

Achievement in Science

Glenn S Chamberlain

chapter 5

CHAPTER 5

ACHIEVEMENT IN SCIENCE

Masures of students' science achievement were collected by administering tests and performance assessment tasks to provide a description of the *attained* curriculum in participating countries. The focus of this chapter is on the science achievement of New Zealand's standards 2 and 3 students. Included in this chapter will be an analysis of the achievement data by gender and ethnic groups. The content areas and test items which were relatively 'easy' and 'difficult' for New Zealand standards 2 and 3 students are also reviewed. Another area investigated is the relative 'growth' from standard 2 to standard 3, and any related gender differences. To enable the performance of New Zealand students relative to that of students in the same age-group from other participating countries to be judged, selected international results are included.

The Science Content Reporting Categories

The areas of science to be tested were those areas considered to be most crucial in the curricula of the countries taking part. These content areas, or *reporting categories*, were developed for the purpose of forming reliable sub-scales as well as incorporating specific 'in-depth' topics of interest to the international science education community. Most of the content areas included items contextualised in 'real world' situations. Not all topics were tested at this population level. The categories along with respective content areas and descriptions follow:

1. *Earth Science*

- a. *Earth Features* — with particular emphasis on the Earth's landforms, atmosphere, and rock and soil types.
- b. *Other Earth Science* — Earth processes including, for example, weather and climate, earth processes (eg water cycle), and Earth in the universe.

2. *Life Science*

- a. *Human Biology* — including many of the life science categories but specifically focusing on organs and tissues (eg circulatory systems, eyes, ears), and energy handling (eg respiration, digestion).
- b. *Other Life Science* — diversity, organisation, structure of living things, life processes and systems enabling life functions, life spirals, genetic continuity and diversity, and interactions of living things.

3. *Physical Science*

- a. *Energy and Physical Processes* — an emphasis on energy types, sources, and conversions.

- b. *Other Physical Sciences* — including matter, physical transformations, chemical transformations, and forces and motions.
4. *Environmental Issues and the Nature of Science*
- a. *Environment* — environmental and resource issues related to science including pollution, conservation, world population, food production and storage, and effects of natural disasters.
 - b. *Other Science Content* — concepts of technology; interactions of science, mathematics, and technology; and interactions of science, technology, and society.

Source: Minor adaptations from McKnight et al, 1993; Robitaille et al, 1993; TIMSS Study Centre, 1994.

TEST-CURRICULUM MATCHING ANALYSIS

For population 1 a total of 97 science items in multiple-choice, short answer, and extended response formats were rotated across eight separate booklets, with a set of ‘core’ items common to all booklets. To ensure the best possible match between the tests and each country’s science curriculum for the upper class level being tested, as well as ensuring a relatively uniform match across countries, the items were examined during several extensive selection stages (see Chapter 3). A Test-Curriculum Matching Analysis exercise was undertaken after the test was administered in 1994 to determine which of the items were judged to be in the *intended* curriculum for the population 1 class levels being tested in each of the participating systems. This enables the between-countries achievement comparisons based on the whole test to be complemented by comparisons based on items which excluded those not clearly in a national science curriculum.

The inclusion or exclusion of an item from the analysis was based on two factors. Firstly, whether or not the content of the item was part of the *intended* curriculum for more than 50 percent of the students and, if so, whether or not the item content would have been covered at the time of testing (TIMSS Study Centre, 1995). The acquisition of knowledge or skills required to answer items correctly from sources other than in science classes, or those not part of the *intended* curriculum for population 1 but taught in science classes, was not taken into account. Overall, 87 percent of the science items were thought to form part of the *intended* curriculum for New Zealand standard 3 students, based on the *Science Syllabus and Guide: Primary to Standard Four* (1979), *Health Education in Primary and Secondary Schools* (1985), and *Science in the New Zealand Curriculum* (1993). The other 13 percent of items represent content taught at this level in a majority of countries, but not in New Zealand. In contrast, 65 percent of the items were judged to be in the *intended* curriculum for New Zealand standard 2 students.

The results of the Test-Curriculum Matching Analysis for science can be seen in Martin et al (1997). As in mathematics, it can be seen that if the TIMSS test had been composed only of those items judged to be in the New Zealand *intended* curriculum, the international results would have been substantially the same as for the test used. In New Zealand, as in other countries, much of the science knowledge 9-year-olds have has been acquired from sources other than school science lessons.

OVERALL ACHIEVEMENT TEST RESULTS

This section will present the results of the students' performance on the science items by content area and class level. A mean score has been calculated from all responses, including both multiple-choice and free-response items, across all eight booklets at the population 1 level. Each student's data has been statistically adjusted by the use of weights so that results better reflect the average achievement of the 9-year-old population the sample was drawn from. Each correct multiple-choice item was scored one mark, while the short answer and extended response items were scored on a correctness scale — usually one, two, or three marks.

Figure 5.1



* Countries in italics did not meet all sampling criteria

▲	Mean achievement significantly higher than NZ
●	Mean achievement not significantly different from NZ
▼	Mean achievement significantly lower than NZ

Source: Adapted from Martin et al, 1997.

