# *MCEETYA NATIONAL ASSESSMENT PROGRAM, SCIENCE, YEAR 6 2003: TECHNICAL REPORT*

## *14 June 2005*

**Ministerial Council on Education, Employment, Training and Youth Affairs**

**Australian Council for Educational Research**
**Sydney**

# TABLE OF CONTENTS

# List of Tables and Figures

# CHAPTER 1: OVERVIEW OF THE NATIONAL ASSESSMENT

## 1.1 Introduction

In 1999, the State, Territory and Commonwealth Ministers of Education agreed to the *National Goals* for *Schooling in the Twenty-first Century* (the Adelaide Declaration). The National Goals provide the framework for reporting on student achievement through the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) publication, the *National Report on Schooling in Australia*.

Also in 1999, Ministers established the National Education Performance Monitoring Taskforce (NEPMT) to develop key performance measures to monitor and report on progress toward the achievement of the Goals on a nationally comparable basis. They identified six priority areas for the development of performance measures: literacy, numeracy, science, information technology, vocational education and training in schools and participation and attainment.

As a first step, in early 2000 NEPMT commissioned a project to develop options for the assessment and reporting of the achievements of primary school students in science. The outcome of this process was a report to NEPMT entitled *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* (Ball et al., 2000).

The Ball report recommended that student achievement in science literacy (namely, science concepts and science process skills) rather than the acquisition of factual information should be assessed and reported at the primary level. In particular, the report suggested adoption of the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) definition of science literacy (1999) for the purposes of primary science monitoring.

In July 2001, MCEETYA agreed to the development of assessment instruments and key performance measures for reporting on student skills, knowledge and understandings in science at primary school. MCEETYA requested the Performance Measurement and Reporting Taskforce (PMRT), a newly established taskforce that had replaced NEPMT, to undertake the national assessment program.

The Ball report highly recommended that the assessment be conducted at the end of primary school because:

*… in Science with the purpose of monitoring, delay until the end of Primary schooling has the advantages of being able to assess a more mature learner who has had greater opportunity to develop scientific skills and processes and develop a better understanding of basic scientific principles*

(Ball et al., 2000, p.44)

Implementation of the National Year 6 Science Assessment required a large number of separate but related steps, including the development of an assessment domain and accompanying items and instruments, trialing of these items and assessments, the construction of key performance measures for measuring and reporting on the achievement of students in science, the administration of the assessments to a sample of year 6 students and marking, analysing and reporting the results.

## 1.2 The National Year 6 Science Assessment

The assessment measured scientific literacy. Scientific literacy is a construct that:

*encompasses the use of broad conceptual understandings of science for making sense of the world, understanding natural phenomena, interpreting media reports about scientific issues. It also encompasses competencies related to asking investigable questions, conducting investigations, collecting and interpreting data and making decisions.*

(Hackling, 2002).

The science items and instruments therefore assess outcomes that contribute to scientific literacy, such as conceptual understandings (rather than facts) and investigation competencies in realistic situations. As such, it relates to the ability to think scientifically in a world in which science and technology are increasingly shaping children's lives.

A scientific literacy assessment domain was developed for the assessment in consultation with curriculum experts from each jurisdiction and representatives of the Catholic and independent schools sectors. This domain includes the definition of scientific literacy and outlines the development of scientific literacy across three main areas. A copy of the scientific literacy assessment domain is provided at Appendix 1, MCEETYA Scientific Literacy Framework.

Three strands of scientific literacy were assessed:

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

*Strand B:* interpreting evidence and drawing conclusions, critiquing the trustworthiness of evidence and claims made by others and communicating findings.

*Strand C:* using science understandings for describing and explaining natural phenomena, interpreting reports and making decisions.

There was a conscious effort to develop assessment items that related to everyday contexts rather than to laboratory situations.

The science assessment items draw on four concept areas: Life and Living, Earth and Beyond, Natural and Processed Materials and Energy and Change. These evolved from a review of the National Statements and Profiles and were generally common across Australian curricula. It is interesting to note that the same concept areas are also fairly common internationally.

The strands of scientific literacy and the concepts to be assessed were informed by a thorough analysis and mapping of the curriculum documents of all States and Territories. The intention was to ensure that all year 6 students would be familiar with the types of materials and experiences involved in the assessment. This process was conducted in order to avoid any systematic bias in the assessment instruments being developed.

## 1.3 Participants in the assessment

The target population for the study was year 6 students enrolled in educational institutions across Australia. A grade-based population was chosen. As shown in Table 1.1, there are structural differences between the States and Territories in the ages of entry to full-time formal schooling. This is consistent with the reporting of literacy and numeracy performance in the *National Report on Schooling in Australia*. There was no adjustment for any age factors in the analysis and reporting of the assessment results.

– **Table 1.1: Years of formal schooling, by State and Territory**

| State/Territory | Age at commencement of school | Starting class | Years of primary schooling to year 6 |
|---|---|---|---|
| NSW | 5 years by 31 July | Kindergarten | 7 |
| VIC | 5 years by April 30 | Preparatory class | 7 |
| QLD | 6 years by 31 December | Year 1 | 6 |
| SA | 6 years | Preparatory Year | 6 |
| WA | 5 years by April 30 | Year 1 | 6 |
| TAS | 5 years to commence school | Grade 1 | 6 |
| NT | 5 years to commence school | Year 1 | 6 |
| ACT | 5 years by April 30 | Kindergarten | 7 |

Approximately 6 per cent of the national year 6 student population were sampled randomly and assessed. All States and Territories, and a majority of government, Catholic and independent schools within them participated. Table 1.2 shows the number of schools and students, by State and Territory, in the final sample from which performance comparisons were reported.

– **Table 1.2: Number of schools and students in the final sample, by State and Territory**

| State/Territory | Number of schools in target sample | Number and % of schools in final sample | Number of students in final sample |
|---|---|---|---|
| NSW | 122 | 103 (84%) | 2,466 |
| VIC | 122 | 100 (82%) | 2,130 |
| QLD | 122 | 110 (90%) | 2,607 |
| SA | 130 | 115 (88%) | 2,032 |
| WA | 126 | 103 (81%) | 2,347 |
| TAS | 64 | 60 (94%) | 1,240 |
| NT | 32 | 23 (72%) | 496 |
| ACT | 44 | 36 (82%) | 854 |
| **All** | **762** | **650 (85%)** | **14,172** |

In a number of cases, schools nominated students to participate in the assessment who were outside the target population. These students were not included in the results. There were also instances of schools that wished to participate in the study as volunteer schools. The students from these schools were not included in the results, even though their assessments were marked and feedback was provided to the schools on the performance of their students. Some schools with multi-level classes requested that their years 5 and 7 students complete the assessment. Similarly, these students' tests were identified and removed from the sample. This meant that many students who attempted the tests could not validly be included in the final sample.

## 1.4 The assessment test format

The students' regular classroom teachers administered the assessment between 20 and 31 October 2003. The assessment comprised two pencil-and-paper assessments with multiple-choice and short answer items and two practical assessment tasks. The assessment papers were distributed randomly so that half the students in each class completed one of the pencil-and-paper assessments and the other half completed the other pencil-and-paper assessment. All students in the one class took the same practical task, but the tasks were assigned randomly across Australia in a way that ensured that approximately equal numbers of classes attempted the two practical assessment tasks.

The practical task required the students to work in groups of three. The teachers, using a procedure outlined in the Assessment Administrator's Manual, allocated students randomly to groups. Students conducted the experiment in these groups and responded to a set of questions designed to stimulate group discussion about the experiment. The students then answered a further set of items independently. Only the individual student responses were used in the analysis and generation of proficiency data.

Equating the two objective assessments onto the one scale was achieved by the use of ten common items (four common units or sets of items) shared between the two objective assessments. The practical items were then linked onto this scale by results obtained from common students doing one objective assessment and one practical task.

Students were allowed 60 minutes to complete the pencil-and-paper assessments and 45 minutes for the practical assessment tasks.

## 1.5 Reporting of the assessment results

The results of the assessment were reported in the *MCEETYA National Year 6 Science Assessment Report 2003*. Mean scores and distributions of scores are shown at the national level and by State and Territory. The results are also described in terms of the understandings and skills that students demonstrated in the assessment, which are mapped against the scientific literacy assessment framework.

## 1.6 Structure of the Technical Report

This report describes the technical aspects of the National Year 6 Science Assessment and summarises the main activities associated with the data collection, the data collection instruments and the analysis and reporting of the results.

Chapter 2 reviews the sample design for the assessment and describes the sampling process and the sample achieved.

Chapter 3 summarises the test development and implementation procedures and the procedures for instrument construction and compliance with the test specification.

Chapter 4 reviews the assessment administration procedures, discusses the cleaning of data collected in the assessment and the treatment of missing data and invalid students.

Chapter 5 addresses the main features of the procedures used for weighting the student data and the replication procedures used to account for the sample design.

Chapter 6 summarises the results of the analyses undertaken, concentrating on the results of the Rasch analysis and providing information about the calibration procedures for item locations and student ability estimates.

Chapter 7 describes and analyses the procedures undertaken to review the quality of the links between the various test forms and for the equating of the various test forms and their scaling for reporting.

Chapter 8 discusses the results in terms of students' proficiency on the scientific literacy scale. The scale links students' results to descriptions of their understandings and skills in the assessment domain.

Chapter 9 comments on some issues that may be relevant in the 2006 iteration of the sample study and may need further consideration in terms of the test construction, its equating to the 2003 assessment and its analysis.

# CHAPTER 2: SAMPLE DESIGN

## 2.1 The overall sampling procedure

The sample was selected using procedures similar to those of PISA and the Trends in International Mathematics and Science Study (TIMSS), (1999). The distribution of schools from the sectors within each jurisdiction was drawn proportional to the student representation within that jurisdiction. In smaller States and Territories, the selection of a sample size necessary to achieve the same degree of confidence in the results as in larger States and Territories would have meant that almost all year 6 students would have had to participate in the assessment. Consequently, the sample numbers in the Australian Capital Territory, Northern Territory and Tasmania were reduced. This had the effect of increasing the level of uncertainty around the results.

The sampling procedures helped to ensure that the data were of a high standard, so that valid comparisons of results between States and Territories could be made. Personnel drawing on their sampling expertise from the PISA project provided advice on the sample design.

## 2.2 Sample frame and design

The design implemented was a two-stage stratified cluster sample. The cluster size was assumed to be 25 students.

*The first-stage sampling units* consisted of individual schools. Schools were selected with probabilities proportional to their size (PPS), with 'size' being defined as the estimated number of eligible students enrolled.

The comprehensive list of all eligible schools is the school sampling frame.

The national sample frame has been developed by the Australian Council of Educational Research (ACER) by coordinating information from multiple sources, including the Australian Bureau of Statistics and State and Territory databases.

*The second-stage sampling units* were classes within sampled schools. The use of whole classes minimised disruption within schools. In addition, this design was most suitable for the administration of the practical task, given that it was dependent upon group work.

A disadvantage of using whole classes was the interaction effect associated with the teacher (and school). This is acknowledged by the degree of homogeneity ($\rho$) within the cluster (where $\rho$ is referred to as the intra-cluster correlation coefficient). Consequently, the sample size needed to be increased over that of a simple random sample, in order to compensate for the effect of the sampling design on the sampling variance and hence the standard error of the means and percentages.

The ACER sample frame was based on 2002 statistical data.

The sample design was influenced by two factors:

- the sampling standards specified in the tender document and contract; and

- the requirement for a level of face validity of the information generated from the study.

The sample design was derived using the principles articulated in the *TIMSS 1999 Technical Report* and was based on a calculated national random sample size of 400 students. Given a cluster size of 25 students, Exhibit 2.2 in the *TIMMS Technical Report* indicates that in order to provide 95 per cent confidence limits of +/- 0.2s for estimated means, a sample size of 2,800 cases per State or Territory is required. As such large samples would involve high proportions of schools in the smaller States and Territories, it was decided that smaller samples would be drawn for Tasmania, the Australian Capital Territory and the Northern Territory, despite the potential for non-compliance with the sampling standards.

The sample frame was partitioned into 24 strata (eight States and Territories and three sectors within each), as indicated in Table 2.1.

## 2.3 Achieved sample

The achieved sample at State and Territory level is shown in Table 2.1. It provides the global sample weights, targeted sample and achieved samples from the study. The ACER sample frame details the relative weights that were applied to each candidate to indicate his or her individual contribution to the total sample.

The initial contract specified that the achieved sample must provide a minimum sampling precision at the 95 per cent confidence limits of +/- 0.2s for estimated means in relation to the effective sample achieved.

– **Table 2.1: Target and achieved samples for the National Year 6 Science Assessment**

| State/ Territory | School sector | Year 6 enrolment | Percentage of total enrolment | Year 6 target sample | Percent age of year 6 in target sample | Number of year 6 in achieved sample |
|---|---|---|---|---|---|---|
| NSW | Government | 63,182 | 23.5 | 1,987 | 11.62 | 1,631 |
| | Catholic | 17,585 | 6.5 | 553 | 3.24 | 586 |
| | Independent | 8,457 | 3.1 | 265 | 1.55 | 249 |
| | **Totals** | **89,226** | | **2,805** | | **2,466** |
| VIC | Government | 44,424 | 16.5 | 1,913 | 11.19 | 1,612 |
| | Catholic | 14,631 | 5.4 | 630 | 3.68 | 399 |
| | Independent | 6,210 | 2.3 | 227 | 1.33 | 119 |
| | **Totals** | **65,265** | | **2,770** | | **2,130** |
| QLD | Government | 40,160 | 14.9 | 2100 | 12.28 | 1847 |
| | Catholic | 8,128 | 3.0 | 428 | 2.50 | 467 |
| | Independent | 5,136 | 1.9 | 277 | 1.62 | 293 |
| | **Totals** | **53,424** | | **2,805** | | **2,607** |
| SA | Government | 13,538 | 5.0 | 1,947 | 11.38 | 1,192 |
| | Catholic | 3,290 | 1.2 | 492 | 2.88 | 490 |
| | Independent | 2,469 | 0.9 | 368 | 2.15 | 350 |
| | **Totals** | **19,297** | | **2807** | | **2,032** |
| WA | Government | 19,859 | 7.4 | 2,042 | 11.94 | 1,607 |
| | Catholic | 4,606 | 1.7 | 478 | 2.80 | 525 |
| | Independent | 2,691 | 1.0 | 292 | 1.71 | 215 |
| | **Totals** | **27,156** | | **2,812** | | **2,347** |
| TAS | Government | 5,135 | 1.9 | 1,048 | 6.13 | 843 |
| | Catholic | 1,084 | 0.4 | 215 | 1.26 | 287 |
| | Independent | 595 | 0.2 | 115 | 0.67 | 110 |
| | **Totals** | **6,814** | | **1,378** | | **1,240** |
| NT | Government | 2,272 | 0.8 | 528 | 3.09 | 420 |
| | Catholic | 409 | 0.2 | 88 | 0.52 | 43 |
| | Independent | 172 | 0.1 | 65 | 0.38 | 33 |
| | **Totals** | **2,853** | | **681** | | **496** |
| ACT | Government | 2,995 | 1.1 | 653 | 3.82 | 483 |
| | Catholic | 1,408 | 0.5 | 300 | 1.75 | 288 |
| | Independent | 380 | 0.1 | 88 | 0.52 | 83 |
| | **Totals** | **4,783** | | **1,041** | | **854** |
| **National** | **Totals** | **268,818** | **100.0** | **17,099** | **100.00** | **14,172** |

Table 2.2 shows that, despite the reduction in the achieved sample in smaller States and Territories, the required precision was achieved in all States and Territories. This result could have implications for the determination of the sample size for the next iteration of the program. However, it should be considered in the context of the 2003 program, which took into account clarifying validity issues in public debate if smaller samples were implemented.

– **Table 2.2: Comparison of achieved sampling precision with nominated standard**

| State/ Terr | Cases | Effective sample size | Mean ability estimate | Standard error | 95% CI[(a)] | Standard deviation (SD) | 0.2 Standard deviation |
|---|---|---|---|---|---|---|---|
| NSW | 2,466 | 570 | 0.110 | 0.040 | 0.078 | 0.949 | 0.190 |
| VIC | 2,130 | 559 | -0.010 | 0.041 | 0.080 | 0.956 | 0.191 |
| QLD | 2,607 | 669 | -0.080 | 0.037 | 0.073 | 0.971 | 0.194 |
| SA | 2,032 | 652 | -0.070 | 0.040 | 0.078 | 1.000 | 0.200 |
| WA | 2,347 | 440 | -0.100 | 0.047 | 0.092 | 0.970 | 0.194 |
| TAS | 1,240 | 314 | 0.070 | 0.060 | 0.118 | 1.090 | 0.218 |
| NT | 496 | 155 | -0.210 | 0.099 | 0.194 | 1.230 | 0.246 |
| ACT | 854 | 297 | 0.300 | 0.062 | 0.122 | 1.040 | 0.208 |
| **All** | **14,172** | **2,625** | **0.000** | **0.019** | **0.037** | **1.000** | **0.200** |

(a) Comparison of 95% confidence interval with 0.2 SD achieved in all jurisdictions.

## 2.4 School exclusions and within-school exclusions

There are three classes of non-participation:

- *exclusions:* those excluded from the sample frame due to remoteness or size;
- *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.

The test design required that a practical task be completed and that a minimum of three students was required for the task. Given the normal absentee patterns of schools, it was decided that schools with year 6 populations of fewer than five students would be excluded from the target sample.

In addition, because of logistical issues and concerns about the reliability of communications with schools defined within the MCEETYA sample as 'very remote', they were excluded from the sample, as were any special-purpose schools, including hospital schools.

At the school level, all students in the nominated class were to be included in the sample, with the principal having the prerogative to exempt students with defined disabilities.

Table 2.3 shows the student exclusion/exemption categories.

