclusion Code (see overleaf)
613505	Student 613505	2	F	4	24/10/1996					
613513	Student 613513	3	M	4	3/12/1996					
613521	Student 613521	4	F	9	1/01/1997					
613539	Student 613539	5	F	4	27/01/1997					
613547	Student 613547	6	M	4	26/02/1997					
613554	Student 613554	7	F	9	1/03/1997					
613562	Student 613562	1	F	4	7/03/1997					
613570	Student 613570	2	F	9	16/03/1997					
613588	Student 613588	3	M	4	22/03/1997					
613596	Student 613596	4	M	4	22/03/1997					
613604	Student 613604	5	M	4	11/04/1997					
613612	Student 613612	6	F	4	16/04/1997					
613620	Student 613620	7	M	4	24/04/1997					
613638	Student 613638	1	F	4	27/04/1997					
613646	Student 613646	2	F	4	31/05/1997					
613653	Student 613653	3	M	4	2/07/1997					
613661	Student 613661	4	F	9	8/07/1997					
613679	Student 613679	5	F	4	28/07/1997					
613687	Student 613687	6	M	4	4/08/1997					
613695	Student 613695	7	M	4	19/08/1997					
613703	Student 613703	1	F	4	24/08/1997					
613711	Student 613711	2	F	4	3/09/1997					
613729	Student 613729	3	F	9	13/10/1997					
613737	Student 613737	4	M	4	21/10/1997					

Appendix G
Variables in File

Table G.1 File Name: *NAPSL2009_PV_2010-04-13.sav*

Variable names	Description
Barcode	Student Barcode
BookNumber	Book Number (Objective Task)
B_Q01 to A_Q14	Practical Task Question
ID0B184 to IDOB426	Objective Task Question
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
LBOTE	LBOTE recode for stratification
LBOTE1	LBOTE regression variable 1
LBOTE2	LBOTE regression variable 2
Prac	Practical Task identifier
ObjectiveTest	Did not sit code for Objective Task
PracticalTask	Did not sit code for Practical Task
StudentID	Student ID
SchoolID	School ID
State	State
Stratum	Stratum Identifier
CQC_CaseID	Conquest ID
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

Table G.1 (Cont.) *NAPSL2009_PV_2010-04-13.sav*

Variable names	Description
tmpPair	Temporary variable for sampling zone
DblWgtPairNum	Sampling Zone Pair weight flag
RW0 - RW318	Replicate weight 0 to 318
RUMMWLE	Weighted Likelihood estimate from RUMM2020
SchWleRUMM	School Mean WLE from RUMM2020
Free_PV1 - PV10	2009 Plausible Value calibrated free (1 to 10)
EAP	EAP value
EAP_SE	EAP SE Value
PV1 - PV10	2009 PV1 -10 (on 2006 scale)
Level1 - Level 10	2009 Level for PV1 - 10
YearLevel	School Year Level
Class	School Class
Gender	Student Gender
DOB	Student Date of Birth
IsParticipating	Student particpation flag (from SPF)
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
LboteP1ID	Language Background of Parent 1
LboteP2ID	Language Background of Parent 2
SchoolStudentID	Student ID allocated by school
SENCode	Special Education Needs code
ObjNonInclusionCode	Objective Task non-inclusion code
PracNonInclusionCode	Practical Task non-inclusion code
Sector	School Sector
CaseID	Record Number

Appendix H
ConQuest Control File for Producing Plausible Values

Table H.1 File Name: *NAPSL2009_Produce_2009_PV.cqc*

```
reset;

Data NAPSL2009_PV_Check_2010-04-06.dat;

format id 1-6 bookID 8 response 9-121 PWeight 122-129 SchWLE 130-137

          State1 138
          State2 139
          State3 140
          State4 141
          State5 142
          State6 143
          State7 144
          Gender1 145
          Gender2 146
          ATSI1 148
          ATSI2 149
          Sector1 150
          Sector2 151
          Geolocation1 152
          Geolocation2 153
          LBOTE1 155
          LBOTE2 156;

Labels << NAPSL2009.lab;

set constraint=none;
Set n_plausible=10;
caseweight PWeight;

key
11111141113111111111112121113341141114143411121111131122411121211111214132134121112131111143111112141113111211124 !
1;

key
2XX2XXX2XXXXXXXXXXX2XX2XXXXXXXXXX2XXXXXXX2XXXX2XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX2XXXXXXXXXXXXXXXXXX !
2;

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

import anchor_param << NAPSL2009_Calib_Anchor_2010-03-23.anc;

model bookID + item + item*step;

regression
 SchWLE
 State1
 State2
 State3
```

```
  State4
  State5
  State6
  State7
  Gender1
  Gender2
  ATSI1
  ATSI2
  Sector1
  Sector2
  Geolocation1
  Geolocation2
  LBOTE1
  LBOTE2;

set warnings = no;

Estimate! iterations = 2000, fit=no, nodes=30;

show !estimate=latent >> NAPSL2009_PV_Check_2010-04-06.shw;

itanal >> NAPSL2009_PV_Check_2010-04-06.itn;

show cases !estimate=latent >> NAPSL2009_PV_Check_2010-04-06.pls;

show cases !estimate=WLE >> NAPSL2009_PV_Check_2010-04-06.wle;
```