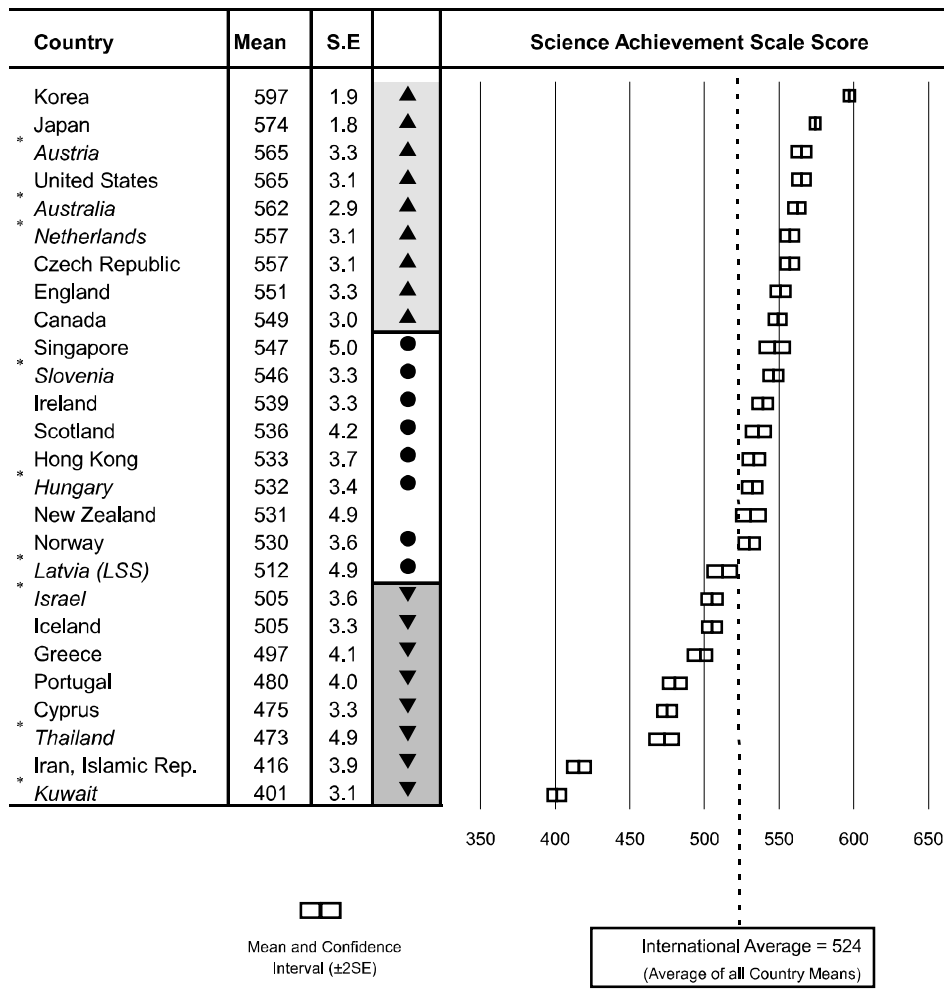
As in the mathematics chapter, scales based on Item Response Theory (IRT) were constructed to give comparisons between means of the test scores for each country, as shown in Figures 5.1 and 5.2. These scales are not linear. In percent correct terms, the highest mean for the upper grade

(standard 3 in New Zealand) is that of Korea (74%); while the lowest, that of Kuwait, is 39 percent. The range of mean scores is thus very great. Nevertheless, several countries in the middle of the range have mean scores which are not significantly different. The mean for New Zealand (61%) is not significantly different from those of Singapore, Slovenia, Ireland, Scotland, Hong Kong, Norway, Hungary, and Latvia. English-speaking countries outperforming New Zealand are Australia, Canada, England, and the United States. No English-speaking countries have significantly lower mean scores than New Zealand.

At the lower grade level (standard 2) the situation is similar, except that Canada's mean score is not significantly different from that of New Zealand, while that of Norway is significantly lower. It should be remembered that 9-year-olds in most countries have had either one or two years less schooling than New Zealand children.

Figure 5.2

Science achievement scores for the class level equivalent to standard 3 in New Zealand



* Countries in italics did not meet all sampling criteria

Percent Correct Scores and Practical Significance

As in Chapter 4, percent correct scores will be used for comparisons between New Zealand groups. Standard errors (s.e.) have been calculated, as described in Chapter 4, using the jack-knife procedure (based on replicated samples) for the reporting of the mean percent scores. This approach was necessary because the design of the study incorporated multiple-matrix-sampling, that is, each student received only a sample of the test items. As in Chapter 4, a 95 percent confidence interval applies, but again a level of ‘practical significance’ of three percent has been applied for some differences, with five percent for differences between smaller groups.

Table 5.1 contains the overall results of the New Zealand standards 2 and 3 students’ performance in science, by class level and gender.

Table 5.1

Mean science scores, by class level and gender

Class level	Girls’ mean percent (s.e)	Boys’ mean percent (s.e)	Overall mean percent (s.e)	Items in intended curriculum (%)
Standard 2	51.4 (0.9)	50.7 (1.1)	51.1 (0.9)	65
Standard 3	61.9 (0.9)	59.7 (1.2)	60.8 (0.9)	87

Notes: 1 Standard 2 girls N = 1290, boys N = 1214; standard 3 girls N = 1238; boys N = 1183.
2 (s.e) standard error.

The overall mean score of standard 2 students for science was 51.1 percent (s.e. 0.9%), whereas the equivalent figure at the standard 3 level was 60.8 percent (s.e. 0.9%). Differences between mean scores for boys and girls were not significant, although the difference for standard 3 of 2.2 percent in favour of girls approached significance. In contrast, Chamberlain (1996e) reported that population 2 boys scored higher than girls in science at both class levels (form 2 = 2.1% higher; form 3 = 4.2% higher). Gender analyses from past international studies of science achievement have produced mixed results. Lapointe et al (1992b) found that for 13 of the 14 countries taking part in the IAEP study, boys scored higher on average. However, results from the second IEA science study (of 10-year-olds) reported in Postlethwaite and Wiley (1992) were less clear-cut, with just 10 of the 18 participating countries having significant gender differences, all favouring boys. New Zealand did not take part in either of these studies.

In TIMSS population 1 only New Zealand, Latvia, and Thailand means were greater for girls than for boys. Although the differences are not significantly different, they do occur at both class levels. All significant differences were in favour of boys, and these occurred in about half of the countries. However, gender difference at this level is less pervasive internationally than at the population 2 (forms 2 and 3) level (Martin et al, 1997).

Differences in Achievement According to Types of Items

As mentioned earlier, the test items used in the TIMSS written tests comprised both multiple-choice and free-response (short answer and extended response) formats. Having a variety of item formats allows for a broader range of mathematics and science education outcomes to be measured. Furthermore, previous research has also suggested that boys may be unduly advantaged by the use of multiple-choice items for assessment purposes (eg Harding, 1979; Murphy, 1982). The results of the New Zealand mean achievement scores for science by item type and gender follow in Table 5.2.

Table 5.2

Mean science scores by item type, class level, and gender

Student gender	Standard 2		Standard 3	
	Multiple-choice mean percent	Free-response mean percent	Multiple-choice mean percent	Free-response mean percent
Girls	55	43	64	56
Boys	54	42	63	53

Notes: 1 Standard 2 girls $N = 1290$, standard 2 boys $N = 1214$.
 2 Standard 3 girls $N = 1238$; standard 3 boys $N = 1183$.

At the standard 2 level, there was a one percent difference favouring girls for both the multiple-choice and free-response formats. A similar pattern was evident for standard 3 students with girls doing slightly better again, although the difference for the free-response items was slightly larger, in the order of three percent. The results here suggest that neither sex was significantly disadvantaged by the use of either multiple-choice or free-response item formats at the 9-year-old level. A similar finding was reported at the forms 2 and 3 level (Chamberlain, 1996e).