– **Table 2.3: Coding of the student exclusion/exemption categories**

| 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 PSAP 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 PSAP 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 requested that student not participate OR student refusal. |

## 2.5 Small schools

To ensure that the sample was not biased against smaller schools by the Probability Proportional to Size (PPS) methodology, small schools (those with target populations of fewer than 15 students) were combined into a separate stratum. By combining two or more small schools, the target cluster size could be achieved. This process was performed within each State and Territory and the combined schools formed 'pseudo-schools'. In cases where a pseudo-school was chosen by the sampling process, schools that made up that 'pseudo-school' were included in the sample.

## 2.6 Replacement schools

To allow for the possibility that a sample school might be unable to participate in the assessment, a replacement school was nominated for each school selected by the PPS process. At least one potential replacement school was selected for each sampled school across all jurisdictions in the original sample. A substitution was made only if the principal of a selected school indicated that the school was unable to participate in the assessment.

Table 2.4 details the participation rate achieved in the sample. For the purpose of reporting participation rates, replacement schools were reported as sampled schools.

– **Table 2.4: Summary of sampling standards achieved**

| State/Territory | Target Population Schools | Target Population Students | Target Sample Schools | Target Sample Students | Achieved Schools | Achieved Students | School participation rate (%) | Reported absent | Reported refusals | Reported exempt |
|---|---|---|---|---|---|---|---|---|---|---|
| NSW | 2,050 | 85,834 | 122 | 2,805 | 103 | 2,466 | 84.4 | 220 | 72 | 30 |
| VIC | 1,647 | 62,258 | 122 | 2,770 | 100 | 2,130 | 82.0 | 165 | 55 | 30 |
| QLD | 1,103 | 50,927 | 122 | 2,805 | 110 | 2,607 | 90.2 | 197 | 52 | 42 |
| SA | 566 | 18,668 | 130 | 2,807 | 115 | 2,032 | 88.5 | 221 | 67 | 28 |
| WA | 666 | 24,131 | 126 | 2,812 | 103 | 2,347 | 81.7 | 213 | 38 | 22 |
| TAS | 209 | 6,700 | 64 | 1,378 | 60 | 1,240 | 93.8 | 125 | 20 | 22 |
| NT | 62 | 2,195 | 32 | 706 | 23 | 496 | 71.9 | 60 | 7 | 12 |
| ACT | 101 | 4,766 | 44 | 1,041 | 36 | 854 | 81.8 | 90 | 18 | 17 |
| **All** | **6,404** | **255,479** | **762** | **17,124** | **649** | **14,172** | **85.2** | **1,291** | **329** | **203** |

*Schools accepted invitation to participate and returned materials from the assessment.
** Effective sample net of excluded schools.

Tables 2.5 and 2.6 provide further details about individual State and Territory participation rates. Table 2.5 describes the levels of absenteeism in the target population and Table 2.6 outlines exemptions present in the target population.

– **Table 2.5: Number of students, by reasons of absenteeism, by State and Territory**

| State / Territory | Absent | Functional disability | Intellectual disability | Linguistic challenge | Other | Refusal | Totals |
|---|---|---|---|---|---|---|---|
| NSW | 220 | 2 | 15 | 13 | 0 | 72 | 322 |
| VIC | 165 | 4 | 21 | 4 | 1 | 55 | 250 |
| QLD | 197 | 9 | 16 | 16 | 1 | 52 | 291 |
| SA | 221 | 6 | 14 | 5 | 3 | 67 | 316 |
| WA | 213 | 5 | 10 | 5 | 2 | 38 | 273 |
| TAS | 125 | 8 | 9 | 4 | 1 | 20 | 167 |
| NT | 60 | 1 | 3 | 8 | 0 | 7 | 79 |
| ACT | 90 | 2 | 6 | 6 | 3 | 18 | 125 |
| **Totals** | **1,291** | **37** | **94** | **61** | **11** | **329** | **1,823** |

– **Table 2.6: Percentage distribution of non-participation, by State and Territory**

| State / Territory | Absent (%) | Refusal (%) | Exempt (%) | Totals (%) |
|---|---|---|---|---|
| NSW | 7.9 | 2.6 | 1.1 | 11.5 |
| VIC | 6.9 | 2.3 | 1.3 | 10.5 |
| QLD | 6.8 | 1.8 | 1.4 | 10.0 |
| SA | 9.4 | 2.9 | 1.2 | 13.5 |
| WA | 8.1 | 1.5 | 0.8 | 10.4 |
| TAS | 8.9 | 1.4 | 1.6 | 11.9 |
| NT | 10.4 | 1.2 | 2.1 | 13.7 |
| ACT | 9.2 | 1.8 | 1.7 | 12.8 |
| **Totals** | **8.1** | **2.1** | **1.3** | **11.4** |

These tables reveal that:

- on average, there was about 8 per cent absenteeism in schools on the test date;
- the project assumes passive participation. Approximately 2 per cent of parents requested that their children not participate in the assessment; and
- on average, 1.3 per cent of students were exempted from the program by principal nomination in accordance with the categories described in Table 2.3.

These data were reported in the *MCEETYA National Year 6 Science Assessment Report 2003*, but were not included in any calculations regarding the achievement of proficiency levels (Chapter 8).

# CHAPTER 3: TEST DESIGN

## 3.1 Background to test development

The PMRT established a number of national committees to ensure that the assessments and results were valid across States and Territories and to advise it on critical aspects of the study. In addition, the contractor, the Australian Council for Educational Research (ACER), established a number of its own committees and ad hoc advisory groups. The key role of all of these committees and groups was to ensure that the scientific literacy assessment domain was inclusive of the different curricula across States and Territories, and that the items that comprised the assessments were fair for the students (valid), irrespective of the State or Territory in which they attended school.

The following sections provide a brief description of the steps that were used to define and deploy the scientific literacy scale. The intention of providing this overview is to make clear:

- the steps that were used to ensure the quality of the assessment tools used to derive the national, State and Territory results; and
- the extent of State and Territory cooperation and involvement in the process.

## 3.2 Test design

### 3.2.1 Defining the assessment domain for scientific literacy

The PISA (1999) definition of scientific literacy formed the basis for the work to assess the scientific literacy of year 6 students in Australia. Associate Professor Mark Hackling of Edith Cowan University prepared an assessment domain for scientific literacy that includes descriptions of the strands, the initial hierarchy of students' understandings and skills and the concept areas (see Appendix 1, MCEETYA Scientific Literacy Framework). These provided the basis for the construction of the assessment items for the study.

The development of the assessment was characterised by a high level of communication with stakeholders and regular feedback and consultation with representatives from each State and Territory, and nominated members of the PMRT. It was through these committees and groups that the experts from States and Territories commented on and contributed to all aspects of the study.

### 3.2.2 Describing increasingly complex student understandings and skills within the domain

As with measurement in the physical sciences, the measurement of students' proficiency in scientific literacy required the development of a measurement scale. The scale was conceptualised by describing the main understandings and skills that students were expected to develop during the compulsory years of schooling. To create the measurement scale, descriptions were developed to form a hierarchy of increasingly complex understandings and skills. To give additional meaning to the hierarchy, the descriptions were later linked to the items from the assessment. Difficult tasks or items that challenged the most able students were located at the upper end of the hierarchy and define the upper end of the measurement scale.

Conversely, those items that were relatively easy and could be answered by students with few scientific literacy understandings and skills were located toward the base of the hierarchy, defining the lower end of the measurement scale. This continuum of easy-to-difficult items represents increasingly more complex skills and understandings and defines the scientific literacy measurement scale that is central to the analysis of results in this report.

### 3.2.3 Constructing assessments comprising items and tasks that operationally defined the assessment domain and covered the full range of proficiency expected to be represented in year 6 classes

As stated above, the hierarchy of students' understandings and skills, together with the concept areas, described what was to be assessed. Items and tasks were then constructed to provide an operational definition of scientific literacy.

Test constructors developed items and tasks that enabled students at different points along the scale to demonstrate what they knew and could do in terms of scientific literacy. They had to ensure that the tasks assessed the outcomes articulated in the assessment domain and that items assessing higher-order understandings and skills at the top of the scale were, in fact, harder than items located at the bottom of the scale.

The advisory committees and working groups had significant input into this part of the assessment. ACER and its review panels initially reviewed the items, prior to State and Territory advisory committees and other key staff reviewing them. At a later stage the items were reviewed again, following trialing of the items and tasks with samples of students in four States and Territories.

The PMRT set the policy objectives of the assessment and the policy priorities for the implementation of the program. This included endorsing the definition of scientific literacy, the assessment domain, items developed for the assessment and the plans for reporting the results. It also endorsed, after receiving advice from the contractor, the more technical aspects of the design, including, for example, the number of assessment booklets, the ratio of multiple-choice to open-ended items in the booklets and the number of items per domain per test booklet.

Teachers and curriculum experts had a number of opportunities during this stage to review and suggest modifications to the tasks. Teachers were also involved when marking the tasks. The emphasis during these review exercises was on ensuring that the items and tasks reflected the understandings and skills in the assessment domain and were not biased unduly for or against particular groups of students.

### 3.2.4 Trialing and final item selection

When the items and tasks had been written, they were trialed with students in a sample of 24 schools selected from the government, Catholic and independent sectors in New South Wales, the Northern Territory, Victoria, Queensland and Western Australia. The results were analysed in a systematic way to determine the degree to which the items and tasks measured the scientific literacy domain. The committees then reviewed the data from the trial testing, gauged the validity of the assessments and suggested modifications where necessary. These suggestions were then included in the revised assessments.

Ten trial forms were prepared, which allowed a selection of item types to be assessed. These included multiple-choice items, short-response items requiring single word or single sentence responses, and extended-response items that

required an explanation or description (typically of more than one sentence) to achieve maximum scores. Of the ten forms, six involved content areas and four were practical tasks.

A total of 164 items was trialed in the first iteration of the process. A second iteration was requested to include items from content areas that were considered to be imperatives in science learning. This resulted in a further 22 items being prepared, reviewed, trialed and linked to the original scale. Detailed information regarding this stage of the project was presented to PMRT in a report entitled, *Assessment Instruments and Key Performance Measures for Primary School Science* (May 2003). This report explained changes that were made to items and scoring keys as a result of the trial.

Teachers involved in the assessment process were requested to complete a PSAP Session Report Form (see Appendix 2, National Year 6 Science Assessment 2003: Session Report Form). This document gave them the opportunity to comment on issues such as the context, content and timing of the assessment. It was returned with the test booklets and reviewed by ACER. Overall, the feedback was highly positive and informed the final item selection process.

The Teacher's Administration Guide gave teachers the opportunity to react to various aspects of the assessment as well as to report implementation compliance and issues of concern. These tended to be supportive of the assessment and, in particular, of the practical exercises. There was no evidence of inadequate time being allowed for either component of the test.

A final item set, totalling 70 items, was selected after extensive reviews by PMRT representatives, the Australian Science Academy and State and Territory representatives on the PSAP Review Committee.

Table 3.1 shows how the final selection of items compared with the test specification in the contract, which was changed as a result of negotiations with States and Territories and PMRT.

**Table 3.1: Comparison between initial and final tables of specification**

| Category | Ideal (%) | Final items (%) |
|---|---|---|
| **Scientific literacy strands** | | |
| Earth and Beyond | 25 | 20 |
| Energy and Change | 25 | 24 |
| Life and Living | 25 | 30 |
| Natural and Processed Materials | 25 | 26 |
| **Conceptual domains** | | |
| A | 40 | 46 |
| B and C | 60 * | 54 |
| **Item types** | | |
| Multiple choice/short answer | 50 | 51 |
| Extended response | 50 | 49 |

Table 3.1 shows, for example, that the initial target was to have 25 per cent of the items assessing the Earth and Beyond strand and in the final test 20 per cent of the items assessed this strand. Similarly, the initial requirements were that 50 per cent of the assessments would comprise multiple-choice/short answer type items: in the final test, they comprised 51 per cent.

The final assessment was made up of four test forms: Objective Form A, Objective Form B, Practical Form A (Craters) and Practical Form B (Parachutes).

To maximise the interactions between items and students, the test forms were distributed in a rotational pattern so that within a classroom students would be participating in the same Practical task, but alternate students would be undertaking different forms of the Objective tasks.

For the purpose of data analysis, this assessment design produced four different combinations of test data to reflect the combinations of test forms undertaken at the class level:

- Objective Form A with Practical Form A (AA)
- Objective Form A with Practical Form B (AB)
- Objective Form B with Practical Form A (BA)
- Objective Form B with Practical Form B (BB).

Ten items were common to Objective Forms A and B. Of the 70 items, 24 were multiple-choice and 46 items were hand-scored. The marking scheme that had been trialed was revised so the final rubric recommended that 36 of the hand-scored items be scored dichotomously (0/1 scale) and ten scored polytomously (0/0.5 scale).

## 3.3 Structure of the final test

The distribution of items across the assessment domain for scientific literacy (strands of the domain and major concept areas) is shown in Table 3.2.

The 70 items in the two pencil-and-paper tests and two practical tasks were worth a total of 78 marks. Each student had to sit one pencil-and-paper test and one practical task.

– **Table 3.2: Structure of the final test**

| Domain aspect | Item type and number of items | | | |
|---|---|---|---|---|
| | Multiple-choice | Short-answer | Extended response | Totals |
| **Distribution of items by strand** | | | | |
| Strand A | 14 | 5 | 13 | **32** |
| Strand B | 4 | 2 | 10 | **16** |
| Strand C | 6 | 3 | 13 | **22** |
| **Totals** | **21** | **12** | **37** | **70** |
| ***Distribution of items by major science conceptual area*** | | | | |
| Life and Living | 6 | 5 | 10 | **21** |
| Earth and Beyond | 5 | 1 | 8 | **14** |
| Natural and Processed Materials | 10 | 1 | 7 | **18** |
| Energy and Change | 3 | 3 | 11 | **17** |
| **Totals** | **24** | **10** | **36** | **70** |

The scientific literacy domain has specified concepts in terms of major thematic areas rather than within traditional subject boundaries such as physics, chemistry, biology, etc.

These thematic areas are articulated above. They are considered to be of more relevance to students at primary school and, according to PISA, 'to all people in their lives beyond school than the more traditional subject areas' (PISA, 1999, p. 97).

Table 3.2 shows that the items were distributed relatively evenly across conceptual areas, and that the strands and major science concepts of scientific literacy were assessed through a range of item types. Thirty-seven of the items were classified as being in the extended-response format, 21 in the multiple-choice format and 12 in the short-answer format.

Almost all of the items were presented in item sets or units, with two or more items relating to each stimulus text and/or diagram.

# CHAPTER 4: MAIN TEST AND DATA PREPARATION

## 4.1 Administering the tests to students

Assessment materials were sent to 672 schools and included a School Coordinator's Guide, a Test Administrator's Guide and the assessment instrument, with the appropriate practical materials for the particular combination of tasks being undertaken.

The final assessments were administered to a stratified random sample of students between 21 and 30 October 2003 and 14,172 valid responses were received from 649 schools.

## 4.2 The assessment

The assessment consisted of two booklets that students responded to in a pencil-and-paper format and two booklets in which the students first performed a practical activity in small groups and then responded individually to pencil-and-paper items related to the activity.

The multiple-choice [MC] items had only one correct answer. The open-ended items required students to construct their own responses. Some of these items had only one correct answer, while others provided a partial credit-marking scheme to accommodate a range of response levels. These latter items allowed for a wider range of skills to be assessed. The open-ended items have been further classified into those that require a single word or short sentence response [SS] and those that require a more substantive response [SL].

Each of the items and tasks included stimulus material to contextualise a series of questions relating to the material. Both pencil-and-paper objective assessments contained 13 units (a unit is stimulus material and a number of associated questions). The two practical assessments comprised a practical task, group activities and either four or five items to be answered individually by students.

## 4.3 Test administration procedures

The regular class teacher administered the assessment. This was done to minimise the disruption to the normal class environment.

Standardised administration procedures were developed by the contractor and brought together in an Assessment Administrator's Manual. In all schools where students were to complete the assessments, teachers and school administrators were provided with the manual. Detailed instructions were also provided regarding the participation and exclusion of students with disabilities and students from non-English speaking backgrounds.

To assist in standardising the assessment conditions and familiarising teachers with the procedures and requirements of the practical assessment, training videos were produced.

The teachers were able to review the manual and the training videos before the assessment date and raise any questions they had about the procedures with the State and Territory Coordinators responsible for the program.

As a result, it was expected that standardised administration of the assessments would be achieved.

A quality-monitoring program was also implemented, to gauge the extent to which class teachers followed the administration procedures. This involved trained monitors observing the administration of the assessments in a random sample of classes across the nation. Forty-eight of the 672 schools (included 23 duplicate stratum representation schools) were observed and the monitors reported a high degree of conformity with the administration procedures.

## 4.4 Marking of responses to open-ended items

The assessment tasks were marked centrally. Approximately two-thirds of the items were open-ended, necessitating the use of trained markers, and most required a single answer or phrase that could be marked objectively.

Marking guides were prepared by the contractor and refined during the trial process. The marking team included nominated representatives from most States and Territories and a team of experienced markers employed by the contractor.

The markers participated in a five-hour training session conducted by a member of the test construction team. In addition, the markers undertook two hours of marking in which a pair of markers marked the same student answer books and moderators reconciled differences in discussion with the markers. The session involved formal presentations by the trainers, followed by hands-on practice with sample student answer books.

Markers were monitored constantly for reliability by having samples of their student answer books remarked by senior markers. In cases where there were differences between the markers and senior markers, the scoring was reconciled jointly. This approach, coupled with the intensive training at the beginning of the marking exercise, ensured that markers were applying the criteria consistently.

## 4.5 Data entry procedures

Individual student responses to multiple-choice questions were captured by scanner. The contractor was required to perform a validation of the scanning process to ensure the accuracy of the recording. This displayed 100 per cent accuracy in data capture.

Images of all student responses were collected and stored.