Performance Assessment tasks

As mentioned in Chapter 3, another type of item used in TIMSS to provide a broader measure of student achievement was Performance Assessment. In addition to the mathematics tasks (see Chapter 4), there were five science and two combined (science and mathematics) tasks. The mean scores for the science tasks were *Pulse* (21%), *Magnets* (66%), *Batteries* (30%), *Rubber Band* (48%), *Containers* (37%), *Shadows* (25%), and *Plasticine* (33%). Pearson correlations revealed significant positive relationships between each of the performance tasks and students' achievement in science (overall $r = 0.40$, $p < 0.0001$). The strongest correlation was reported for *Rubber Band* (0.54) and the weakest with *Shadows* (0.18). Full details of the international administration of Performance Assessment can be found in Martin and Kelly (1996). Caygill (1995) conducted a preliminary analysis of the New Zealand results, while a more comprehensive report will be released in the near future.

Urban-Rural Location of Schools

Factors associated with the location of schools in some countries often have a major influence on the academic performance of the students. For example, schools in rural or remote areas (a community which is geographically isolated in a village or rural farm area) may have difficulties in attracting and retaining teachers, and gaining access to educational resources (eg libraries, museums) found in urban areas (a community which is on the outskirts of a town/city or located close to the centre of a town/city). A summary of the results comparing mean achievement for students in New Zealand schools in urban and rural locations can be found in Table 5.3.

Table 5.3

Students' mean science scores, by urban or rural location

Location of community	Standard 2 mean percent	Standard 3 mean percent
Urban	51	60
Rural	51	62

Notes: 1 Urban: standard 2 N = 2072, standard 3 N = 2029.
2 Rural: standard 2 N = 432, standard 3 N = 392.

There was no significant difference at either level but standard 3 students situated in rural schools did slightly better on average on the science test than their urban-based counterparts. Since the size of the difference (2%) is not of practical significance, this result supports the findings from past IEA studies of mathematics and reading (Slyfield,1986; Wagemaker, 1993b) that New Zealand students are not disadvantaged by attending schools in rural locations.

Home Language

When the main language spoken in the home is different from the language of the classroom, students face a disadvantage. Earlier studies by Lamb (1987) and Wagemaker (1993b) have demonstrated that when the home language differs from that used for instruction by the teacher, the level of achievement is likely to be lower. Population 1 students were asked to indicate to what extent they speak English at home. The TIMSS results are presented in Table 5.4.

Table 5.4

Mean science scores, by how often English is spoken in the home

Speaking English at home	Standard 2 mean percent	Standard 3 mean percent
Always, almost always	53	63
Sometimes, never	43	47

Notes: 1 Always, almost always: standard 2 N= 1975, standard 3 N = 2030.
2 Sometimes, never: standard 2 N= 399, standard 3 N = 297.

The data show a clear advantage in science achievement for those students who said they speak English ‘always or almost always’ at home. The differences between the two groups are very significant — 10 percent at the standard 2 level and 16 percent for the upper class level. At population 2 the results were very similar, with a difference of 12 percent for both form 2 and form 3 students who speak English at home on a frequent basis (Chamberlain, 1996e). Students who are less familiar with the language of instruction are more likely to be achieving poorly in science.

Achievement in Science Content Areas

Table 5.5 contains a summary of the standards 2 and 3 students’ achievement scores for the science content areas. International means for the topic areas are also included.

Table 5.5

**New Zealand and international mean scores
for reporting categories and content areas***

Reporting Category (content area)	Standard 2 equivalent		Standard 3 equivalent	
	NZ mean percent (s.e.)	International mean percent (s.e.)	NZ mean percent (s.e.)	International mean percent (s.e.)
<i>Earth Science</i>	47.8 (0.8)	48.9 (0.2)	57.5 (0.9)	57.2 (0.1)
• earth feature	43.1 (1.6)		52.7 (2.0)	
• other content	55.3 (1.9)		65.0 (2.1)	
<i>Life Science</i>	56.8 (1.0)	55.4 (0.2)	66.0 (0.9)	64.3 (0.1)
• human biology	46.2 (2.0)		56.4 (2.0)	
• other content	62.2 (2.2)		71.1 (1.9)	
<i>Physical Science</i>	46.7 (0.9)	47.8 (0.2)	56.7 (1.1)	57.1 (0.2)
<i>Environmental Issues and the Nature of Science</i>	42.8 (1.2)	40.5 (0.2)	53.9 (1.2)	51.3 (0.2)
• environment	44.7 (2.4)		55.1 (2.3)	
• other science content	39.5 (3.7)		51.8 (4.0)	

Note: * The content area means are based on the New Zealand national data and do not take into account the decision made by the International Study Centre to recode one extended response item. The reporting category figures reported in the table for New Zealand are those that are reported in the international report (Martin et al, 1997). The national New Zealand reporting category figure for the category affected was *Life Science* (S2: 57.6% and S3: 66.9%).

The international averages show that different content areas in the TIMSS tests were not equally difficult for the students taking the test. TIMSS has developed profiles of relative performance designed to show whether participating countries performed better or worse in some content areas than they did on the test as a whole, after adjusting for the differing difficulty for the items in each of the content areas (Martin et al, 1997). New Zealand standard 2 students performed better on *Environmental Issues and the Nature of Science* and worse on *Physical Science* than they did on the test as a whole, ie these were respectively areas of relative strength and weakness. In *Earth Science* and *Life Science*, mean performance was in line with overall performance. The only content area in which standard 3 performance deviated from overall performance was *Physical Science*, in which

performance was relatively weak. New Zealand patterns of achievement across topic areas were more similar to those of the United States and Ireland than to those of other participating countries.

It may be the case that a number of standards 2 and 3 students are missing the opportunity to learn important science knowledge and skills that are included as part of the *intended* curriculum. On the other hand, many students have knowledge not included in the *intended* curriculum. Students obtain much of their science knowledge from sources other than the school. Nevertheless, discrepancies between what was *intended*, what was taught, and what was learned seem to have been substantial.

‘Growth’ from Standard 2 to Standard 3

The study design used in TIMSS, which incorporates two adjacent class levels, allows for a quasi-measure of ‘growth’ by comparing the differences in mean achievement between standard 2 and standard 3. The overall growth across all seven content areas was approximately 10 percent, with the lowest improvement occurring in *Other Life Science* (9%) and the highest growth in *Other Science Content* (12%).

Differences between class levels were about nine percent for many of the participating countries (Martin et al, 1997). Generally, performance differences between grades were similar across content areas. Greatest overall growth occurred in Norway (15%), and the least in Korea (7%).

Gender Differences in Achievement

Another area of interest is a comparison of the respective performances of the boys and girls for each of the science content areas. The results of the gender analyses for New Zealand students are presented in Table 5.6.

Table 5.6

Mean scores in each science content area, by gender

Content area	Standard 2		Standard 3	
	Girls' mean percent (s.e)	Boys' mean percent (s.e.)	Girls' mean percent (s.e.)	Boys' mean percent (s.e.)
<i>Earth Features</i>	42.3 (0.9)	43.9 (1.1)	52.3 (1.1)	53.0 (1.3)
<i>Other Earth Science</i>	53.7 (1.3)	57.0 (1.3)	63.7 (1.3)	66.4 (1.5)
<i>Human Biology</i>	47.5 (1.2)	44.7 (1.3)	57.2 (1.3)	55.6 (1.2)
<i>Other Life Science</i>	64.1 (1.2)	60.1 (1.4)	74.1 (0.9)	68.0 (1.4)
<i>Physical Science</i>	45.7 (1.0)	47.6 (1.1)	56.4 (1.1)	56.9 (1.5)
<i>Environment</i>	45.3 (1.5)	44.0 (1.6)	56.2 (1.3)	54.0 (1.7)
<i>Other Science Content</i>	39.9 (2.3)	38.8 (2.9)	58.0 (2.4)	45.7 (2.8)

Notes: 1 Standard 2 girls N = 1290; boys N = 1214; standard 3 girls N = 1238; boys N = 1183.

2 (s.e) standard error.

At each class level, girls outperformed boys on four of the science content areas. Six of these differences were large enough to be of practical significance. At the standard 2 level girls did better in *Human Biology* (3%) and *Other Life Science* (4%), while boys typically achieved scores three percent higher than girls in *Other Earth Science*. In contrast, standard 3 boys achieved well in *Other Earth Science* (3%), whereas the girls scored higher in *Other Life Science* (6%) and *Other Science Content* (12%). Only two of the content areas with differences of practical significance were consistent across both class levels — *Other Earth Science* (boys) and *Other Life Science* (girls). One science content area — *Human Biology* — showed a practical difference in favour of girls at standard 2 but this was no longer evident by standard 3. Conversely, *Other Science Content* had no practical difference at the lower level but standard 3 girls scored, on average, 12 percent higher than the boys.

Martin and his colleagues reported that significant gender differences occurred in few countries for *Life Science* or *Environmental Issues and the Nature of Science* at either class level. In *Life Science*, at the upper class level (standard 3):

“ . . . boys did better than girls in the Netherlands, and third grade [standard 2] boys did better (than girls) in the Czech Republic, Hong Kong, Korea, and Portugal. Girls performed better than boys at both grade levels in New Zealand. For the items in the area of *Environmental Issues and the Nature of Science*, fourth-grade [standard 3] girls had higher achievement than boys in two countries — New Zealand and Slovenia, while boys had higher achievement in Austria.” (Martin et al, 1997.)

On the other hand, Martin et al (1997) reported many significant gender differences in *Earth Science* and *Physical Science*, all advantages for boys. In neither of these areas was there a significant difference between girls’ and boys’ performance in New Zealand, even though boys performed better on a subset of the *Earth Science* items, as discussed above.

ACHIEVEMENT OF STUDENTS FROM DIFFERENT ETHNIC GROUPINGS

The achievement in science of the different ethnic groups found in New Zealand is another major point of interest. Garden (1984) and Wagemaker (1993b) have reported on the relatively low average performance of the Maori and Pacific Islands students in comparison to that of their Pakeha/European counterparts in previous IEA studies of mathematics and reading literacy respectively. However, there is no information from past international studies involving New Zealand on the relationship between ethnicity and science achievement.

Standard 2

The mean percent scores for science by ethnicity and gender for the standard 2 students are summarised in Table 5.7. The 'Other' ethnic groups are excluded from the discussion because of the small numbers and the diverse range of backgrounds observed.

Table 5.7

Standard 2 students' mean science scores, by ethnic grouping

Student gender	Pakeha/ European mean percent	NZ Maori mean percent	Pacific Islands mean percent	Asian mean percent	Other ethnic groups mean percent
Girls	55	46	40	51	59
Boys	56	41	43	55	52
Total	56	43	42	54	55

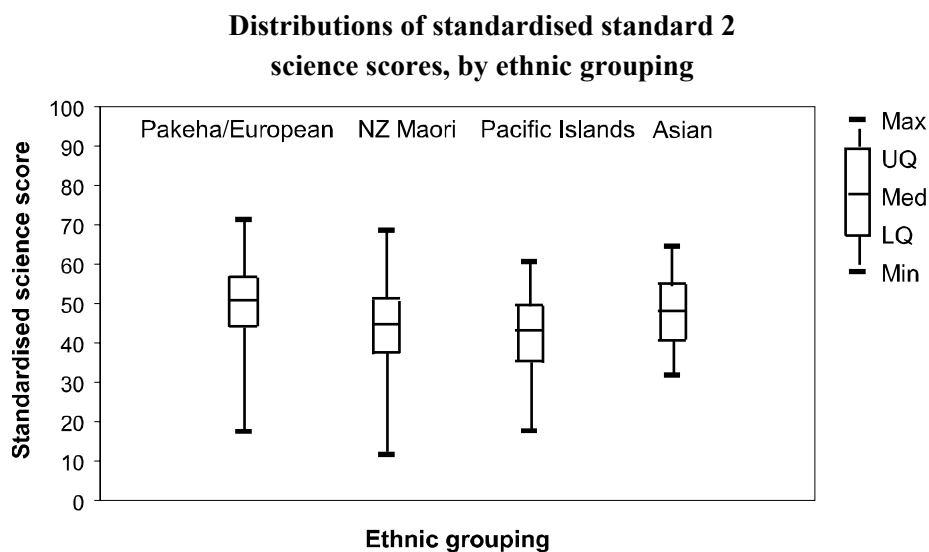
- Notes: 1 Pakeha/European N = 1574; Maori N = 592; Pacific Islands N = 145; Asian N = 90; Other ethnic groups N = 44; and missing (ethnicity) N = 59.
2 Standard errors (by gender) ranged from 0.9% to 1.6% for Pakeha/European and Maori students; and 2.3% to 6.4% for Pacific Islands, Asian students, and Other ethnic groups.
3 Overall standard errors: Pakeha/European 0.9%, Maori 1.3%, Pacific Islands 2.0%, Asian 2.2%, and Other ethnic groups 3.4%.

Overall, the standard 2 students from Pakeha/European and Asian ethnic groupings performed at comparable levels (mean percent scores of 56 and 54 respectively) of achievement in science. However, the Maori and Pacific Islands students were achieving, on average, some 12 to 13 percent behind the other groups. The pattern of achievement was quite similar at the New Zealand form 2 level (Chamberlain, 1996e). The results here confirm for science the findings from previous studies which focused on different subject areas.