## 4.6 Quality assurance checks used to check all data

### 4.6.1 Data coding rules

#### 4.6.1.1 Rules for counting and categorising student groups

Several schools that had initially declined to participate in the assessment, and had been replaced in the sample, subsequently indicated that they were able to

participate. In such cases, both schools participated and the data from the replacement school was removed from the data used for analysis.

In several schools, whole multi-year classes were allowed to complete the assessment. Data for students who were outside the target group of year 6 students were eliminated from the analysis.

Table 4.1 indicates the effect of these procedures.

– **Table 4.1: Total number of tests returned, by year level and test form**

| | Test form | | | | |
|---|---|---|---|---|---|
| **Year level** | **AA** | **AB** | **BA** | **BB** | **All** |
| 5 | 122 | 71 | 93 | 66 | 352 |
| 6 | 3,633 | 3,602 | 3,783 | 3,548 | 14,566 |
| 7 | 34 | 25 | 36 | 35 | 130 |
| Other | 37 | 24 | 31 | 18 | 110 |
| **Totals** | **3,826** | **3,722** | **3,943** | **3,667** | **15,158** |

As Table 4.1 shows, there was a total of 14,566 year 6 candidates for the analysis of the study. Three hundred and ninety-four students were excluded from the data pool following conditioning of the data to remove duplicate stratum representation. This resulted in a final data set of 14,172 students.

#### 4.6.1.2 Student demographics

Student-level demographic data were collected through the test forms. The types of data are shown in Table 4.2.

– **Table 4.2: Student demographic data**

| Student data collected | Format |
|---|---|
| Gender | Boy (1) or Girl (2) |
| Aboriginal status | No (1), Aboriginal (2), Torres Strait Islander (3), Both Aboriginal AND Torres Strait Islander (4) |
| Age | Free response, 2 digits |
| Grade/Level | Year 5 (1), Year 6 (2), Year 7 (3), Other (4) |
| Language other than English at home | No (1), Yes (2) if Yes specify. |
| Country born (student) | Australia (1), Other (2) if Other specify |
| Country born (Mother) | Australia (1), Other (2) if Other specify |
| Country born (Father) | Australia (1), Other (2) if Other specify. |

The coding rules shown in Table 4.3 were applied to the demographic information.

– **Table 4.3: Coding rules for demographic responses**

| Group | Rule |
|---|---|
| Gender | Classified by response; missing data treated as missing unless the student was present at a single-sex school |
| Grade | Classified by response; only Year 6 included in the analysis. All other cases were removed |
| ATSI | Coded as ATSI if response was 'yes' to Aboriginal, OR Torres Strait Islander OR both |
| LBOTE | Coded as LBOTE if response was 'yes' to 'Do you speak another language at home?' |

Table 4.4 summarises the demographic profile of the sample achieved:

– **Table 4.4: Summary of the sample demographics**

| Gender | Frequency | Percentage |
|---|---|---|
| Female | 6,875 | 48.5 |
| Male | 7,246 | 51.1 |
| Not identified | 19 | 0.1 |
| Missing | 32 | 0.2 |
| **Totals** | **14,172** | **100.0** |
| Non-ATSI | 13,368 | 94.3 |
| ATSI | 593 | 4.2 |
| Not identified | 211 | 1.5 |
| **Totals** | **14,172** | **100.0** |
| English-speaking background | 12,397 | 87.5 |
| LBOTE | 1,662 | 11.7 |
| Not identified | 113 | 0.8 |
| **Totals** | **14,172** | **100.0** |

The next chapter considers the actual sampling procedure that was used in the final study.

# CHAPTER 5: SAMPLING WEIGHTS AND SAMPLING VARIANCE

## 5.1 Weighted sample

The sample frame provides data on each school and its population demographics.

The weighting principles and calculations follow those established for international programs such as TIMSS and PISA, and are documented in Chapter 11 of the *TIMSS 1999 Technical Report* (2000), the main source for the following description.

The basic sample design used in sample was a two-stage stratified cluster design, with schools as the first stage and classrooms as the second. The design required schools to be sampled using a PPS systematic method and classrooms to be sampled by sorting class groups alphabetically by teacher name and selecting the second class on the sorted list. This was considered an efficient methodology of a pseudo-random sample at class level. The cluster size was assumed to be 25.

While the multi-stage stratified cluster design provides very economical and effective data collection process in a school environment, it results in differential probabilities of selection for the ultimate sampling elements, the students. Consequently, one student in the assessment does not necessarily represent the same proportion of students in the population as another, as would be the case with a simple random sampling approach. To account for differential probabilities of selection, due to the design and to ensure proper survey estimates, a sampling weight was computed for each participating student. The ability to provide proper sampling weights was an essential characteristic of an acceptable sample design, since appropriate sampling weights were essential for the computation of accurate population estimates.

## 5.2 Weighting procedure

The weighting procedure required three steps, reflecting the sample design.

The first step consisted of calculating a school weight; this also incorporated weighting factors from any additional front-end sampling stages, such as districts or regions. A school-level participation adjustment was then made to the school weight to compensate for any sampled schools that did not participate. That adjustment was calculated independently for each stratum.

In the second step, a classroom weight was calculated. No classroom-level participation adjustment was necessary, as in most cases a single classroom was sampled in each school. The classroom weight was calculated independently for each school.

The third and final step consisted of calculating a student weight.

A non-participation adjustment was made to compensate for students who did not take part in the testing. The student weight was calculated independently for each sampled classroom. The basic sampling weight attached to each student record was the product of the three intermediate weights: the first stage (school) weight, the second stage (classroom) weight and the third stage (student) weight. The

overall student sampling weight was the product of these three weights and the two non-participation adjustments, school-level and student-level.

## 5.3 Replication procedures

In cases where the sampling design involved multi-stage cluster sampling, there were several options for estimating sampling errors that avoided the assumption of simple random sampling (Wolter, 1985).

The jackknife repeated replication technique (JRR) was chosen for the sample analysis.

JRR is internationally accepted and was the procedure used in TIMSS in both 1995 and 1999. JRR is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals and percentages.

The general use of JRR entails assigning pairs of schools systematically to sampling zones and selecting one of these schools randomly to have its contribution doubled and the other to have it zeroed, so as to construct a number of 'pseudo-replicates' of the original sample. The statistic of interest is computed once for the entire original sample, and once again for each pseudo-replicate sample. The variation between the estimates for each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

The variation on the JRR technique used followed the procedures of TIMSS 1999, as described by Johnson and Rust (1992). It assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique accounts for the combined effect of the between-and-within PSU contributions to the sampling variance. The process is described in detail in Chapter 12 of the *TIMSS 1999 Technical Report* (2000).

The next chapter considers some of the features of the analysis that were used to construct the measurement scale.

# CHAPTER 6: ITEM AND RASCH ANALYSIS

## 6.1 Introduction

The Rasch measurement model was used to analyse the results from the sample of students who attempted the assessment. This model is used in all State and Territory testing programs and in the major international testing programs such as PISA and TIMSS.

The following steps were implemented to screen and prepare the data for analysis.

### 6.1.1 Data cleaning: out-of-range responses

SPSS statistical analysis software was used to clean the data. Minor cases of spurious student demographic data were detected and these were coded as unidentified. The scanning contractor had adhered rigorously to the codebook provided and delivered a clean data set. All anomalies were resolved in the scanning and data validation phases.

### 6.1.2 Consistency checks

In order to ensure the completeness of the data set provided by the scanning contractor, comparisons were undertaken between the logging sheets from marking with the data returned from the scanning process. The checks confirmed that all booklets submitted had been scanned and that data had been provided for each student. There were no duplicate records in the final data set. A comparison of counts by school was also undertaken.

### 6.1.3 Validation of item keys and link items

The response frequencies for all items and percentage response by category were determined. Initial item analyses were undertaken to identify any incorrect keys in the codebook. Frequency checks were carried out to ensure that the response range and patterns were consistent in the link items between Objective Form A and Objective Form B.

## 6.2 Item analysis

Before the Rasch analysis was performed, item analyses were undertaken to determine the traditional parameters for each item. These are reported at Appendix 3, Item Level Statistics and include the percentage correct overall and by group, the incidence of missing responses and the Findlay Index of Discrimination.

## 6.3 Rasch analysis

### 6.3.1 Rasch analysis and missing values

A consequence of the assessment design was that the data matrix contained significant amounts of missing data for each case, representing those components that had not been attempted. The codebook provided to the scanning contractor

required all missing data to be coded '9'. These data needed to be conditioned to reflect the level of student test engagement.

There were two classifications of missing data:

- those due to non-participation in a test form; and,
- those due to a failure to attempt an item on a form for a test in which a student had participated.

In cases where data was missing due to a student's non-participation in a test form, all data were coded as '9' (missing). For the purpose of calibration, analysis and reporting, these data were treated as missing and not included in any calculations pertaining to the items or the estimation of student ability.

For analysis purposes, the approach used to treat missing data was as follows:

- If a student had not attempted all items in the form, the point at which he or she had discontinued engagement was determined. It was assumed that the first missing response in a student response string that indicated disengagement was the last item the student had attempted. This item was coded as wrong. All items beyond this point in the student response string were treated as missing and coded '9'.
- Missing items within the test were coded as incorrect. It was assumed that the student had attempted the item for which a missing response was recorded but could not provide an answer.

  These rules were used for the calibration of the item parameters and the generation of the raw score to ability tables.

  In the actual measurement process, all missing responses within a test form were treated as wrong. The raw score obtained in this way was a sufficient statistic for the generation of the ability estimates.

### 6.3.2 Item calibration

After consultation with MCEETYA's Benchmarking and Educational Measurement Unit (BEMU), it was decided that the item calibration would be performed by generating equally sized samples from each State and Territory.

Item calibration was undertaken on a random sample of data comprising 200 students from each State and Territory, resulting in a calibration sample of 1,600 students. Equal numbers from each jurisdiction were used to ensure that larger States did not bias the construct in any way.

A second sample of 500 students was drawn from each State and Territory. This sample of 4,000 students was used to check the fit of the data to the model.

The equally weighted jurisdiction sample of 1,600 was used to generate the item locations (item difficulties) and the raw score-to-person measure (ability) estimates for each test combination. Common items were used to link Form A and Form B. The full data matrix was run concurrently using the Rasch Unidimensional Mathematical Models (RUMM) program with missing data treated as described above.

Table 6.1 shows the centralised item locations (item difficulties) for each item. The shaded items in Table 6.1 are the items that are common to the objective assessments of both Form A and Form B.

A number of items did not fit the model. Reverse thresholds were identified in two polytomously scored items.

After consultation with BEMU, the following actions were taken:

- Item 15 in Objective Form A proved to be ambiguous and two responses were accepted as correct;
- Item 29 in Objective Form A displayed a reverse threshold and was collapsed to a 0/1 response category; and
- Item 11 in Objective Form B displayed a reverse threshold and was collapsed to a 0/1 response category item.

The item calibrations were then used with the test equating routine in RUMM 2020 to produce raw score-to-ability tables for each of the combinations of test forms.

**Table 6.1: Item locations based on a sample of 1,600 students (200 from each State and Territory)**

| Item code | Item type | Item name | Item location | Standard error |
|---|---|---|---|---|
| I0001 | Poly | a1ic1r | -2.287 | 0.151 |
| I0002 | Poly | a2ic2r | -0.124 | 0.082 |
| I0003 | MC | a3b1m1r | -1.060 | 0.069 |
| I0004 | Poly | a4b2m2r | -0.158 | 0.057 |
| I0005 | Poly | a5b3m3r | -0.463 | 0.060 |
| I0006 | MC | a6fl1r | -1.318 | 0.108 |
| I0007 | Poly | a7fl2r | -1.295 | 0.107 |
| I0008 | MC | a8fl3r | -1.608 | 0.118 |
| I0009 | MC | a9e1r | -0.925 | 0.096 |
| I0010 | MC | a10e2r | -2.607 | 0.172 |
| I0011 | MC | a11e3r | 0.344 | 0.078 |
| I0012 | Poly | a12b4p1r | 0.809 | 0.054 |
| I0013 | MC | a13b5p2r | -0.610 | 0.062 |
| I0014 | MC | a14b20c1 | -3.155 | 0.151 |
| I0015 | MC | rec_a15 | -1.990 | 0.094 |
| I0016 | Poly | a16b22c3 | -0.030 | 0.056 |
| I0017 | Poly | a17m1r | -0.167 | 0.082 |
| I0018 | Poly | a18m2r | 0.943 | 0.077 |
| I0019 | Poly | a19m3r | 0.091 | 0.079 |
| I0020 | Poly | a20m4r | 2.064 | 0.064 |
| I0021 | MC | a21pc1r | -1.198 | 0.104 |
| I0022 | Poly | a22pc2r | 1.088 | 0.077 |
| I0023 | MC | a23bu1r | 0.413 | 0.077 |
| I0024 | Poly | a24bu2r | -0.541 | 0.088 |
| I0025 | Poly | a25bu3r | 0.524 | 0.063 |
| I0026 | Poly | a26wb1r | -0.027 | 0.056 |
| I0027 | MC | a27wb2r | -0.814 | 0.094 |
| I0028 | Poly | a28sp1r | -1.786 | 0.127 |
| I0029 | Poly | rec_a29 | -0.684 | 0.092 |
| I0030 | Poly | a30tt1r | 1.054 | 0.078 |
| I0031 | Poly | a31tt2r | -0.133 | 0.083 |
| I0032 | Poly | a32tt3r | 1.263 | 0.080 |
| I0033 | Poly | a33tt4r | 1.592 | 0.083 |
| I0034 | Poly | a34b28p1 | 0.651 | 0.035 |

| | | | | |
|---|---|---|---|---|
| I0035 | MC | a35b29p2 | 0.692 | 0.056 |
| I0036 | MC | b6cf1r | -0.664 | 0.087 |
| I0037 | Poly | b7cf2r | -0.974 | 0.094 |
| I0038 | MC | b8cf3r | -0.028 | 0.079 |
| I0039 | Poly | b9cd1r | 0.970 | 0.077 |
| I0040 | MC | b10cd2r | 0.704 | 0.076 |
| I0041 | Poly | rec_b11 | -1.556 | 0.111 |
| I0042 | Poly | b12aw1r | 0.431 | 0.076 |
| I0043 | Poly | b12bw2r | 0.467 | 0.076 |
| I0044 | Poly | b13w3r | -1.027 | 0.095 |
| I0045 | Poly | b14w4r | 0.181 | 0.077 |
| I0046 | Poly | b15sh1r | 1.354 | 0.079 |
| I0047 | MC | b16sh2r | -1.886 | 0.125 |
| I0048 | Poly | b17er1r | 0.749 | 0.052 |
| I0049 | MC | b18er2r | -1.738 | 0.119 |
| I0050 | Poly | b19er3r | 1.191 | 0.078 |
| I0051 | MC | b23se1r | 0.692 | 0.076 |
| I0052 | MC | b24se2r | 0.452 | 0.076 |
| I0053 | Poly | b25se3r | 1.927 | 0.087 |
| I0054 | Poly | b26bp1r | -0.159 | 0.067 |
| I0055 | Poly | b27bp2r | 1.481 | 0.081 |
| I0056 | Poly | b30c1r | 1.490 | 0.061 |
| I0057 | MC | b31c2r | -0.391 | 0.086 |
| I0058 | MC | b32c3r | 0.528 | 0.079 |
| I0059 | Poly | b33ex1r | 0.833 | 0.057 |
| I0060 | MC | b34ex2r | -0.958 | 0.100 |
| I0061 | MC | b35ex3r | -0.279 | 0.087 |
| I0062 | Poly | ap36_1r | 0.231 | 0.078 |
| I0063 | Poly | ap37_2r | 0.812 | 0.076 |
| I0064 | Poly | ap38_3r | 0.866 | 0.077 |
| I0065 | Poly | ap39_4r | 2.302 | 0.095 |
| I0066 | Poly | bp36_1r | 1.166 | 0.079 |
| I0067 | Poly | bp37_2r | -1.085 | 0.099 |
| I0068 | Poly | bp38_3r | 1.909 | 0.087 |
| I0069 | Poly | bp39_4r | 1.172 | 0.079 |
| I0070 | Poly | bp40_5r | 0.288 | 0.079 |

Note: The items shaded in Table 6.1 are the items common to both Objective Forms of the assessment.

The next section examines the fit of the data to the Rasch model.

### 6.3.3 Item fit to the Rasch model

The initial analysis showed that Item 15a required to be rescored, with options A and B both being marked correct, and Items 11b and 29a and 11b to be rescored because there were reverse thresholds identified. These refinements were carried out.

A more appropriate measure in this instance is to examine the misfit associated with individual items. Table 6.2 shows the item fit for all the items in the tests after the changes outlined above had been carried out.