Gender differences at the standard 2 level showed that Pacific Islands and Asian boys attained higher average scores of three and four percent respectively than the girls from these groups, while Maori girls performed better on average (5%) than Maori boys. The level of achievement for Pakeha/European boys and girls was nearly the same.

Measures of central tendency, such as the mean, can inform us how well particular groups have done relative to others. But it is also important to ascertain the variability of the achievement scores, that is to what extent they are spread out. Each ethnic group has some students who score either very well or very poorly, regardless of the mean score. Figure 5.3 presents the distribution of standardised scores for the standard 2 students in the different ethnic groups.

Figure 5.3



Note: Max = maximum; UQ = upper quartile; Med = median; LQ = lower quartile; Min = minimum.

The widest spread of science scores was observed in the Maori grouping, while students of Asian origin had the smallest range.

Standard 3

Table 5.8 contains equivalent information about science achievement by ethnicity and gender for standard 3. Again, there is no discussion of ‘Other’ ethnic groups.

Table 5.8

Standard 3 students’ mean science scores, by ethnic grouping

Student gender	Pakeha/ European mean percent	NZ Maori mean percent	Pacific Islands mean percent	Asian mean percent	Other ethnic groups mean percent
Girls	66	57	52	57	58
Boys	66	48	50	59	61
Total	66	52	50	58	60

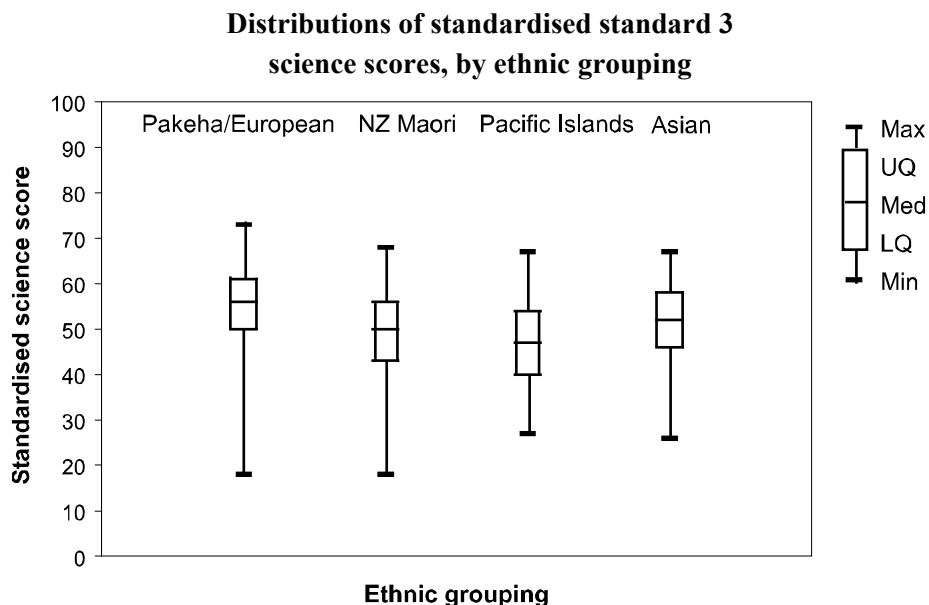
- Notes: 1 Pakeha/European N = 1531; Ma ori N = 563; Pacific Islands N = 132; Asian N = 101; Other ethnic groups N = 57; and missing (ethnicity) N = 37.
 2 Standard errors (by gender) ranged from 0.8% to 2.0% for Pakeha/European and Maori students; and 2.4% to 5.7% for Pacific Islands, Asian students, and Other ethnic groups.
 3 Overall standard errors: Pakeha/European 0.7%, Maori 1.7%, Pacific Islands 2.1%, Asian 2.6%, and Other ethnic groups 3.4%.

As at the lower level, the standard 3 Pakeha/European and Asian students performed markedly better in science than those from Maori and Pacific Islands ethnic groups. However at this level, a difference of practical significance was noted, with Pakeha/European students achieving some eight percent higher in comparison to their Asian class-mates. Similar results were observed at the form 3 level as well (Chamberlain, 1996e).

Standard 3 girls from Maori and Pacific Islands backgrounds scored higher, on average, than boys from these groupings, and Asian boys scored higher in science than Asian girls. Only one of these was of practical significance — Maori boys lagging about nine percent below the Maori girls. No gender difference was found for the Pakeha/European students.

The distribution of standardised science scores by each ethnic grouping for the standard 3 students is presented in Figure 5.4.

Figure 5.4



Note: Max = maximum; UQ = upper quartile; Med = median; LQ = lower quartile; Min = minimum.

At the standard 3 level Pakeha/European students produced the biggest range of scores, while those from Pacific Islands and Asian backgrounds had the smallest spread.

‘Growth’ from Standard 2 to Standard 3 in Different Ethnic Groupings

Data showing the growth in science achievement from the lower class level to the upper class level by ethnicity for population 1 is presented in Figures 5.5 (girls) and 5.6 (boys).