– **Table 6.2: Item fit based on Rasch analysis of all Primary Science items**

| Seq | Item code | Type | Item name | Location | SE | Fit residual | Chi squ | Prob |
|---|---|---|---|---|---|---|---|---|
| 1 | I0001 | Poly | a1ic1r | -2.287 | 0.151 | 0.692 | 8.775 | 0.032 |
| 2 | I0002 | Poly | a2ic2r | -0.124 | 0.082 | -1.473 | 5.380 | 0.146 |
| 3 | I0003 | MC | a3b1m1r | -1.060 | 0.069 | 3.441 | 54.825 | 0.000 |
| 4 | I0004 | Poly | a4b2m2r | -0.158 | 0.057 | 5.347 | 36.144 | 0.000 |
| 5 | I0005 | Poly | a5b3m3r | -0.463 | 0.060 | 0.495 | 3.697 | 0.296 |
| 6 | I0006 | MC | a6fl1r | -1.318 | 0.108 | -0.490 | 2.498 | 0.476 |
| 7 | I0007 | Poly | a7fl2r | -1.295 | 0.107 | 0.456 | 2.225 | 0.527 |
| 8 | I0008 | MC | a8fl3r | -1.608 | 0.118 | -0.360 | 2.393 | 0.495 |
| 9 | I0009 | MC | a9e1r | -0.925 | 0.096 | 1.251 | 7.085 | 0.069 |
| 10 | I0010 | MC | a10e2r | -2.607 | 0.172 | -1.593 | 7.148 | 0.067 |
| 11 | I0011 | MC | a11e3r | 0.344 | 0.078 | 2.559 | 5.474 | 0.140 |
| 12 | I0012 | Poly | a12b4p1r | 0.809 | 0.054 | 2.401 | 3.870 | 0.276 |
| 13 | I0013 | MC | a13b5p2r | -0.610 | 0.062 | 2.616 | 13.151 | 0.004 |
| 14 | I0014 | MC | a14b20c1 | -3.155 | 0.151 | -0.604 | 1.031 | 0.794 |
| 15 | I0015 | MC | rec_a15 | -1.990 | 0.094 | -0.700 | 1.957 | 0.581 |
| 16 | I0016 | Poly | a16b22c3 | -0.030 | 0.056 | -0.598 | 11.204 | 0.011 |
| 17 | I0017 | Poly | a17m1r | -0.167 | 0.082 | -0.264 | 3.298 | 0.348 |
| 18 | I0018 | Poly | a18m2r | 0.943 | 0.077 | -2.767 | 12.123 | 0.007 |
| 19 | I0019 | Poly | a19m3r | 0.091 | 0.079 | -1.334 | 9.717 | 0.021 |
| 20 | I0020 | Poly | a20m4r | 2.064 | 0.064 | -2.488 | 15.447 | 0.001 |
| 21 | I0021 | MC | a21pc1r | -1.198 | 0.104 | 1.163 | 17.212 | 0.001 |
| 22 | I0022 | Poly | a22pc2r | 1.088 | 0.077 | -0.548 | 0.337 | 0.953 |
| 23 | I0023 | MC | a23bu1r | 0.413 | 0.077 | 1.833 | 1.787 | 0.618 |
| 24 | I0024 | Poly | a24bu2r | -0.541 | 0.088 | -0.702 | 2.636 | 0.451 |
| 25 | I0025 | Poly | a25bu3r | 0.524 | 0.063 | 0.635 | 11.788 | 0.008 |
| 26 | I0026 | Poly | a26wb1r | -0.027 | 0.056 | -0.629 | 5.319 | 0.150 |
| 27 | I0027 | MC | a27wb2r | -0.814 | 0.094 | -1.388 | 4.743 | 0.192 |
| 28 | I0028 | Poly | a28sp1r | -1.786 | 0.127 | -0.414 | 1.534 | 0.674 |
| 29 | I0029 | Poly | rec_a29 | -0.684 | 0.092 | -2.316 | 16.973 | 0.001 |
| 30 | I0030 | Poly | a30tt1r | 1.054 | 0.078 | 2.353 | 5.991 | 0.112 |
| 31 | I0031 | Poly | a31tt2r | -0.133 | 0.083 | 1.940 | 14.228 | 0.003 |
| 32 | I0032 | Poly | a32tt3r | 1.263 | 0.080 | -1.296 | 7.896 | 0.048 |
| 33 | I0033 | Poly | a33tt4r | 1.592 | 0.083 | -1.548 | 10.151 | 0.017 |
| 34 | I0034 | Poly | a34b28p1 | 0.651 | 0.035 | 4.102 | 20.967 | 0.000 |
| 35 | I0035 | MC | a35b29p2 | 0.692 | 0.056 | 0.110 | 1.135 | 0.769 |
| 36 | I0036 | MC | b6cf1r | -0.664 | 0.087 | -0.921 | 4.087 | 0.252 |
| 37 | I0037 | Poly | b7cf2r | -0.974 | 0.094 | -0.581 | 1.639 | 0.651 |
| 38 | I0038 | MC | b8cf3r | -0.028 | 0.079 | 1.598 | 1.286 | 0.733 |
| 39 | I0039 | Poly | b9cd1r | 0.970 | 0.077 | 1.166 | 0.196 | 0.978 |
| 40 | I0040 | MC | b10cd2r | 0.704 | 0.076 | 1.030 | 2.905 | 0.407 |
| 41 | I0041 | Poly | rec_b11 | -1.556 | 0.111 | -1.313 | 5.685 | 0.128 |
| 42 | I0042 | Poly | b12aw1r | 0.431 | 0.076 | 0.335 | 2.828 | 0.419 |
| 43 | I0043 | Poly | b12bw2r | 0.467 | 0.076 | 1.114 | 0.602 | 0.896 |
| 44 | I0044 | Poly | b13w3r | -1.027 | 0.095 | -2.073 | 33.944 | 0.000 |
| 45 | I0045 | Poly | b14w4r | 0.181 | 0.077 | -1.743 | 13.615 | 0.003 |
| 46 | I0046 | Poly | b15sh1r | 1.354 | 0.079 | -0.850 | 11.084 | 0.011 |
| 47 | I0047 | MC | b16sh2r | -1.886 | 0.125 | 0.187 | 7.296 | 0.063 |
| 48 | I0048 | Poly | b17er1r | 0.749 | 0.052 | 1.235 | 8.733 | 0.033 |
| 49 | I0049 | MC | b18er2r | -1.738 | 0.119 | -0.967 | 2.292 | 0.514 |
| 50 | I0050 | Poly | b19er3r | 1.191 | 0.078 | -3.832 | 28.830 | 0.000 |
| 51 | I0051 | MC | b23se1r | 0.692 | 0.076 | 1.639 | 9.963 | 0.019 |
| 52 | I0052 | MC | b24se2r | 0.452 | 0.076 | -1.085 | 11.425 | 0.010 |
| 53 | I0053 | Poly | b25se3r | 1.927 | 0.087 | -2.540 | 26.222 | 0.000 |
| 54 | I0054 | Poly | b26bp1r | -0.159 | 0.067 | 0.022 | 1.362 | 0.714 |
| 55 | I0055 | Poly | b27bp2r | 1.481 | 0.081 | -0.730 | 7.297 | 0.063 |
| 56 | I0056 | Poly | b30c1r | 1.490 | 0.061 | -0.992 | 5.478 | 0.140 |
| 57 | I0057 | MC | b31c2r | -0.391 | 0.086 | 1.784 | 7.856 | 0.049 |
| 58 | I0058 | MC | b32c3r | 0.528 | 0.079 | 2.196 | 7.557 | 0.056 |
| 59 | I0059 | Poly | b33ex1r | 0.833 | 0.057 | -1.323 | 2.843 | 0.417 |
| 60 | I0060 | MC | b34ex2r | -0.958 | 0.100 | -1.983 | 18.679 | 0.000 |
| 61 | I0061 | MC | b35ex3r | -0.279 | 0.087 | -2.227 | 11.678 | 0.009 |
| 62 | I0062 | Poly | ap36_1r | 0.231 | 0.078 | 2.234 | 5.201 | 0.158 |
| 63 | I0063 | Poly | ap37_2r | 0.812 | 0.076 | 1.252 | 1.234 | 0.745 |
| 64 | I0064 | Poly | ap38_3r | 0.866 | 0.077 | -0.096 | 1.557 | 0.669 |
| 65 | I0065 | Poly | ap39_4r | 2.302 | 0.095 | -0.828 | 3.034 | 0.386 |
| 66 | I0066 | Poly | bp36_1r | 1.166 | 0.079 | -1.453 | 12.069 | 0.007 |
| 67 | I0067 | Poly | bp37_2r | -1.085 | 0.099 | -0.270 | 9.670 | 0.022 |
| 68 | I0068 | Poly | bp38_3r | 1.909 | 0.087 | -0.189 | 4.535 | 0.209 |
| 69 | I0069 | Poly | bp39_4r | 1.172 | 0.079 | -0.524 | 8.910 | 0.031 |
| 70 | I0070 | Poly | bp40_5r | 0.288 | 0.079 | 0.931 | 0.221 | 0.974 |

Analysis of the information in Table 6.2 shows that the following items are the worst fitting (See Table 6.3). They have relatively large chi square values and relatively large fit residuals.

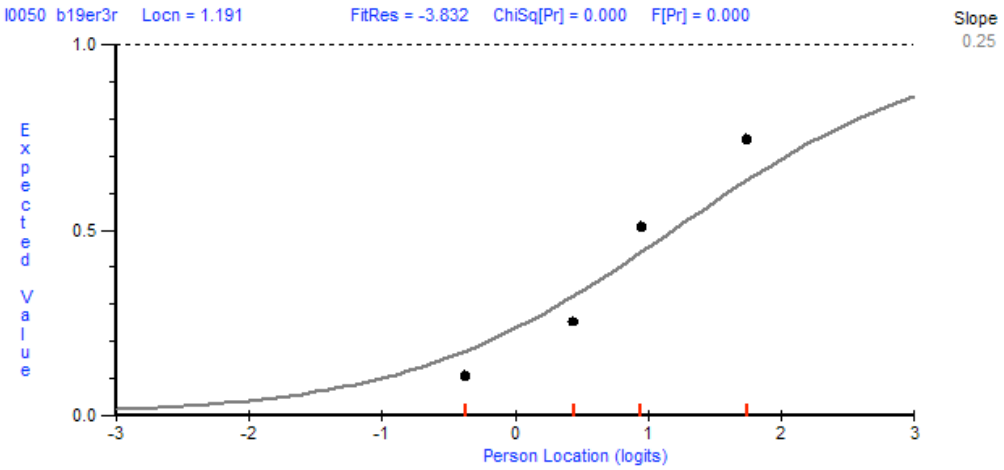
– **Table 6.3: Worst-fitting items**

| Item number | Item name | Item | Item location | Standard error | Fit residual | Df | Chi square |
|---|---|---|---|---|---|---|---|
| 34 | A35B29 Paper Plane 2 | Poly | 0.651 | 0.035 | 4.102 | 1463.76 | 20.967 |
| 44 | B13 Wheat 2 | Poly | -1.027 | 0.095 | -2.073 | 780.54 | 33.944 |
| 50 | B19 Erosion 3 | Poly | 1.191 | 0.078 | -3.832 | 777.62 | 28.830 |
| 4 | A04B02 Bar Magnets | Poly | -0.158 | 0.057 | 5.347 | 1545.51 | 36.144 |
| 3 | A03B01 Bar Magnets | MC | -1.060 | 0.069 | 3.441 | 1545.51 | 54.825 |
| 53 | B25 School Electricity 3 | Poly | 1.927 | 0.087 | -2.540 | 769.84 | 26.222 |

It is interesting to note that the worst-fitting items are either those that require an answer to be supplied (polytomously scored items), or those that appear as link items in two forms of the test or both.

Once again, the sample size made the interpretation of the fit statistics in the 'normal' way problematic.

– **Figure 6.1: Item characteristic curve for item 50 (B19 Erosion)**




Item 50 in Figure 6.1 is an example of the typical sort of Item Characteristic Curve (ICC) that one would expect from an extended response type item that is relatively difficult. The over-discrimination is probably due to the item type requiring the students to respond in writing to an open-ended type question. These questions measured something more than scientific reasoning: the ability to articulate an answer. The over-discrimination would probably have been more pronounced if the missing responses had been recorded as wrong and not missing.

The next section examines the differential item functioning of the items in the tests.

### 6.3.4 Differential item functioning

Differential item functioning analysis (DIF) was conducted on the data to assess whether there was any significant difference in the performance of the defined groups. It was performed by gender (boys: girls), LBOTE (English-speaking background: language background other than English) and ATSI (students of Aboriginal and/or Torres Strait Islander background: students of non-ATSI background).

Systematic differential item functioning on individual items would suggest that an item is measuring something different for that group and that this difference can be attributed to a curriculum difference.

To ascertain the performance of items in the study, a Rasch analysis was performed on the data from the 1,600-student calibration sample, using RUMM. An item analysis was undertaken for each item on the test.

Based on the findings from the Rasch and item analysis as noted earlier, Item 15 of Objective A was rescored such that response options A and B were considered valid responses.

To account for reverse thresholds, the response keys for Item 29 of Objective A and Item 11 of Objective B were collapsed.

Rasch analysis was undertaken again to derive item difficulties and person-abilities for the entire national year 6 sample.

In addition, a detailed analysis was undertaken to determine evidence of differential item functioning. The results provided some evidence of DIF for jurisdictions on certain items. Given that the large and unequal sample size may have contributed to the DIF reported, further analyses were performed using an equal number of students from each jurisdiction. Specifically, SPSS software was used to select a random sample of 4,000 (500 per jurisdiction) and 1,600 (200 per jurisdiction) from the national year 6 sample.

A review of the DIF analyses by jurisdiction, gender, language and Aboriginality indicated that none of the items was problematic.

After consultation with BEMU representatives, it was decided not to undertake supplementary analysis or to make any adjustments to the calibrations and student ability estimates to account for DIF, as it was considered that the factors contributing to DIF were artefacts of curriculum coverage and/or student learning in various jurisdictions and, as such, should be reported globally.

## 6.4 Traditional item analyses

Item analysis was performed on each sub-test (AA, AB, BA, BB) separately to determine the traditional item parameters of facility rate, facility by group, point biserial correlation, percentage missing and reliability estimates.

These initial analyses were used to check keys and examine any item characteristics that would inform the performance of the items. These parameters are included at Appendix 3, Item Level Statistics.

## 6.5 Student statistics

As indicated earlier, 15,158 students participated in the assessment. However, data from 592 students were excluded from the analyses because the students were enrolled in year levels other than the year 6 target group, and data from a further 392 students were omitted, as their demographic profile duplicated those of students in another sampled school. The resultant national sample comprised 14,172 year 6 students.

A primary aim of the study was to compare student performance across States and Territories and between the individual difference variables of gender, first language and Aboriginality. Difference by gender was the only variable to be reported at the jurisdiction level.

Hence, the results based on the national sample had to yield a sample providing accurately-weighted estimates of population parameters for which estimates of sampling variance could be made. The final phase of the analyses involved the application of weighting factors to develop the required national and State and Territory statistics.

Ability estimates were calculated in order to derive the final proficiency levels and standards. The ability parameters were derived from a sample of 1,600 students, comprising 200 students from each jurisdiction, selected at random from the national sample of 14,172 (all year 6) cases.

In the calibration of the ability estimates, items not attempted in each test were considered wrong until the student abandoned the test, the last item attempted being coded as wrong and all subsequent items coded as missing.

Data for all 14,172 students were included in the analysis. No data were omitted on the basis of poor fit or extreme scores. Zero scores were included in the analysis.

# CHAPTER 7: EQUATING AND SCALING

## 7.1 Analysis of the performance of the common items

The equating of common items was performed programmatically by concurrent analysis of all of the items using RUMM 2020.

A preliminary analysis of each of the Objective Forms was performed to test the quality of the links between the two test forms.

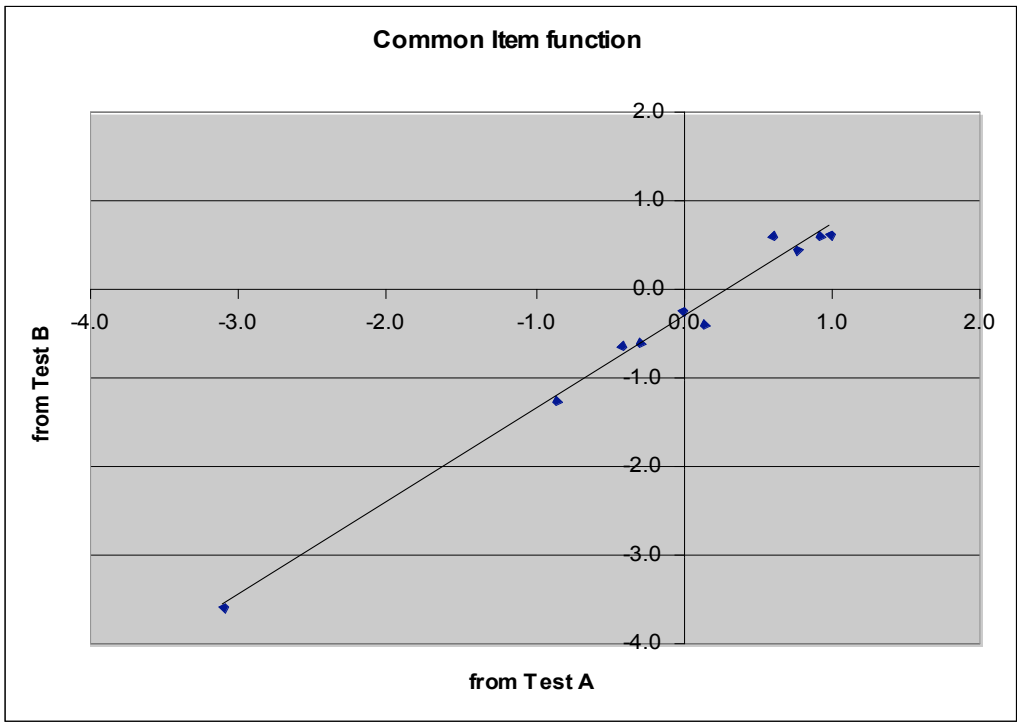
Figure 7.1 shows the relative locations of the items in each form.

It was considered that the items had functioned acceptably within each test to allow all of the common items to be included in the final calibration of the item locations for the concurrent analysis.

– **Table 7.1: Locations of common items**

| Item code | Form A | SE a | Form B | SE b |
|---|---|---|---|---|
| a03b01 | -0.864 | 0.046 | -1.237 | 0.046 |
| a04b02 | 0.130 | 0.037 | -0.381 | 0.039 |
| a05b03 | -0.308 | 0.040 | -0.595 | 0.040 |
| a12b04 | 0.987 | 0.036 | 0.638 | 0.037 |
| a13b05 | -0.417 | 0.041 | -0.625 | 0.040 |
| a14b21 | -3.102 | 0.118 | -3.569 | 0.042 |
| a15b22 | 0.752 | 0.035 | 0.473 | 0.045 |
| a16b23 | -0.009 | 0.039 | -0.227 | 0.038 |
| a34b29 | 0.591 | 0.024 | 0.615 | 0.024 |
| a35b30 | 0.913 | 0.037 | 0.615 | 0.037 |

– **Figure 7.1: Scattergram of common item function**



## 7.2 Person-ability estimates

The test equating function of RUMM was used to generate raw score-to-ability tables for each combination of test forms.

As the feedback from classroom teachers and quality control monitors indicated that there was sufficient time for students to complete the assessment, missing data were considered as wrong in the calibration of person-ability estimates.

Table 7.2 provides the raw score-to-ability translations for each test.

– **Table 7.2: Test raw score-to-ability translation (Logits P = 0.50)**

| Raw score | Test AA | Test AB | Test BA | Test BB | Obj A only | Obj B only |
|---|---|---|---|---|---|---|
| Maximum possible score | 43 | 44 | 45 | 46 | 39 | 41 |
| 0 | -5.07 | -5.10 | -4.87 | -4.90 | -5.06 | -4.85 |
| 1 | -4.22 | -4.25 | -4.00 | -4.03 | -4.20 | -3.98 |
| 2 | -3.62 | -3.65 | -3.39 | -3.43 | -3.60 | -3.37 |
| 3 | -3.20 | -3.24 | -2.97 | -3.01 | -3.18 | -2.94 |
| 4 | -2.87 | -2.91 | -2.64 | -2.69 | -2.85 | -2.60 |
| 5 | -2.60 | -2.64 | -2.36 | -2.41 | -2.57 | -2.32 |
| 6 | -2.36 | -2.40 | -2.12 | -2.18 | -2.32 | -2.08 |
| 7 | -2.14 | -2.19 | -1.91 | -1.97 | -2.10 | -1.86 |
| 8 | -1.94 | -1.99 | -1.72 | -1.77 | -1.90 | -1.67 |
| 9 | -1.76 | -1.82 | -1.54 | -1.60 | -1.72 | -1.48 |
| 10 | -1.59 | -1.65 | -1.37 | -1.43 | -1.54 | -1.31 |
| 11 | -1.44 | -1.49 | -1.22 | -1.28 | -1.38 | -1.15 |
| 12 | -1.28 | -1.34 | -1.07 | -1.13 | -1.22 | -1.00 |
| 13 | -1.14 | -1.20 | -0.93 | -1.00 | -1.07 | -0.85 |
| 14 | -1.00 | -1.06 | -0.80 | -0.86 | -0.93 | -0.71 |
| 15 | -0.87 | -0.93 | -0.67 | -0.73 | -0.78 | -0.57 |
| 16 | -0.74 | -0.80 | -0.54 | -0.61 | -0.65 | -0.44 |
| 17 | -0.61 | -0.68 | -0.42 | -0.49 | -0.51 | -0.31 |
| 18 | -0.48 | -0.55 | -0.30 | -0.37 | -0.38 | -0.19 |
| 19 | -0.36 | -0.43 | -0.19 | -0.26 | -0.25 | -0.07 |
| 20 | -0.24 | -0.31 | -0.08 | -0.15 | -0.12 | 0.05 |
| 21 | -0.12 | -0.19 | 0.04 | -0.04 | 0.01 | 0.17 |
| 22 | 0.00 | -0.08 | 0.15 | 0.07 | 0.15 | 0.29 |
| 23 | 0.12 | 0.04 | 0.25 | 0.18 | 0.28 | 0.41 |
| 24 | 0.24 | 0.16 | 0.36 | 0.29 | 0.41 | 0.53 |
| 25 | 0.36 | 0.27 | 0.47 | 0.40 | 0.54 | 0.65 |
| 26 | 0.48 | 0.39 | 0.58 | 0.50 | 0.68 | 0.77 |
| 27 | 0.60 | 0.51 | 0.69 | 0.61 | 0.82 | 0.90 |
| 28 | 0.72 | 0.63 | 0.80 | 0.72 | 0.97 | 1.02 |
| 29 | 0.85 | 0.75 | 0.91 | 0.83 | 1.13 | 1.16 |
| 30 | 0.98 | 0.88 | 1.03 | 0.94 | 1.29 | 1.29 |
| 31 | 1.12 | 1.01 | 1.14 | 1.05 | 1.46 | 1.44 |
| 32 | 1.26 | 1.14 | 1.27 | 1.17 | 1.65 | 1.59 |
| 33 | 1.41 | 1.28 | 1.39 | 1.29 | 1.86 | 1.75 |
| 34 | 1.57 | 1.43 | 1.52 | 1.41 | 2.10 | 1.93 |
| 35 | 1.74 | 1.59 | 1.66 | 1.54 | 2.37 | 2.12 |
| 36 | 1.92 | 1.76 | 1.81 | 1.68 | 2.70 | 2.34 |
| 37 | 2.13 | 1.94 | 1.97 | 1.83 | 3.11 | 2.59 |
| 38 | 2.36 | 2.14 | 2.14 | 1.98 | 3.71 | 2.90 |
| 39 | 2.63 | 2.37 | 2.33 | 2.16 | 4.57 | 3.29 |
| 40 | 2.95 | 2.63 | 2.55 | 2.34 | | 3.85 |
| 41 | 3.36 | 2.95 | 2.80 | 2.56 | | 4.67 |
| 42 | 3.94 | 3.35 | 3.10 | 2.80 | | |
| 43 | 4.79 | 3.93 | 3.49 | 3.10 | | |
| 44 | | 4.77 | 4.05 | 3.48 | | |
| 45 | | | 4.87 | 4.04 | | |
| 46 | | | | 4.86 | | |

### 7.2.1 Person-ability estimates and scaled scores

Individual student abilities were generated using the parameters of Table 7.2, taking into consideration the particular combination of tests undertaken by each student. The ability estimate of each student in the total sample was then standardised to provide a mean of zero and a standard deviation of 1 logit for the whole sample.

In the documentation, the parameter that relates to this ability is called ZRSA 1600 (standardised ability from raw score from 1600-sample calibration). This parameter estimate then allowed the generation of scaled scores with a nominal mean of 400 and a standard deviation of 100, which was computed by applying the linear equation:

**Scaled Score = Standardised Ability x 100 + 400**

## 7.2. Plausible values

The assessment instruments generated an item pool that was far too extensive to be administered in its entirety to any one student, so a test design was developed whereby each student was given a single test booklet containing a part of the entire assessment.

The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Each student responded to just a few items from each content area, and therefore multiple imputation, or 'plausible values', was used to derive reliable indicators of student proficiency.

The maximum number of score points provided by the assessment structure was 46 from the combination BB. When all possible combinations of the assessments were considered, there were 178 unique ability estimates generated by the assessment, as indicated in Table 6.6. If these values were to be extrapolated to the population by the weighting process described in Chapter 5, the performance of 260,000 students would be summarised by just 176 possible ability estimates.

Since every proficiency estimate incorporates some uncertainty, the analysis of the national sample mirrored TIMSS by following the customary procedure of generating five estimates for each student and using the variability among them as a measure of this imputation uncertainty or error. In the assessment report, the imputation error for each variable has been combined with the sampling error for that variable to provide a standard error incorporating both. The process is further described in Chapter 12 of the *TIMSS 1999 Technical Report* (2000).

The five ability estimates are derived using ConQuest. Pre-determined cut-scores define levels of achievement (See Chapter 8). For each set of plausible values generated, a count of students achieving each level is computed. An algorithm is used to estimate the proportions of students within each set of plausible values that have achieved the level and the parameters of standard error and standard deviation calculated. The mean of these calculations is then reported for each level.

When the five plausible values have been computed for a student, they are used for calculations of all achievement levels for each group to which the student belongs (gender, ATSI, LBOTE, geolocation, etc).

# CHAPTER 8: STANDARDS OF STUDENT ACHIEVEMENT

## 8.1 Scaling and standards-setting procedures

The standard for year 6 science was endorsed by the Key Performance Measures sub-group of the PMRT. This chapter describes the process by which that standard was set.

The process for setting standards in areas such as primary science, information and communications technologies, civics and citizenship and secondary (15-year-old) reading, mathematics and science was endorsed by the PMRT at its 6 March 2003 meeting and is described in the PMRT paper, *Setting National Standards*.

This process, referred to as the 'empirical judgemental technique', requires stakeholders to examine the test items and the results from the national assessments and agree on a proficient level of performance.

PMRT members were invited to nominate up to two representatives to participate in a standard-setting workshop on 23 September 2004.

### 8.1.1 Standards-setting process

The standards-setting process required expert judges to first identify and discuss factors that influenced the difficulty of the items in addition to science skills and understandings by examining test items from other standardised primary school science assessments.

The expert judges then had to decide independently whether a marginally-proficient year 6 student would be expected to answer each of the questions from the national assessment correctly. The term 'marginally' was added to 'proficient' to focus judges' attention on the lower end of the 'proficient' range, rather than on exemplary performances. Conceptually, this matched with the lower end of the proficiency bands in the report.

The results from the rating session, which showed the percentage of judges who expected marginally-proficient students to answer each question correctly, were summarised and returned to the judges. The results were rearranged in order of test item difficulty (as calculated from the national assessment) so the judges could determine whether there was a point on the science literacy scale where the group's rating decreased and signalled the location of the proficient standard cut-point.

Judges were then requested to work in groups of three to identify a question or small group of questions that best represented the most difficult items that a marginally-proficient student could be expected to answer correctly.

In coming to a decision, judges were expected to use the national test data, their initial ratings and the summary ratings for the group. The information from judges would locate the base of the 'proficient' band in the draft assessment; that is, the cut-point for the year 6 standard.

To conclude the standards-setting process, judges were required to identify independently and record the most difficult items that a marginally-proficient student would be expected to answer correctly, and give reasons for their decisions. These

results were collated by BEMU and formed the basis for the standard adopted for the project.

### 8.1.2 Outcomes of the standards-setting process

Eighteen of 21 expert judges identified the same small group of items on the Science literacy scale (Items 26-34 from the difficulty order item bank).

The location of this item was then defined as the standard for minimal proficiency at the p = 0.50 probability level. This cut-score was -0.13 on the initial calibration of the items shown in Table 6.1. This defined the lower lever of the proficiency band for Level 3.1 in Table 8.1 below.

The width of the band between proficiency levels was negotiated between BEMU and an expert panel of psychometricians and set at 1.25 logits (125 scaled points). This panel also recommended that in order for a student to have demonstrated proficiency at a particular level, he or she should have a probability of 65 per cent of achieving that level.

In order to adjust for this change in probability (p = 0.65) a constant of 0.62 (two decimal places) logits was added to the locations of the cut-scores determined in the standards-setting process, while the student ability estimates were maintained at their p = 0.50 calibrated levels. Hence the cut-score for proficiency at Level 3.1 became 0.45 logits.

The recalibrated cut-scores to provide for the p = 0.65 probability level are reflected in Table 8.1. The scaled scores have been calculated using the linear equation:

**Scaled Item Location = Adjusted Difficulty $_{(p=0.65)}$ x 100 + 400**

## 8.2 Proficiency levels

One of the key objectives of the MCEETYA National Assessment Program is to monitor trends in scientific literacy performance over time. One convenient and informative way of describing student performance over time is to reference the results to proficiency levels.

Students whose results are located within a particular level of proficiency are typically able to demonstrate the understandings and skills associated with that level, and also typically possess the understandings and skills defined as applying at lower proficiency levels.

To establish the proficiency levels, a combination of experts' knowledge of the skills required to answer each scientific literacy item and information from the analysis of students' responses was utilised.

Initially, three proficiency levels were identified, to correspond with Levels 2, 3 and 4 of the assessment domain (see Appendix 1, MCEETYA Scientific Literacy Framework). However, as 90 per cent of students' scores fell in Level 3, three further proficiency levels within Level 3 were created. This defined five proficiency levels for reporting student performances from the assessment.

– **Table 8.1: Student proficiency by bands (P = 0.65)**

| Level | Approximate percentage of students in the band | Cut-point (logits) |
|---|---|---|
| L 4 and above | 0.10 | |
| L 3.3 | 7.60 | 2.95 |
| L 3.2 (proficient) | 50.5 | 1.70 |
| L 3.1 | 37.2 | 0.45 |
| L 2 and below | 4.60 | -0.80 |

These cut-scores were used, in conjunction with the plausible values derived for each student (p = 0.5) described in Section 7.2.2, to determine the proportion of students who fell within each proficiency band.

### 8.2.1 Describing proficiency levels

Appendix 4, Descriptors of proficiency levels, provides the descriptions of the knowledge and skills required of students at each proficiency level. The descriptions reflect the skills assessed by the full range of scientific literacy items, including the three domains of scientific literacy.

It can be seen from Appendix 4 that the descriptors come from the scientific literacy assessment domain presented at Appendix 1, MCEETYA Scientific Literacy Framework. However, descriptions have only been provided in the assessment domain for Levels 1 to 5. There has been no description provided in the table for the sub-levels in Level 3, i.e. Levels 3.1 and 3.2.

### 8.2.2 Distribution of students across proficiency levels

The distributions of students within proficiency levels are shown in Table 8.2.

At the national level, approximately 4.6 per cent (SE = 0.2 per cent) of students performed at Proficiency Level 2 and below.

When the information from the judges was translated into a cut-point to represent the standard for year 6 science (and the lower end of the 'proficient' band), approximately 59.4 per cent of students achieved the standard. The corresponding figures for the proficiency bands are shown in Table 8.1.

Table 8.2 shows the percentage of students in each of the jurisdictions at Proficiency Levels 3.1, 3.2, 3.3 and the highest and lowest Proficiency Levels. It also shows in brackets the 95 per cent confidence interval about the mean estimates for each Proficiency Level. This has been calculated using the formula:

**95% confidence interval = 1.96*standard error.**

– **Table 8.2: Percentage of students, by proficiency band, by State and Territory**

| State/ Territory/ 95% confidence interval (+/-) | Level 2 and below | Level 3.1 | Level 3.2 | Level 3.3 | Level 4 and above |
|---|---|---|---|---|---|
| NSW | 3.4 | 33.7 | 52.6 | 10.1 | 0.1 |
| 95% CI (+/-) | 0.8 | 2.1 | 2.4 | 1.6 | 0.2 |
| VIC | 4.4 | 36.9 | 52.3 | 6.3 | 0.0 |
| 95% CI (+/-) | 1.0 | 2.7 | 2.7 | 1.2 | 0.1 |
| QLD | 5.1 | 40.0 | 49.0 | 5.8 | 0.0 |
| 95% CI (+/-) | 0.9 | 2.2 | 2.0 | 1.1 | 0.0 |
| SA | 4.4 | 38.6 | 50.1 | 6.8 | 0.0 |
| 95% CI (+/-) | 1.2 | 2.5 | 2.3 | 1.3 | 0.1 |
| WA | 5.1 | 40.3 | 48.7 | 5.9 | 0.0 |
| 95% CI (+/-) | 1.0 | 2.2 | 2.3 | 1.2 | 0.0 |
| TAS | 5.0 | 35.7 | 49.9 | 9.3 | 0.1 |
| 95% CI (+/-) | 1.4 | 2.9 | 2.9 | 1.8 | 0.3 |
| NT | 10.7 | 39.9 | 42.5 | 6.9 | 0.0 |
| 95% CI (+/-) | 3.6 | 5.6 | 4.8 | 2.8 | 0.0 |
| ACT | 2.7 | 27.5 | 56.1 | 13.3 | 0.2 |
| 95% CI (+/-) | 1.1 | 3.9 | 4.8 | 2.7 | 0.5 |
| **All** | **4.6** | **37.2** | **50.5** | **7.6** | **0.1** |
| **95% CI (+/-)** | **0.4** | **0.9** | **0.9** | **0.5** | **0.1** |

The assessment instruments were originally constructed with the expectation that most students in year 6 would demonstrate the understanding and skills of Proficiency Level 3. The results of Table 8.3 show that approximately three out of five Australian year 6 students are proficient at Level 3.2 and above.

– **Table 8.3: Percentage of students achieving defined proficiency level, by State and Territory**

| State/ Territory/ 95% confidence interval (+/-) | Non-proficient | Proficient |
|---|---|---|
| NSW | 37.2 | 62.8 |
| 95% CI (+/-) | 2.1 | 2.1 |
| VIC | 41.3 | 58.7 |
| 95% CI (+/-) | 2.5 | 2.5 |
| QLD | 45.1 | 54.9 |
| 95% CI (+/-) | 2.1 | 2.1 |
| SA | 43.0 | 57.0 |
| 95% CI (+/-) | 2.4 | 2.4 |
| WA | 45.4 | 54.6 |
| 95% CI (+/-) | 2.5 | 2.2 |
| TAS | 40.7 | 59.3 |
| 95% CI (+/-) | 2.9 | 2.9 |
| NT | 50.6 | 49.4 |
| 95% CI (+/-) | 5.5 | 5.5 |
| ACT | 30.2 | 69.8 |
| 95% CI (+/-) | 3.6 | 3.6 |
| **All** | **41.8** | **58.2** |
| **95% CI (+/-)** | **0.9** | **0.9** |

Table 8.4 shows the performance of groups in the assessment, revealing that there is little difference between the performance of boys and girls at each of the proficiency levels.

It can be seen that the performance of students from English-speaking backgrounds is higher than that of students with ATSI backgrounds and those from language backgrounds other than English.

Thirty per cent of ATSI students and 48.1 per cent of LBOTE students performed at Level 3.2 or above, compared with 59.6 per cent of non-ATSI students from English-speaking backgrounds.