Figure 5.5

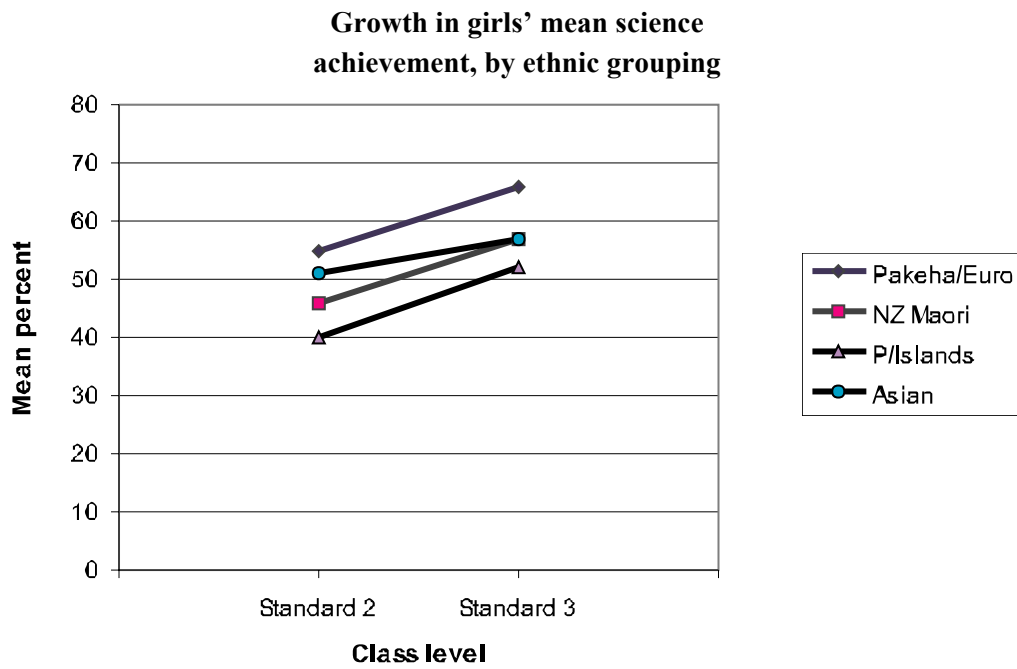
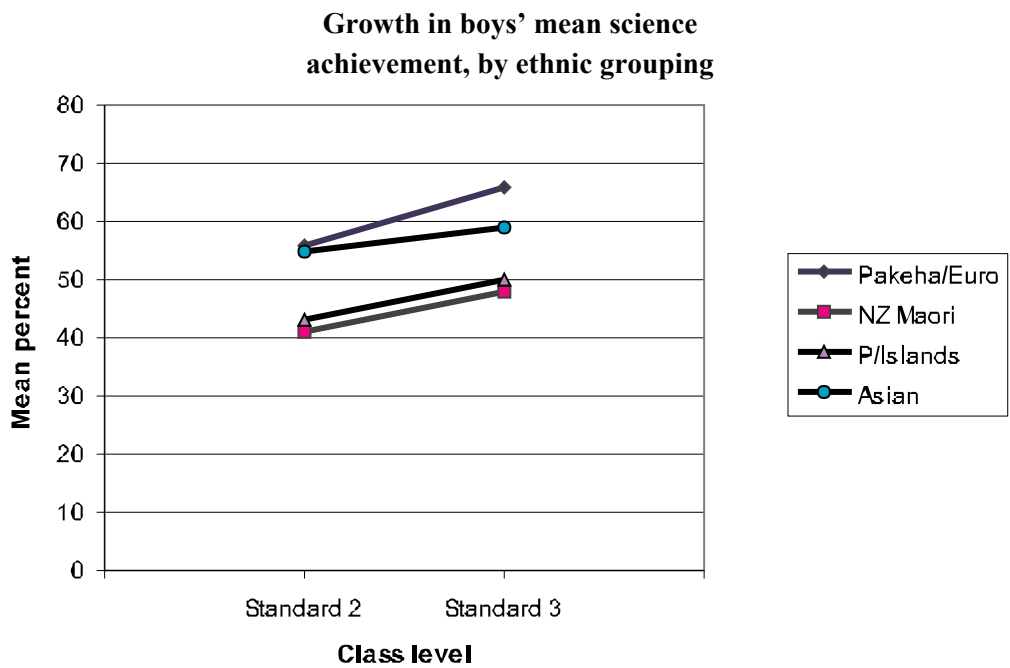


Figure 5.6



As shown in Figure 5.5, the amount of growth in science achievement for population 1 girls was consistent across three of the four ethnic groupings — Pakeha/European, Maori, and Pacific Islands — all recording an 11 or 12 percent increase. By comparison, the Asian students only achieved about half (6%) this amount of growth, but the number of Asian students in the sample is quite small and the apparent difference may be due to sampling error.

In contrast, the growth in science achievement for the boys (Figure 5.6) from standard 2 to standard 3 was more variable — Pakeha/European (10%), Maori and Pacific Islands (both 7%), and Asian (4%). Overall the girls achieved a higher level of growth in science than the boys, especially students from Maori and Pacific Islands backgrounds, where the differences of four and five percent respectively were of practical significance. At population 2, the extent of growth from form 2 to form 3 favoured the boys for all groups except for Asians (Chamberlain, 1996e).

Performance in Content Areas by Gender and Ethnic Grouping

To provide more in-depth detail of the performance amongst and within the ethnic groups, the students' mean achievement scores for each of the science content areas were examined. Summaries of the results are provided in Tables 5.9 (standard 2) and 5.10 (standard 3).

Table 5.9

Standard 2 mean percent scores for each science content area, by gender and ethnic grouping

Content area	Pakeha/European		NZ Maori		Pacific Islands		Asian	
	Girls (%)	Boys (%)	Girls (%)	Boys (%)	Girls (%)	Boys (%)	Girls (%)	Boys (%)
<i>Earth Features</i>	45	48	39	36	31	37	39	47
<i>Other Earth Science</i>	57	61	48	52	44	47	53	60
<i>Human Biology</i>	52	51	41	34	37	39	45	46
<i>Other Life Science</i>	69	65	57	49	53	58	69	66
<i>Physical Science</i>	49	53	42	39	34	33	44	52
<i>Environment</i>	49	50	40	33	33	36	43	51
<i>Other Science Content</i>	44	43	34	25	30	46	35	60

- Notes: 1 Other ethnic groups (N = 44) excluded from table.
 2 Pakeha/European N = 1574; Maori N = 592; Pacific Islands N = 145; Asian N = 90; and missing (ethnicity) N = 59.
 3 Standard errors (by gender) ranged from 1.0% to 2.6% for Pakeha/European; 1.2% to 3.3% for Maori; 1.9% to 6.0% for Pacific Islands; and 3.0% to 5.7% for Asian students.

Table 5.10

Standard 3 mean percent scores for each science content area, by gender and ethnic grouping

Content area	Pakeha/European		NZ Maori		Pacific Islands		Asian	
	Girls (%)	Boys (%)	Girls (%)	Boys (%)	Girls (%)	Boys (%)	Girls (%)	Boys (%)
<i>Earth Features</i>	56	59	46	44	44	35	51	50
<i>Other Earth Science</i>	68	70	57	58	56	65	60	69
<i>Human Biology</i>	62	62	50	44	49	45	51	54
<i>Other Life Science</i>	77	74	70	56	65	62	71	68
<i>Physical Science</i>	60	64	51	44	45	45	53	57
<i>Environment</i>	61	62	51	40	47	37	41	48
<i>Other Science Content</i>	63	53	54	28	38	43	54	55

- Notes: 1 Other ethnic groups (N = 57) excluded from table.
 2 Pakeha/European N = 1531; Maori N = 563; Pacific Islands N = 132; Asian N = 101; and missing (ethnicity) N = 37.
 3 Standard errors (by gender) ranged from 0.8% to 2.0% for Pakeha/European; 1.6% to 3.8% for Maori; 2.6% to 4.6% for Pacific Islands; and 2.9% to 7.8% for Asian students.

Pakeha/European and Asian students at the standard 2 level, on average, consistently outperformed their Maori and Pacific Islands counterparts in each of the science content areas tested. The former achieved higher scores of practical significance over the Asian group in the following content areas: *Earth Features*, *Human Biology*, *Physical Science*, and *Environment*. Meanwhile, Asian students did significantly better in *Other Science Content*. Overall, the gender results revealed that standard 2 boys from Pakeha/European (3 differences of practical significance), Pacific Islands (5), and Asian (5) backgrounds were doing better in comparison to the girls from those groups. In contrast, Maori girls generally performed better than the Maori boys in six content areas.