– **Table 8.4: Percentage of students, by proficiency band, by group**

| Group 95% confidence interval (+/-) | Level 2 and below | Level 3.1 | Level 3.2 | Level 3.3 | Level 4 and above |
|---|---|---|---|---|---|
| Females | 4.8 | 37.8 | 50.5 | 6.9 | 0.0 |
| 95% CI (+/-) | 0.6 | 1.2 | 1.2 | 0.6 | 0.1 |
| Males | 4.4 | 36.5 | 50.7 | 8.3 | 0.1 |
| 95% CI (+/-) | 0.6 | 1.3 | 1.3 | 0.7 | 0.1 |
| ATSI | 18.2 | 51.9 | 28.3 | 1.7 | 0.0 |
| 95% CI (+/-) | 3.9 | 5.9 | 4.3 | 1.3 | 0.0 |
| Non-ATSI | 4.0 | 36.3 | 51.7 | 7.9 | 0.1 |
| 95% CI (+/-) | 0.4 | 1.0 | 0.9 | 0.5 | 0.1 |
| ESB | 4.1 | 36.2 | 51.6 | 8.0 | 0.1 |
| 95% CI (+/-) | 0.4 | 1.0 | 1.0 | 0.7 | 0.1 |
| LBOTE | 7.9 | 44.0 | 43.6 | 4.5 | 0.0 |
| 95% CI (+/-) | 1.5 | 3.1 | 2.9 | 1.6 | 0.0 |
| **All** | **4.6** | **37.2** | **50.5** | **7.6** | **0.1** |
| **95% CI (+/-)** | **0.4** | **0.9** | **0.9** | **0.5** | **0.1** |

Table 8.5 compares the performance of students from the different geolocations, as defined by the MCEETYA Geographical Classification Location Framework (2001).

– **Table 8.5: Percentage of students, by proficiency band, by geolocation**

| Geolocation 95% confidence interval (+/-) | Level 2 and below | Level 3.1 | Level 3.2 | Level 3.3 | Level 4 and above |
|---|---|---|---|---|---|
| Mainland State capital city regions | 4.3 | 37.6 | 50.9 | 7.1 | 0.0 |
| 95% CI (+/-) | 0.6 | 1.5 | 1.4 | 0.9 | 0.1 |
| Major urban statistical districts | 3.8 | 33.4 | 52.4 | 10.3 | 0.1 |
| 95% CI (+/-) | 0.9 | 2.0 | 2.1 | 1.3 | 0.2 |
| Provincial city statistical districts | 5.5 | 39.1 | 48.4 | 7.0 | 0.1 |
| 95% CI (+/-) | 1.2 | 2.6 | 2.8 | 1.5 | 0.2 |
| Other regional areas | 4.8 | 37.8 | 50.5 | 7.0 | 0.0 |
| 95% CI (+/-) | 1.1 | 2.4 | 2.1 | 1.3 | 0.1 |
| Remote zone | 11.5 | 40.2 | 41.3 | 7.3 | 0.0 |
| 95% CI (+/-) | 3.6 | 5.8 | 5.5 | 3.1 | 0.0 |
| **All** | **4.6** | **37.2** | **50.5** | **7.6** | **0.1** |
| **95% CI (+/-)** | **0.4** | **0.9** | **0.9** | **0.5** | **0.1** |

## 8.3 Proficiency bands: related technical information

### 8.3.1 Distribution of students across proficiency levels

To facilitate the reporting of results, several of the technical standards from PISA have been adopted. Two of the key mathematically-linked standards are:

- setting the response probability for the analysis of data at p = 0.65; and
- setting the width of the proficiency bands at 1.25 logits.

These two technical standards flow from a consideration of the conceptual issues associated with reporting student performance against a standard and in terms of bands.

As a consequence of adopting these standards for the report, the following inferences can be made about students' proficiency in relation to the bands:

- A student whose result places him/her at the lowest possible point of the proficiency band is likely to get 50 per cent correct on a test made up of items spread uniformly across the band, from the easiest to the most difficult.

- A student whose result places him/her at the lowest possible point of the proficiency band is likely to get 65 per cent correct on a test made up of items similar to the easiest items in the band.

- A student at the top of the proficiency band is likely to get 85 per cent correct on a test made up of items similar to the easiest items in the band.

- A student whose result places him or her at the same point on the science literacy scale as the cut-point for the science standard is likely to get 65 per cent correct on a test made up of items similar to the items at the cut-point for the standard.

Clearly it is possible to change the two mathematically interrelated technical standards in order to vary the inferences about the likely percentage correct on tests. The position taken by PISA, and adopted by PMRT, attempts to balance the notions of mastery and 'pass' in a way that is likely to be understood by the community.

# CHAPTER 9: IMPLICATIONS FOR SCALING AND EQUATING THE NATIONAL YEAR 6 SCIENCE ASSESSMENT IN 2006

## 9.1 Horizontal equating: common items 2003 and 2006

With the exception of the common link items, Objective Form A and Practical Form A have been released into the public domain for school use.

Objective Form B and Practical Form B are secure forms that may be used as an intact test, or drawn upon to enable the 2006 assessment to be calibrated on the 2003 scale.

## 9.2 Considerations regarding item fit 2006

Discussions regarding fit and DIF have been addressed in this report and in the main report, the *MCEETYA National Year 6 Science Assessment Report 2003*. These factors may require consideration in assessing the suitability of items for inclusion in the set of equating items for the links between the 2003 and 2006 assessments.

There is obviously an issue regarding the 'thickness' of the variable. Possible outcomes include:

- keeping the thicker variable and accepting the greater degree of misfit to the model;

- reconfiguring the scientific literacy variable so that at least two measures are presented for each student – one for the multiple-choice class of items and the other for the open-ended class of items. This would improve the fit of data to the model; or

- adjusting the model in some way to account for the systematic difference in the classes of items, before producing a single scientific reasoning score for each student.

Table 9.1 highlights those items that exhibited significant DIF among States and Territories. It is included as many of the items are present in the 'secure' test items (Objective B and Practical B) that may be used in the equating and calibration of the 2006 assessment.

– **Table 9.1: Items showing most significant DIF among States and Territories**

| Item number | Item name | Item type | Mean square | F statistic | DF | Prob |
|---|---|---|---|---|---|---|
| 3 | A03B01 Bar Magnets | Poly | 18.16 | 15.28 | 7 | 0.000 |
| 4 | A04B02 Bar Magnets | Poly | 13.39 | 11.76 | 7 | 0.000 |
| 44 | B13 Wheat Growth 3 | Poly | 10.54 | 13.44 | 7 | 0.000 |
| 50 | B19 Erosion 3 | Poly | 9.95 | 12.45 | 7 | 0.000 |
| 53 | B25 School Electricity 3 | Poly | 8.54 | 10.92 | 7 | 0.000 |
| 34 | A34B28 Paper Plane 1 | Poly | 7.76 | 6.94 | 7 | 0.000 |
| 60 | B34 Exercise 2 | MC | 6.01 | 7.58 | 7 | 0.000 |

## 9.3 Considerations regarding item type 2006

When marking the assessments, the markers commented on an apparently large number of students who did not respond to items requiring an extended answer to be supplied. Appendix 3, Item Level Statistics, contains a summary of the item level information on each of the items in the assessments.

– **Table 9.2: Percentages of students omitting responses, 2003**

| | | Item types and percentages omitted | | |
|---|---|---|---|---|
| State/Territory | Gender | Multiple-choice | Short-answer | Extended-response |
| NSW | Female | 3.4 | 2.7 | 3.7 |
| | Male | 2.1 | 3.4 | 4.5 |
| VIC | Female | 3.1 | 4.3 | 6.2 |
| | Male | 3.1 | 4.5 | 5.9 |
| QLD | Female | 4.2 | 4.8 | 6.8 |
| | Male | 3.6 | 5.6 | 7.1 |
| SA | Female | 3.2 | 4.7 | 6.4 |
| | Male | 3.7 | 6.3 | 7.0 |
| WA | Female | 3.1 | 4.7 | 6.4 |
| | Male | 3.3 | 4.9 | 6.7 |
| TAS | Female | 2.9 | 4.6 | 5.6 |
| | Males | 2.7 | 6.0 | 6.9 |
| NT | Female | 6.4 | 7.6 | 7.8 |
| | Male | 7.3 | 8.2 | 11.1 |
| ACT | Female | 3.5 | 4.4 | 5.9 |
| | Male | 2.8 | 3.5 | 5.0 |
| All | Female | 3.3 | 4.4 | 6.0 |
| | Male | 3.2 | 5.1 | 6.4 |

The trend in Table 9.2 reveals that in nearly all cases, the percentage of students omitting responses for extended-response items is double the percentage omitting responses for multiple-choice items. The percentage omitting responses in short answer items is generally higher than the percentage omitting responses for

multiple-choice items, but not as high as the percentage omitting responses for extended response items. There is no evidence that there is a gender effect with response type.

These results do suggest, however, that there is a systematic effect throughout the scientific literacy scale associated with the literacy demands of the extended response item types which may, in turn, affect the level of student engagement with the test items.

## 9.4 Considerations regarding item inclusion for equating 2006

Table 7.2 includes the raw score-to-ability translation for Objective Form B only.

Should issues that affect the assessment program result in only the secure objective form of the 2003 implementation being used for equating in 2006, these parameters will then be relevant.

# REFERENCES

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

Foy, Pierre (2000). *TIMSS 1999 Technical Report. Chapter 11.* Chestnut Hill, MA: International Study Center, Boston College.

Gregory, Kelvin G, Martin, Michael O. &, Stemler, Steven E. (eds) (2000). *TIMSS 1999 Technical Report.* Chestnut Hill, MA: International Study Center, Boston College.

Hackling, M.W., Goodrum, D. & Rennie, L. (2001). The state of science in Australian secondary schools. *Australian Science Teachers Journal*, 47(4), 6-17.

Johnson, E.G., & Rust, K.F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics*, *17*, 175-190, cited in Kelvin G. Gregory et al., *TIMMS 1999 Technical Report*. Chestnut Hill, MA: International Study Center, Boston College, p. 204.

Jones, R. (2004). *Geolocation Questions and Coding Index*. Performance and Reporting Taskforce: Ministerial Council on Education, Employment, Training and Youth Affairs.

Organisation for Economic Co-operation and Development (1999). Programme for International Student Asssessment. *Measuring student knowledge and skills: a new framework for assessment*. Paris: OECD.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag. cited in Kelvin G. Gregory et al., *TIMMS 1999 Technical Report*. Chestnut Hill, MA: International Study Center, Boston College, p. 204.

# APPENDIX 1

## MCEETYA Scientific Literacy Framework

## Scientific Literacy Assessment Domain

| Level | Scientific literacy | | |
|---|---|---|---|
| | **Strand A**<br>Formulating or identifying investigable questions and hypotheses, planning investigations and collecting evidence<br><br>Process Domain: experimental design and data gathering | **Strand B**<br>Interpreting evidence and drawing conclusions from their own or others' data, critiquing the trustworthiness of evidence and claims made by others, and communicating findings<br>Process Domain: interpreting experimental data | **Strand C**<br>Using understandings for describing and explaining natural phenomena, and for interpreting reports about phenomena<br>Conceptual Domain: applies conceptual understanding |
| **1** | Responds to the teacher's questions, observes and describes. | Describes what happened. Identifies one aspect of the data. | Describes (or recognises) one aspect or property of an individual object or event that has been experienced or reported<br>SOLO: Concrete unistructural |
| **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.<br><br>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.<br>SOLO: Concrete multistructural |
| **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. | 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.<br>SOLO: Concrete relational |
| **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 data. Able to make general suggestions for improving an investigation (eg. 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.<br>SOLO: Abstract unistructural |
| **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.<br>SOLO: Abstract multistructural |
| **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 (eg. 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.<br>SOLO: Abstract relational |

*The conceptual strand (C) has been abstracted across conceptual strands and makes no reference to particular concepts or contexts.*

# APPENDIX 2

## National Year 6 Science Assessment 2003: Session Report

### Appendix 2. PSAP SESSION REPORT FORM

1. **School Name:** _____
2. **School ID:** _____
3. **Test Administrator:** _____
4. **School Contact:** _____

### Session Information

4. **Date of Testing:** _____/_____/_____
   Day Month Year
5. **Scheduled Start Time:** _____:_____
   24:00

### Session Timing

| | | Start | End | Not Applicable |
|---|---|---|---|---|
| 6. | Introduction to the Assessment (Preparation of Students, Instructions, Materials Distribution) | ___:___ 24:00 | ___:___ 24:00 | ☐ |
| 7. | Objective Assessment Part 1 (60 Minutes) | ___:___ 24:00 | ___:___ 24:00 | ☐ |
| 8. | Practical Assessment Part 2 (45 Minutes)<br>Date of Testing: _____/_____/_____ | ___:___ 24:00 | ___:___ 24:00 | ☐ |

9. **Were any of the following present during the testing session?**

| | | Yes | No |
|---|---|---|---|
| a) | PSAP Quality Monitor | ☐1 | ☐2 |
| b) | Other_____ (please specify) | ☐1 | ☐2 |

### Disruptions

10. **Did any of the following affect the test session?**

| | | Yes | No |
|---|---|---|---|
| a) | Announcements over the loudspeaker / Alarms | ☐1 | ☐2 |
| b) | Class changeover in the school | ☐1 | ☐2 |
| c) | Other students not participating in the test session | ☐1 | ☐2 |
| d) | Students or Teachers visiting the testing room | ☐1 | ☐2 |

## Assessment Booklet Format and Content

**11. Were there any problems with the Assessment Booklets (e.g. errors or omissions, unclear directions, confusing format, too long, too hard, boring, tiring etc.)?**

☐ No ☐ Yes. Specify

_____
_____
_____
_____
_____
_____

**12. Were there any problems with specific test items?**

☐ No ☐ Yes. Specify (include booklet number and item number):

| BOOK# | ITEM# | PROBLEM |
|---|---|---|
| _____ | _____ | _____ |
| _____ | _____ | _____ |
| _____ | _____ | _____ |
| _____ | _____ | _____ |
| _____ | _____ | _____ |
| _____ | _____ | _____ |

**13. Please note other comments that you think would help improve the assessment:**

_____
_____
_____
_____
_____

THANK YOU VERY MUCH

# APPENDIX 3

## Item Level Statistics

**National Year 6 Science sample assessment 2003**

**Click here to download Excel version of the following table:**

## Appendix 3 ~ Item Level Statistics

**National Year 6 Science sample assessment 2003**

| ItemID | Form | Question | Qcode | Type | Keys | % CORRECT | % OMIT | % CORRECT - Males | % score 2 - Males | % CORRECT Females | % Score 2 - Females | % CORRECT - INDG - Yes | % score 2 - INDG - Yes | % CORRECT - INDG - No | % score 2 - INDG - No | % CORRECT - ESL - Yes | Score 2 - ESL - Yes | % CORRECT - LBOTE | Score 2 - LBOTE | ItmCode | Location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I0001 | A | ice cubes | a1ic1r | SA | 0,1,9 | 93.76 | 0.98 | 92.70 | | 95.00 | | 92.10 | | 93.90 | | 93.90 | | 92.60 | | a0001 | -2.287 |
| I0002 | A | ice cubes | a2ic2r | SA | 0,1,9 | 66.64 | 1.96 | 66.50 | | 66.80 | | 51.70 | | 67.50 | | 67.30 | | 62.20 | | a0002 | -0.124 |
| I0003 | COMMON | bar magnets | a3b1bm1r | MC | C | 83.21 | 0.83 | 83.30 | | 83.26 | | 77.23 | | 83.53 | | 83.33 | | 82.55 | | a0003 | -1.060 |
| I0004 | COMMON | bar magnets | a4b2bm2r | SA | 0,1,9 | 67.48 | 1.77 | 70.09 | | 64.79 | | 59.36 | | 67.83 | | 67.73 | | 65.88 | | a0004 | -0.158 |
| I0005 | COMMON | bar magnets | a5b3bm3r | SA | 0,1,9 | 73.41 | 1.23 | 78.26 | | 68.28 | | 61.38 | | 73.97 | | 73.76 | | 71.24 | | a0005 | -0.463 |
| I0006 | A | floors | a6fl1r | MC | A | 86.76 | 0.90 | 85.90 | | 87.80 | | 77.80 | | 87.30 | | 87.40 | | 82.40 | | a0006 | -1.318 |
| I0007 | A | floors | a7fl2r | SA | 0,1,9 | 87.10 | 1.35 | 86.10 | | 88.30 | | 77.80 | | 87.70 | | 87.50 | | 84.70 | | a0007 | -1.295 |
| I0008 | A | floors | a8fl3r | MC | A | 88.75 | 0.84 | 89.70 | | 87.70 | | 82.50 | | 89.10 | | 88.90 | | 87.80 | | a0008 | -1.608 |
| I0009 | A | echidna | a9e1r | MC | B | 82.51 | 0.92 | 80.50 | | 84.70 | | 77.80 | | 82.80 | | 83.10 | | 78.10 | | a0009 | -0.925 |
| I0010 | A | echidna | a10e2r | MC | C | 93.66 | 1.15 | 93.80 | | 93.60 | | 88.10 | | 93.90 | | 94.20 | | 89.50 | | a0010 | -2.607 |
| I0011 | A | echidna | a11e3r | MC | B | 58.76 | 0.99 | 60.50 | | 56.90 | | 42.70 | | 59.50 | | 58.90 | | 58.20 | | a0011 | 0.344 |
| I0012 | COMMON | planets | a12b4p1r | SA | 0,1,9 | 45.92 | 3.44 | 45.93 | | 45.99 | | 31.70 | | 46.57 | | 46.33 | | 43.56 | | a0012 | 0.809 |
| I0013 | COMMON | planets | a13b5p2r | MC | A | 75.01 | 0.94 | 75.24 | | 74.75 | | 68.13 | | 75.35 | | 75.57 | | 71.00 | | a0013 | -0.610 |
| I0014 | COMMON | camping holiday | a14b20ch1r | MC | D | 96.29 | 1.18 | 95.93 | | 96.70 | | 89.54 | | 96.66 | | 96.60 | | 94.34 | | a0014 | -3.155 |
| I0015 | COMMON | camping holiday | a15b21ch2r | MC | A,B | 90.42 | 1.31 | 90.15 | | 90.76 | | 80.94 | | 90.88 | | 90.91 | | 87.18 | | a0015 | -1.990 |
| I0016 | COMMON | camping holiday | a16b22ch3r | SA | 0,1,9 | 63.20 | 7.06 | 64.02 | | 62.43 | | 51.94 | | 63.85 | | 64.28 | | 55.29 | | a0016 | -0.030 |
| I0017 | A | mosquito | a17m1r | SA | 0,1,9 | 67.23 | 2.83 | 64.50 | | 70.30 | | 50.70 | | 68.30 | | 67.50 | | 65.60 | | a0017 | -0.167 |
| I0018 | A | mosquito | a18m2r | SA | 0,1,9 | 45.80 | 5.26 | 48.10 | | 43.50 | | 27.50 | | 46.80 | | 46.20 | | 42.90 | | a0018 | 0.943 |
| I0019 | A | mosquito | a19m3r | SA | 0,1,9 | 62.24 | 9.04 | 62.20 | | 62.40 | | 45.70 | | 63.10 | | 62.70 | | 59.20 | | a0019 | 0.091 |
| I0020 | A | mosquito | a20m4r | SA | 0,1,2,9 | 24.24 | 12.50 | 34.81 | 6.54 | 38.11 | 5.55 | 19.21 | 2.65 | 37.29 | 6.25 | 36.47 | 6.28 | 35.03 | 4.73 | a0020 | 2.064 |
| I0021 | A | paper clip | a21pc1r | MC | B | 82.57 | 1.28 | 83.30 | | 81.90 | | 75.20 | | 83.00 | | 83.20 | | 78.10 | | a0021 | -1.198 |
| I0022 | A | paper clip | a22pc2r | SA | 0,1,9 | 41.61 | 3.98 | 42.70 | | 40.60 | | 27.20 | | 42.50 | | 42.90 | | 31.90 | | a0022 | 1.088 |
| I0023 | A | bush pond | a23bu1r | MC | B | 58.02 | 3.89 | 58.00 | | 58.20 | | 45.00 | | 58.70 | | 58.60 | | 54.20 | | a0023 | 0.413 |
| I0024 | A | bush pond | a24bu2r | SA | 0,1,9 | 72.84 | 8.52 | 74.90 | | 70.70 | | 65.60 | | 73.30 | | 73.20 | | 70.20 | | a0024 | -0.541 |
| I0025 | A | bush pond | a25bu3r | SA | 0,1,2,9 | 50.67 | 7.36 | 59.20 | 21.01 | 60.15 | 20.72 | 47.02 | 21.52 | 60.33 | 20.85 | 59.98 | 20.71 | 57.58 | 21.82 | a0025 | 0.524 |
| I0026 | A | wood burning | a26wb1r | SA | 0,1,2,9 | 66.79 | 3.99 | 42.02 | 44.90 | 37.17 | 49.22 | 40.40 | 33.11 | 39.44 | 47.76 | 39.89 | 47.57 | 37.21 | 43.39 | a0026 | -0.027 |
| I0027 | A | wood burning | a27wb2r | MC | A | 76.84 | 3.22 | 78.20 | | 75.50 | | 63.60 | | 77.50 | | 78.00 | | 68.40 | | a0027 | -0.814 |
| I0028 | A | sandpaper | a28sp1r | SA | 0,1,9 | 88.08 | 3.82 | 88.40 | | 87.80 | | 75.20 | | 88.60 | | 88.40 | | 85.60 | | a0028 | -1.786 |
| I0029 | A | sandpaper | a29sp2r | SA | 0,1,9 © | 76.60 | 7.61 | 74.30 | | 79.30 | | 60.60 | | 77.60 | | 77.60 | | 69.70 | | a0029 | -0.684 |
| I0030 | A | toy train | a30tt1r | SA | 0,1,9 | 42.42 | 7.60 | 44.80 | | 40.10 | | 26.50 | | 43.20 | | 43.00 | | 38.70 | | a0030 | 1.054 |
| I0031 | A | toy train | a31tt2r | SA | 0,1,9 | 66.71 | 7.42 | 67.20 | | 66.30 | | 56.60 | | 67.20 | | 67.50 | | 61.10 | | a0031 | -0.133 |
| I0032 | A | toy train | a32tt3r | SA | 0,1,9 | 34.49 | 12.70 | 37.50 | | 31.30 | | 20.50 | | 35.20 | | 35.40 | | 28.20 | | a0032 | 1.263 |
| I0033 | A | toy train | a33tt4r | SA | 0,1,9 | 27.22 | 23.26 | 33.40 | | 20.70 | | 12.30 | | 28.00 | | 27.80 | | 23.40 | | a0033 | 1.592 |
| I0034 | COMMON | paper plane | a34b28pp1r | SA | 0,1,2,9 | 47.27 | 10.35 | 26.59 | 33.12 | 26.34 | 35.04 | 25.63 | 21.92 | 26.41 | 34.67 | 26.52 | 34.68 | 26.11 | 30.26 | a0034 | 0.651 |
| I0035 | COMMON | paper plane | a35b28pp2r | MC | B | 44.71 | 8.98 | 44.05 | | 45.41 | | 29.17 | | 45.48 | | 45.28 | | 41.34 | | a0035 | 0.692 |
| I0036 | B | colour fading | b6cf1 | MC | B | 75.12 | 1.16 | 73.30 | | 77.10 | | 63.90 | | 75.60 | | 76.10 | | 68.20 | | b0006 | -0.664 |
| I0037 | B | colour fading | b7cf2 | SA | 0,1,9 | 78.48 | 3.08 | 76.50 | | 80.50 | | 67.70 | | 79.20 | | 79.20 | | 74.00 | | b0007 | -0.974 |
| I0038 | B | colour fading | b8cf3 | MC | C | 64.26 | 1.23 | 65.20 | | 63.30 | | 56.00 | | 64.70 | | 64.20 | | 64.50 | | b0008 | -0.028 |
| I0039 | B | cave diggers | b9cd1 | SA | 0,1,9 | 42.58 | 2.07 | 41.30 | | 43.80 | | 30.20 | | 43.30 | | 43.00 | | 40.50 | | b0009 | 0.970 |
| I0040 | B | cave diggers | b10cd2 | MC | A | 47.13 | 1.01 | 48.10 | | 46.30 | | 40.20 | | 47.60 | | 47.80 | | 43.50 | | b0010 | 0.704 |
| I0041 | B | cave diggers | b11cd3 | SA | 0,1, 9, © | 86.46 | 2.65 | 88.00 | | 85.00 | | 76.60 | | 86.90 | | **87.00** | | **83.40** | | b0011 | -1.556 |
| I0042 | B | wheat growth | b12aw1 | SA | 0,1,9 | 54.13 | 1.72 | 51.60 | | 56.90 | | 38.10 | | 54.90 | | 54.90 | | 49.20 | | b0012 | 0.431 |
| I0043 | B | wheat growth | b12bw2 | SA | 0,1,9 | 52.52 | 2.45 | 49.60 | | 55.80 | | 40.90 | | 53.10 | | 52.80 | | 51.00 | | b0013 | 0.467 |
| I0044 | B | wheat growth | b13w3 | SA | 0,1,9 | 79.08 | 3.13 | 80.20 | | 77.90 | | 65.30 | | 79.90 | | 80.00 | | 73.40 | | b0014 | -1.027 |
| I0045 | B | wheat growth | b14w4 | SA | 0,1,9 | 56.32 | 4.65 | 58.10 | | 54.40 | | 44.00 | | 57.10 | | 57.00 | | 51.30 | | b0015 | 0.181 |
| I0046 | B | shadows | b15sh1 | SA | 0,1,9 | 33.84 | 1.72 | 39.10 | | 28.40 | | 24.10 | | 34.30 | | 34.30 | | 31.20 | | b0016 | 1.354 |
| I0047 | B | shadows | b16sh2 | MC | C | 89.26 | 1.33 | 90.50 | | 88.00 | | 80.10 | | 89.90 | | 89.70 | | 87.30 | | b0017 | -1.886 |
| I0048 | B | erosion | b17er1 | SA | 0,1,2, 9 | 46.08 | 5.02 | 42.77 | 25.89 | 42.66 | 23.50 | 40.21 | 14.78 | 42.84 | 25.23 | 42.85 | 25.21 | 41.70 | 21.39 | b0018 | 0.749 |
| I0049 | B | erosion | b18er2 | MC | D | 88.38 | 1.74 | 89.00 | | 87.60 | | 76.60 | | 88.90 | | 89.00 | | 84.30 | | b0019 | -1.738 |
| I0050 | B | erosion | b19er3 | SA | 0,1,9 | 37.22 | 12.29 | 38.70 | | 35.70 | | 18.20 | | 38.30 | | 38.50 | | 28.40 | | b0020 | 1.191 |

| ItemID | Form | Question | Qcode | Type | Keys | % CORRECT | % OMIT | % CORRECT - Males | % score 2 - Males | % CORRECT Females | % Score 2 - Females | % CORRECT - INDG - Yes | % score 2 - INDG - Yes | % CORRECT - INDG - No | % score 2 - INDG - No | % CORRECT - ESL - Yes | Score 2 - ESL - Yes | % CORRECT - LBOTE | Score 2 - LBOTE | ItmCode | Location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I0051 | B | school electricity | b23se1 | MC | B | 46.01 | 2.89 | 47.50 | | 44.50 | | 40.20 | | 46.30 | | 46.30 | | 43.60 | | b0024 | 0.692 |
| I0052 | B | school electricity | b24se2 | MC | C | 52.11 | 3.34 | 52.30 | | 51.90 | | 44.00 | | 52.50 | | 53.00 | | 45.90 | | b0025 | 0.452 |
| I0053 | B | school electricity | b25se3 | SA | 0,1,9 | 24.86 | 14.23 | 26.50 | | 23.10 | | 13.70 | | 25.50 | | 25.40 | | 20.90 | | b0026 | 1.927 |
| I0054 | B | bean plant | b26bp1 | SA | 0,1,2, 9 | 60.43 | 4.96 | 57.45 | 32.00 | 57.81 | 31.22 | 57.39 | 19.59 | 57.55 | 32.27 | 57.63 | 32.16 | 57.35 | 28.20 | b0027 | -0.159 |
| I0055 | B | bean plant | b27bp2 | SA | 0,1,9 | 33.03 | 10.02 | 32.80 | | 33.30 | | 21.00 | | 33.70 | | 34.00 | | 26.60 | | b0028 | 1.481 |
| I0056 | B | curtains | b30c1 | SA | 0,1,2, 9 | 33.02 | 14.40 | 47.61 | 8.74 | 47.03 | 9.99 | 34.02 | 3.44 | 47.93 | 9.67 | 48.20 | 9.35 | 41.46 | 9.68 | b0031 | 1.490 |
| I0057 | B | curtains | b31c2 | MC | B | 64.99 | 7.99 | 62.10 | | 68.10 | | 50.90 | | 65.70 | | 65.90 | | 59.00 | | b0032 | -0.391 |
| I0058 | B | curtains | b32c3 | MC | A | 48.20 | 8.52 | 48.80 | | 47.70 | | 33.70 | | 48.90 | | 49.40 | | 40.30 | | b0033 | 0.528 |
| I0059 | B | exercise | b33ex1 | SA | 0,1,2, 9 | 40.22 | 12.42 | 43.73 | 18.97 | 47.96 | 15.58 | 36.77 | 11.00 | 46.09 | 17.74 | 46.43 | 17.63 | 41.46 | 15.29 | b0034 | 0.833 |
| I0060 | B | exercise | b34ex2 | MC | C | 69.54 | 11.30 | 67.40 | | 71.70 | | 49.10 | | 70.50 | | 71.00 | | 59.40 | | b0035 | -0.958 |
| I0061 | B | exercise | b35ex3 | MC | B | 60.83 | 12.13 | 61.10 | | 60.40 | | 41.90 | | 61.80 | | 62.40 | | 50.30 | | b0036 | -0.279 |
| I0062 | OA | craters | ap01_1r | SA | 0,1,9 | 58.82 | 2.31 | 54.58 | | 63.30 | | 43.48 | | 59.63 | | 59.11 | | 56.78 | | ap001 | 0.231 |
| I0063 | OA | craters | ap02_2r | SA | 0,1,9 | 49.44 | 4.33 | 48.14 | | 50.82 | | 36.79 | | 50.17 | | 49.98 | | 46.00 | | ap002 | 0.812 |
| I0064 | OA | craters | ap03_3r | SA | 0,1,9 | 44.52 | 3.94 | 42.81 | | 46.35 | | 25.08 | | 45.53 | | 45.66 | | 36.08 | | ap003 | 0.866 |
| I0065 | OA | craters | ap04_4r | SA | 0,1,9 | 18.59 | 5.12 | 22.87 | | 14.11 | | 11.04 | | 19.01 | | 18.99 | | 15.74 | | ap004 | 2.302 |
| I0066 | OB | parachute | bp01_1 | SA | 0,1,9 | 40.21 | 2.41 | 46.72 | | 33.41 | | 25.51 | | 40.94 | | 40.83 | | 36.84 | | bp001 | 1.166 |
| I0067 | OB | parachute | bp02_2 | SA | 0,1,9 | 81.88 | 2.54 | 86.11 | | 77.41 | | 71.09 | | 82.36 | | 82.40 | | 79.19 | | bp002 | -1.085 |
| I0068 | OB | parachute | bp03_3 | SA | 0,1,9 | 24.29 | 3.39 | 22.84 | | 25.84 | | 14.97 | | 24.81 | | 24.74 | | 20.57 | | bp003 | 1.909 |
| I0069 | OB | parachute | bp04_4 | SA | 0,1,9 | 38.37 | 4.08 | 40.12 | | 36.55 | | 20.75 | | 39.26 | | 39.01 | | 33.73 | | bp004 | 1.172 |
| I0070 | OB | parachute | bp05_5 | SA | 0,1,9 | 56.43 | 3.80 | 57.58 | | 55.32 | | 37.07 | | 57.28 | | 57.82 | | 47.49 | | bp005 | 0.288 |