At the standard 3 level, the pattern was similar. Pakeha/European and, to a lesser extent, Asian students achieved higher scores than Maori and Pacific Islands students across all seven of the science content areas. However, the Pakeha/European group performed better at this level in each content area than the Asian students. The main change noted in the gender differences between the two class levels was the shift in advantage from the boys to the girls for the Pacific Islands students.

REPORTING CATEGORIES FOR THE PERFORMANCE EXPECTATIONS FOR SCIENCE

The *performance expectations* for science are non-hierarchical cognitive behaviours students are expected to be able to demonstrate as an outcome of science education. To assist in the reporting, the broad science performance expectations have been reorganised into the following categories:

1. *Understanding simple information* — demonstrating familiarity with vocabulary, facts, equations, and simple concepts.

2. *Understanding complex information* — integration of several pieces of simple information. Examples include differentiating, comparing, contrasting, synthesising.
3. *Theorising, analysing, and solving problems* — performances which require abstracting and deducing scientific principles; applying scientific principles to solve quantitative problems; applying scientific principles to develop explanations; constructing, interpreting, and applying models; or using scientific skills and knowledge to make decisions regarding personal, local, or societal issues.
4. *Using tools, routine procedures, and science processes* — includes using apparatus; conducting routine experiments; gathering data; and organising, representing, and interpreting data.
5. *Investigating the natural world* — includes identifying questions to investigate; designing investigations; execution of investigations; interpretation of and conclusions from investigational data.
6. *Communicating* — accessing and processing information, and sharing information. This performance was required from students on all extended response items in all reporting categories.

Source: Minor adaptations from McKnight et al, 1993; Robitaille et al, 1993.

‘EASY’ AND ‘DIFFICULT’ ITEMS

Most achievement tests contain items that, for various reasons, most of the students answer correctly and other items that most students are unable to answer correctly. It may be that some scientific terms or concepts are more easily understood than others, or some items may have been presented in a context that makes them less familiar. Another reason is that students may have been repeatedly presented with the content of some items either in school or from other sources, while other content areas they may not have met at all. It is a useful exercise to review those items that standards 2 and 3 students found either very easy or very difficult with regard to the *intended* curriculum. To provide a basis for selecting items, ‘easy’ items were defined as those that were correctly answered by 80 percent or more of the students, while ‘difficult’ items were those correctly answered by only 30 percent or less of the students.

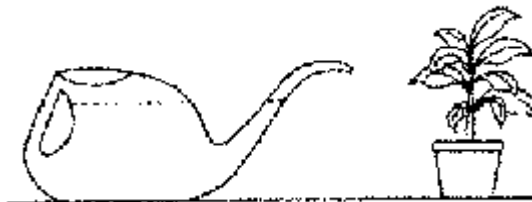
Nineteen science items (or parts of extended response items) from a total of 97 were classified as easy for New Zealand standards 2 and 3 students. Eight of these items were easy at both class levels, whereas the remaining items were easy at the standard 3 level only. The majority of the easy items (11) were found in the *Other Life Science* content area, while two items each came from the following areas — *Other Earth Science*, *Human Biology*, *Physical Science*, and *Environment*. In relation to the performance expectations, all but one of the items were assessing the *understanding of simple information*, and to a lesser extent, *understanding complex information*. Finally, all but four of the easy items were in multiple-choice format.

All of the items that the students found easy formed part of the *intended* curriculum at the standards 2 and 3 level. The ‘easiest’ item (E6) asked students to decide which of the Earth, Mars, the moon, or the sun is the hottest (*Other Earth Science; understanding simple information*). Ninety percent of New Zealand standard 2 students and 95 percent of standard 3 students chose the correct option. International means for the lower and upper grades were 89 percent and 93 percent respectively. Another of the easy items (R4) asked students to select the most important reason for people to use sunscreen when outside in the sunlight (*Other Life Science; understanding complex information*). Eighty-one percent of the standard 3 students scored this item correctly (international mean of 76%).

A total of 17 science items were classified as difficult for students at the population 1 level. Fourteen of these items were difficult at the standard 2 level only, while the other three were hard for both class levels. The difficult items were spread across five of the seven content areas, with the exceptions being *Other Life Science* and *Other Science Content*. All but one of the difficult items were found to be assessing the following performance expectations: *understanding simple information; understanding complex information; and theorising, analysing, and solving problems*. Approximately 70 percent of the difficult items were in a free-response format.

Half of the difficult science items were not included in the *intended* curriculum for the standards 2 and 3 students, thus the poor performance on these items can be explained. Here are two examples of items that students were expected to perform better on. Students were asked to draw a line on a watering can to indicate the surface of the water after the can had been tilted at an angle (*Physical Science; understanding complex information*). Only 13 percent of standard 2 and 17 percent of standard 3 students managed to get this item right. The international means were 16 percent and 21 percent respectively.

R1. A watering can is nearly filled with water as shown.



The watering can is tipped so that the water just begins to drip through the spout.

Draw a line to show where the surface of water in the can is now.



Copyright © 1994 by IEA, Amsterdam

In the other item, students had to work out which town had received snow from a table containing information about the weather (*Other Earth Science; using tools, routine procedures, and science processes*). Just 23 percent of standard 2 students were correct. However, as with the former item, it was found to be relatively difficult at this level in all countries (international mean of 32%).

N1. This table shows the temperature and precipitation (rain or snow) in four different towns on the same day.

	Town A	Town B	Town C	Town D
Lowest Temperature	13°C	-9°C	22°C	-12°C
Highest Temperature	25°C	-1°C	30°C	-4°C
Precipitation (rain or snow)	0 cm	5 cm	2.5 cm	0 cm

Where did it snow?

A. Town A

B. Town B

C. Town C

D. Town D

Copyright © 1994 by IEA, Amsterdam

GROWTH FROM STANDARD 2 TO STANDARD 3

The amount of ‘growth’ in science achievement from the lower class level to the upper class level provides another point of focus for examining the results on an item level basis. Reviewing items that have either very high or very low growth makes it possible to identify content which may need more emphasis in instruction. ‘High’ growth was defined as those items with a difference of 10 percent or more between the students’ average achievement for standard 2 and standard 3, while a difference of less than five percent was employed to defined ‘low or no’ growth.