## Appendix 3 ~ Item Level Statistics

**National Year 6 Science sample assessment 2003**

| ItemID | Form | Question | Qcode | Type | Keys | SE | Residual | ChiSqu | Prob | Threshold 0_1 | Threshold 1_2 | Decsriptor | Strand | Conceptualised Level | Operational Level | Domain | Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I0001 | A | ice cubes | a1ic1r | SA | 0,1,9 | 0.151 | 0.692 | 8.775 | 0.032 | | | Interprets information in a contextualised report by application of relevant scientific knowledge | NP | 2 | 1 | A | SL |
| I0002 | A | ice cubes | a2ic2r | SA | 0,1,9 | 0.082 | -1.473 | 5.380 | 0.146 | | | Explains a choice based on first hand, concrete experience requiring the application of limited knowledge. | NP | 3 | 3 | A | SL |
| I0003 | COMMON | bar magnets | a3b1bm1r | MC | C | 0.069 | 3.441 | 54.825 | 0.000 | | | Identifies relationships of properties between objects that have been experienced or observed | EC | 4 | 2 | A | MC |
| I0004 | COMMON | bar magnets | a4b2bm2r | SA | 0,1,9 | 0.057 | 5.347 | 36.144 | 0.000 | | | Explains the relationship between events that have been observed | EC | 2 | 3 | A | SL |
| I0005 | COMMON | bar magnets | a5b3bm3r | SA | 0,1,9 | 0.060 | 0.495 | 3.697 | 0.296 | | | Describes the interaction between objects in terms of a non observable abstract scientific concept | EC | 3 | 3 | A | SS |
| I0006 | A | floors | a6fl1r | MC | A | 0.108 | -0.490 | 2.498 | 0.476 | | | Chooses between properties based on events that have been experienced or reported | NP | 2 | 2 | A | SS |
| I0007 | A | floors | a7fl2r | SA | 0,1,9 | 0.107 | 0.456 | 2.225 | 0.527 | | | Chooses between properties based on events that have been experienced or reported | NP | 3 | 2 | C | SL |
| I0008 | A | floors | a8fl3r | MC | A | 0.118 | -0.360 | 2.393 | 0.495 | | | Explains differences between properties and events | NP | 2 | 2 | A | SS |
| I0009 | A | echidna | a9e1r | MC | B | 0.096 | 1.251 | 7.085 | 0.069 | | | Interprets diagram containing interrelated elements to identify key elements | NP | 4 | 2 | A | MC |
| I0010 | A | echidna | a10e2r | MC | C | 0.172 | -1.593 | 7.148 | 0.067 | | | Interprets simple data from an image focusing on single aspect | LL | 3 | 1 | A | MC |
| I0011 | A | echidna | a11e3r | MC | B | 0.078 | 2.559 | 5.474 | 0.140 | | | Interprets information in a conceptualized environment by the application of scientific knowledge | LL | 4 | 3 | A | MC |
| I0012 | COMMON | planets | a12b4p1r | SA | 0,1,9 | 0.054 | 2.401 | 3.870 | 0.276 | | | Identifies patterns in scientific data represented in a diagram | EB | 3 | 4 | C | SL |
| I0013 | COMMON | planets | a13b5p2r | MC | A | 0.062 | 2.616 | 13.151 | 0.004 | | | Extrapolates from pattern observed and applies rule to describe expected outcome | EB | 3 | 3 | C | MC |
| I0014 | COMMON | camping holiday | a14b20ch1r | MC | D | 0.151 | -0.604 | 1.031 | 0.794 | | | Selects appropriate reason to explain reported observation related to personal experience | EB | 2 | 1 | A | MC |
| I0015 | COMMON | camping holiday | a15b21ch2r | MC | A,B | 0.094 | -0.700 | 1.957 | 0.581 | | | Extrapolates from an observed pattern and applies rule to describe expected outcome | EB | 3 | 2 | A | MC |
| I0016 | COMMON | camping holiday | a16b22ch3r | SA | 0,1,9 | 0.056 | -0.598 | 11.204 | 0.011 | | | Extrapolates from an observed pattern and applies rule to describe expected outcome | EB | 4 | 3 | A | SL |
| I0017 | A | mosquito | a17m1r | SA | 0,1,9 | 0.082 | -0.264 | 3.298 | 0.348 | | | Extrapolates from an observed pattern and applies rule to describe expected outcome | LL | 3 | 3 | B | SL |
| I0018 | A | mosquito | a18m2r | SA | 0,1,9 | 0.077 | -2.767 | 12.123 | 0.007 | | | Student is required to demonstrate an understanding of what is required for fair testing. | LL | 3 | 4 | C | SL |
| I0019 | A | mosquito | a19m3r | SA | 0,1,9 | 0.079 | -1.334 | 9.717 | 0.021 | | | Conclusion summarises patterns in the data | LL | 3 | 3 | C | SL |
| I0020 | A | mosquito | a20m4r | SA | 0,1,2,9 | 0.064 | -2.488 | 15.447 | 0.001 | 1.107 | 3.020 | identification of the anomaly in the table of data and the forming of a prediction or reason | LL | 3_4 | 5 | C | SL |
| I0021 | A | paper clip | a21pc1r | MC | B | 0.104 | 1.163 | 17.212 | 0.001 | | | Generalise from collected data presented in a table | NP | 3 | 2 | C | MC |
| I0022 | A | paper clip | a22pc2r | SA | 0,1,9 | 0.077 | -0.548 | 0.337 | 0.953 | | | Explanation of the principles of conducting an investigation and controlling variables | NP | 4 | 4 | B | SL |
| I0023 | A | bush pond | a23bu1r | MC | B | 0.077 | 1.833 | 1.787 | 0.618 | | | Identifies and summarises paterns in scientific data | LL | 3 | 3 | C | MC |
| I0024 | A | bush pond | a24bu2r | SA | 0,1,9 | 0.088 | -0.702 | 2.636 | 0.451 | | | Makes conclusions and presents summary of scientific data | LL | 3 | 3 | C | SS |
| I0025 | A | bush pond | a25bu3r | SA | 0,1,2,9 | 0.063 | 0.635 | 11.788 | 0.008 | -0.873 | 1.920 | Interprets reports and predicts changes in interrelationships | LL | 3_4 | 4 | A | SS |
| I0026 | A | wood burning | a26wb1r | SA | 0,1,2,9 | 0.056 | -0.629 | 5.319 | 0.150 | -0.602 | 0.549 | Applies knowledge of relationships of relationships to explain observed phenomenon | NP | 3 | 3 | A | SL |
| I0027 | A | wood burning | a27wb2r | MC | A | 0.094 | -1.388 | 4.743 | 0.192 | | | Application of the principles of conducting an investigation and controlling variables to select most appropriate methodology | NP | 4 | 2 | A | SL |
| I0028 | A | sandpaper | a28sp1r | SA | 0,1,9 | 0.127 | -0.414 | 1.534 | 0.674 | | | Identifies the difference between properties that have been experienced | EC | 2 | 2 | A | SS |
| I0029 | A | sandpaper | a29sp2r | SA | 0,1,9 © | 0.092 | -2.316 | 16.973 | 0.001 | | | Explains interactions that have been reported in terms of an observable property | EC | 3 | 3 | A | SL |
| I0030 | A | toy train | a30tt1r | SA | 0,1,9 | 0.078 | 2.353 | 5.991 | 0.112 | | | The student is required to suggest questions for testing | EC | 3 | 4 | B | SS |
| I0031 | A | toy train | a31tt2r | SA | 0,1,9 | 0.083 | 1.940 | 14.228 | 0.003 | | | The method of testing would be very simple and of the form of a comparison. | EC | 3 | 3 | B | SL |
| I0032 | A | toy train | a32tt3r | SA | 0,1,9 | 0.080 | -1.296 | 7.896 | 0.048 | | | Students are required to identify a pattern in the table. | EC | 3 | 4 | C | SL |
| I0033 | A | toy train | a33tt4r | SA | 0,1,9 | 0.083 | -1.548 | 10.151 | 0.017 | | | predict a relationship between the pattern and the cause | EC | 4 | 4 | C | SL |
| I0034 | COMMON | paper plane | a34b28pp1r | SA | 0,1,2,9 | 0.035 | 4.102 | 20.967 | 0.000 | 0.728 | 0.573 | Identifies and gives reason for controlling a single variable | NP | 3_4 | 4 | B | SL |
| I0035 | COMMON | paper plane | a35b28pp2r | MC | B | 0.056 | 0.110 | 1.135 | 0.769 | | | Applies knowledge of relationships of relationships to explain observed phenomenon | NP | 4 | 4 | B | MC |
| I0036 | B | colour fading | b6cf1 | MC | B | 0.087 | -0.921 | 4.087 | 0.252 | | | Given a question in a familiar context, identifies the variables to be considered | NP | 2 | 3 | B | MC |
| I0037 | B | colour fading | b7cf2 | SA | 0,1,9 | 0.094 | -0.581 | 1.639 | 0.651 | | | Interprets simple data set requiring an element of comparison | NP | 3 | 2 | A | SS |
| I0038 | B | colour fading | b8cf3 | MC | C | 0.079 | 1.598 | 1.286 | 0.733 | | | Explains processes or effects that have been experienced or reported in terms of a non observable property | NP | 4 | 3 | A | MC |
| I0039 | B | cave diggers | b9cd1 | SA | 0,1,9 | 0.077 | 1.166 | 0.196 | 0.978 | | | Describes the interaction between objects in terms of a non observable abstract scientific concept | LL | 3 | 4 | A | SL |
| I0040 | B | cave diggers | b10cd2 | MC | A | 0.076 | 1.030 | 2.905 | 0.407 | | | Interprets abstract diagram situated within an unfamiliar context | LL | 3 | 4 | A | MC |
| I0041 | B | cave diggers | b11cd3 | SA | 0,1, 9, © | 0.111 | -1.313 | 5.685 | 0.128 | | | Predicts the interaction between events in terms of an abstract scientific concept | LL | 3 | 2 | C | SL |
| I0042 | B | wheat growth | b12aw1 | SA | 0,1,9 | 0.076 | 0.335 | 2.828 | 0.419 | | | Summarises patterns in the information provided | LL | 2 | 3 | C | SS |
| I0043 | B | wheat growth | b12bw2 | SA | 0,1,9 | 0.076 | 1.114 | 0.602 | 0.896 | | | Summarises patterns in the information provided | LL | 2 | 3 | C | SS |
| I0044 | B | wheat growth | b13w3 | SA | 0,1,9 | 0.095 | -2.073 | 33.944 | 0.000 | | | Identifies a pattern in the information presented | LL | 3 | 2 | C | SL |
| I0045 | B | wheat growth | b14w4 | SA | 0,1,9 | 0.077 | -1.743 | 13.615 | 0.003 | | | Demonstrates awareness of the need for fair testing | LL | 3 | 3 | B | SL |
| I0046 | B | shadows | b15sh1 | SA | 0,1,9 | 0.079 | -0.850 | 11.084 | 0.011 | | | Identifies relationships between events that have been experienced or observed | EB | 3 | 4 | A | SS |
| I0047 | B | shadows | b16sh2 | MC | C | 0.125 | 0.187 | 7.296 | 0.063 | | | Identifies relationships between events that have been experienced or observed | EB | 2 | 2 | A | MC |
| I0048 | B | erosion | b17er1 | SA | 0,1,2, 9 | 0.052 | 1.235 | 8.733 | 0.033 | 0.237 | 1.261 | Explains interactions and effects that have been observed | EB | 4_4 | 4 | A | SL |
| I0049 | B | erosion | b18er2 | MC | D | 0.119 | -0.967 | 2.292 | 0.514 | | | Identifies and summarises patterns in science data | EB | 3 | 2 | C | MC |
| I0050 | B | erosion | b19er3 | SA | 0,1,9 | 0.078 | -3.832 | 28.830 | 0.000 | | | Applies knowledge of relationships to explain reported phenomenon | EB | 3 | 4 | A | SL |

| ItemID | Form | Question | Qcode | Type | Keys | SE | Residual | ChiSqu | Prob | Threshold 0_1 | Threshold 1_2 | Decsriptor | Strand | Conceptualised Level | Operational Level | Domain | Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I0051 | B | school electricity | b23se1 | MC | B | 0.076 | 1.639 | 9.963 | 0.019 | | | Identifies patterns in the data presented | EC | 3 | 4 | C | MC |
| I0052 | B | school electricity | b24se2 | MC | C | 0.076 | -1.085 | 11.425 | 0.010 | | | Identifies patterns in the data provided | EC | 3 | 3 | C | MC |
| I0053 | B | school electricity | b25se3 | SA | 0,1,9 | 0.087 | -2.540 | 26.222 | 0.000 | | | Identifies and extrapolates patterns in the data provided | EC | 4 | 5 | C | SL |
| I0054 | B | bean plant | b26bp1 | SA | 0,1,2, 9 | 0.067 | 0.022 | 1.362 | 0.714 | -1.788 | 1.470 | Identifies variable to be measured and/or controlled | LL | 3 4 | 3 | B | SL |
| I0055 | B | bean plant | b27bp2 | SA | 0,1,9 | 0.081 | -0.730 | 7.297 | 0.063 | | | Identifies appropriate method to ensure fair testing given abstract representation of context | LL | 3 | 4 | B | SS |
| I0056 | B | curtains | b30c1 | SA | 0,1,2, 9 | 0.061 | -0.992 | 5.478 | 0.140 | -0.419 | 2.561 | Explains interactions that have been reported in terms of an abstract scientific concept | NP | 4 4 | 4 | A | SL |
| I0057 | B | curtains | b31c2 | MC | B | 0.086 | 1.784 | 7.856 | 0.049 | | | Applies knowledge of relationships to make appropriate selection | NP | 2 | 3 | A | MC |
| I0058 | B | curtains | b32c3 | MC | A | 0.079 | 2.196 | 7.557 | 0.056 | | | Applies knowledge of relationships to explain observed phenomenon | NP | 3 | 4 | A | MC |
| I0059 | B | exercise | b33ex1 | SA | 0,1,2, 9 | 0.057 | -1.323 | 2.843 | 0.417 | 0.002 | 1.663 | Interprets information in a contextualised report by application of relevant scientific knowledge | LL | 3 4 | 4 | A | SL |
| I0060 | B | exercise | b34ex2 | MC | C | 0.100 | -1.983 | 18.679 | 0.000 | | | Selects appropriate reason to explain reported observationrelated to personal experience | LL | 3 | 2 | B | MC |
| I0061 | B | exercise | b35ex3 | MC | B | 0.087 | -2.227 | 11.678 | 0.009 | | | Demonstrates awareness of multiple variables in a system | LL | 4 | 3 | B | SL |
| I0062 | OA | craters | ap01_1r | SA | 0,1,9 | 0.078 | 2.234 | 5.201 | 0.158 | | | Explains outcome of scientific investigation | EB | 3 | 3 | C | SL |
| I0063 | OA | craters | ap02_2r | SA | 0,1,9 | 0.076 | 1.252 | 1.234 | 0.745 | | | Explains outcome of scientific investigation | EB | 3 | 4 | b | SL |
| I0064 | OA | craters | ap03_3r | SA | 0,1,9 | 0.077 | -0.096 | 1.557 | 0.669 | | | Explanation of the principles of conducting an investigation and controlling variables | EB | 3 | 4 | B | SL |
| I0065 | OA | craters | ap04_4r | SA | 0,1,9 | 0.095 | -0.828 | 3.034 | 0.386 | | | Extrapolates from experimental evidence to describe a different environment ( multiple variables | EB | 4 | 5 | A | SL |
| I0066 | OB | parachute | bp01_1 | SA | 0,1,9 | 0.079 | -1.453 | 12.069 | 0.007 | | | Proposes suitable method for fair collection of data | EC | 3 | 4 | A | SL |
| I0067 | OB | parachute | bp02_2 | SA | 0,1,9 | 0.099 | -0.270 | 9.670 | 0.022 | | | Generalize from collected experimental data and apply the rule to predict future events | EC | 2 | 2 | B | MC |
| I0068 | OB | parachute | bp03_3 | SA | 0,1,9 | 0.087 | -0.189 | 4.535 | 0.209 | | | Explaining the aim of an investigation with regard to multiple variables | EC | 4 | 5 | B | SL |
| I0069 | OB | parachute | bp04_4 | SA | 0,1,9 | 0.079 | -0.524 | 8.910 | 0.031 | | | Proposes suitable method for fair collection of data (multiple variables) | EC | 4 | 4 | C | SL |
| I0070 | OB | parachute | bp05_5 | SA | 0,1,9 | 0.079 | 0.931 | 0.221 | 0.974 | | | Extrapolates from experimental evidence to describe a different environment ( multiple variables) | EC | 4 | 3 | C | SL |

# APPENDIX 4

**Proficiency level, assessment domain descriptors, illustrative items and item descriptors**

| Proficiency Level (Scaled Location) | Assessment Domain Descriptors | Descriptor: A student at this level may display skills like: | Illustrative Items and Item Descriptor |
|---|---|---|---|
| **Level 4 and above**<br><br>**(δ = > 637)** | **Strand A:** Explains interactions, processes or effects that have been experienced or reported, in terms of a non-observable property or abstract science concept. | Explains interactions that have been observed in terms of an abstract scientific concept.<br><br>Interprets abstract diagrams situated within an unfamiliar context. | Item 48 [Erosion]<br>Explains interactions that have been reported in terms of an abstract concept.<br><br>Item 29 [Sandpaper]<br>Explains processes and effects that have been experienced in terms of a non-observable property. |
| | **Strand B:** Identifies the variable to be changed, the variable to be measured and several variables to be controlled. Uses repeated trials or replicates. | Demonstrates awareness of the need for fair testing by explaining how specific variables must be controlled.<br>Uses repeated trials and replicates in testing. | Item 34 [Paper Planes]<br>Has awareness of the need to control variables to ensure fair testing. |
| | **Strand C:** Can calculate averages from repeat trials, or replicates, plots line graphs where appropriate. Able to make conclusions to summarise and explain the patterns in the data. Able to make general suggestions for improving an investigation (e.g. make more measurements). | Critiques investigations noting design flaws.<br>Makes general suggestions for improving an investigation. | Item 20 [Mosquitos]<br>Identifies an anomaly in a table to formulate a prediction or reason. |
| | **Strand A:** Explains the relationships between individual events that have been experienced or reported and can generalise and apply the rule by predicting future events. | Applies knowledge of relationships to explain reported phenomenon. | Item 25 [Bush Pond]<br>Interprets reports and predicts changes in interrelationships. |

| | | | |
|---|---|---|---|
| **Level 3.3**<br><br>**(512< = δ <637)** | **Strand B:** Formulates scientific questions for testing and makes predictions. Demonstrates awareness of the need for fair testing. Makes simple standard measurements. Records data as tables, diagrams or descriptions. | Demonstrates an awareness of the principles of conducting an experiment and controlling variables.<br>Proposes suitable method for fair collection of data. | Item 22 [Paper clips]<br>Explains of principles of conducting an experiment and controlling variables. |
| | **Strand C:** Displays data as tables or bar graphs, identifies and summarises patterns in science data. Applies the rule by extrapolating or predicting. | Describes key features of a collected set of data, and can predict outcome of next event in series.<br>Extrapolates from an observed pattern to describe an expected outcome or event. | Item 62 [Craters]<br>Identifies and summarises patterns in experimental data. |
| **Level 3.2**<br><br>**(387 < = δ < 512)** | **Strand A:** | Interprets information in a contextualised report by applying relevant science knowledge.<br>Uses observed data and personal experience and applies rule to describe expected outcome. | Item 16 [Camping Holiday]<br>Interprets information from pattern observed and applies rule to describe expected outcome. |
| | **Strand B:** | Collates and compares data set of collected information.<br>Gives reason for controlling a single variable. | Item 61 [Exercise]<br>Identifies multiple variables in a system. |
| | **Strand C:** | Interprets diagrams and graphical data situated in a common or familiar context.<br>Makes conclusions and makes comparisons of scientific data. | Item 24 [Bush Pond]<br>Identifies and summarises patterns in scientific data. |
| | **Strand A:** | Selects appropriate reason to explain reported observation related to personal experience.<br>Identifies the relationship between events that have been observed or experienced. | Item 16 [Camping Holiday] (q1)<br>Selects appropriate reason to explain reported observation related to personal experience. |

| | | | |
|---|---|---|---|
| **Level 3.1**<br><br>**(262 < = δ < 387)** | **Strand B:** | Identifies the variable to be measured or controlled. | Item 54 [Bean Plants]<br>Identifies variables to be measured and/or controlled. |
| | **Strand C:** | Describes the findings of an experiment in simple terms focusing on one variable.<br>Interprets simple data set requiring an element of comparison. | Item 21 [Paper Clips]<br>Generalises from collected data represented in a table. |
| **Level 2 and below**<br><br>**(δ < 262)** | **Strand A:** Describes changes to, differences between or properties of objects or events that have been experienced or reported. | Describes a choice for a situation based on first-hand concrete experience, or requiring the application of limited knowledge.<br>Identifies the difference between properties that have been experienced. | Item 28 [Sandpaper]<br>Identifies the difference between properties that have been experienced. |
| | **Strand B:** Given a question in a familiar context, identifies a variable to be considered, observes and describes, or makes non-standard measurements and limited records of data. | Makes measurements or comparisons involving information or stimulus in a familiar context. | Item 67 [Parachutes]<br>Identifies the variable to be considered. |
| | **Strand C:** Makes comparisons between objects or events observed. | Identifies simple patterns in data and/or interprets a data set containing some interrelated elements. | Item 49 [Erosion]<br>Identifies and summarises patterns in science data. |