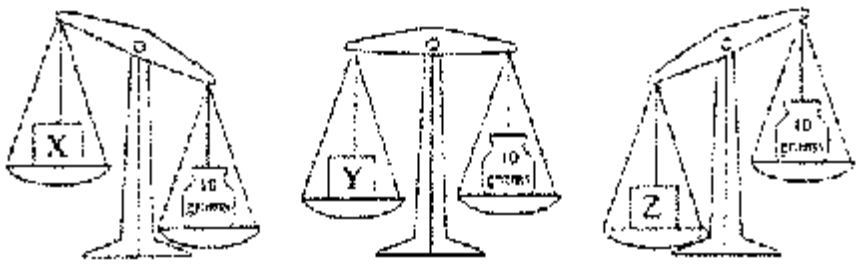
High Growth Items

Fifty-four science items (31 multiple-choice and 23 free-response) from a total of 97 were identified as having high growth for New Zealand students. All of the content areas contained at least three items of high growth, with the largest numbers occurring in *Physical Science* (18) and *Other Life Science* (14). *Earth Features* and *Human Biology* had five and eight items respectively. In relation to the students’ performance expectations, the sorts of knowledge and skills that the high growth items focused on included: *understanding simple information* (21 items); *understanding complex information* (18); *theorising, analysing, and solving problems* (10); and *using tools, routine procedures, and science processes* (4).

Nearly all (46 out of 54) of the high growth items formed part of the *intended* curriculum for science. Of these items, one-fifth were intended for standard 3 only, whereas the rest were taught at both class levels. Thus, students had more opportunity to learn some material at standard 3 than at standard 2.

Here is an example of a high growth item from the *Physical Science* content area assessing the student's ability to *use tools, routine procedures, and science processes*:

N9. Which of the boxes X, Y, or Z weighs the LEAST?



A. X
B. Y
C. Z
D. All three boxes weigh the same.

Copyright © 1994 by IEA, Amsterdam

Standard 2 students had a mean of 56 percent on this item compared with the 72 percent achieved by the standard 3 students, an increase of 16 percent. The item was in the *intended* curriculum for both class levels. The international means were 53 and 63 percent respectively.

Low or No Growth Items

There were only 11 items with growth of less than five percent between the two class levels. With the exception of one item, they were all in a multiple-choice format. One item (O7), asking students to select which of four animals produces milk for its young, provided an example of no growth, that is, the standard 2 students scored the same (73%) on average as their counterparts in standard 3. The low growth items were found in four of the science content areas: *Other Life Science* (5), *Physical Science* (3), *Earth Features* (2), and *Environment* (1). All but one of the items were assessing the students' *understanding of simple information and understanding complex information*.

Eight of the low growth items formed part of the *intended* curriculum for science, and seven of these items were taught at both standard 2 and standard 3. The low growth items were spread right across the full range of student achievement, from the most difficult (15%) through to the very easy (92%). At the forms 2 and 3 level, using the same five percent criteria, 47 science items with low growth were identified (Chamberlain, 1996f).

GENDER DIFFERENCES IN ACHIEVEMENT

Earlier in this chapter, it was reported that there was a difference favouring girls in the overall mean achievement of 0.7 percent at standard 2 and increasing to 2.2 percent for standard 3. These differences were not of practical significance. At population 2, there was a change with boys doing better in science than the girls by 2.1 percent (form 2) and 4.2 percent (form 3). To provide a more in-depth analysis of boys' and girls' results, the data selected will be reviewed in relation to content areas and performance expectations.

Table 5.11 contains a summary of the items for which there was a significant difference between the mean scores of boys and girls. For this section, a difference of at least five percent was set as the level of practical significance. Although the sample size for TIMSS is sufficiently large to produce smaller differences that are of statistical significance, the differences may not be enough to warrant discussion.

Table 5.11

Number of items on which girls or boys scored significantly higher, by class level

Difference in percent correct	Number of items on which more girls were correct		Number of items on which more boys were correct	
	Std 2 (N)	Std 3 (N)	Std 2 (N)	Std 3 (N)
5%–9%	25	13	15	10
10% or more	5	20	6	4
Total	30	33	21	14

Note: Includes parts of free-response items with own mark(s) allocation.

It appears from the results that the science achievement of boys relative to that of the girls declined from standard 2 to standard 3.

Gender Differences by Content and Performance Expectation

Standard 2

There were 30 items on which standard 2 girls did significantly better than the boys did. The items were spread across all of the science content areas, except for *Other Earth Science*. Half of the items were evaluating *Other Life Science*, while most of the remaining items were found in *Physical Science* (6), *Human Biology* (4), and *Environment* (3). Of the performance expectations under consideration, 15 items were assessing the students' ability to *understand simple information*; 10 items were assessing *understanding of complex information*; four items were assessing *theorising, analysing, and solving problems*; and one item was assessing *investigating the natural world*.

The total number of items that standard 2 boys scored significantly higher on in comparison to the girls was 21. The items covered the following content areas: *Physical Science* (8), *Other Life Science* (4), *Earth Features* (3), *Other Earth Science* (3), *Environment* (2), and *Other Science Content* (1). A larger proportion of these items were assessing the performance expectation *understand simple information* in contrast to the result for standard 2 girls. Other performance expectations where items were found included *understand complex information* (5); *theorising, analysing, and solving problems* (2); and *using tools, routine procedures, and science processes*. The items with the largest gender differences at standard 2 tended also to be in a multiple-choice format (40 out of 51).

Standard 3

Standard 3 girls outperformed their male counterparts to a significant level on 33 of the science items. About half of these items were focused on the *Other Life Science* content area. A further six items were evaluating *Physical Science* content, while three items were found in each of *Human Biology*, *Environment*, and *Other Science Content*. Twenty-seven of the 33 items were split evenly between the two performance expectations concerned with *understanding*. The remaining items were spread amongst *theorising, analysing, and solving problems* (3); *investigating the natural world* (2); and *using tools, routine procedures, and science processes*.

Standard 3 boys scored significantly higher than the girls on 14 items. The majority of these items were assessing *Physical Science* (6), *Earth Features* (4), and *Other Life Science* (2) content areas. With regard to the performance expectations, the main focus was on *understanding simple information* (8); *understanding complex information* (4); and *theorising, analysing, and solving problems* (2).

Of the 47 science items with significant gender differences at standard 3, approximately 70 percent were in a multiple-choice format.

Items With Largest Gender Differences

The TIMSS items that recorded the largest differences in scores by gender for standards 2 and 3 are presented here. Firstly, the items that girls achieved better on:

Q5. When an animal breathes faster and its heart beats faster, the animal is MOST likely

- A. cold
- B. frightened
- C. resting
- D. sleeping

Gender	Standard 2 (% correct)	Standard 3 (% correct)
Girls	72	79
Boys	46	67
Difference (%)	+25	+11

Copyright © 1994 by IEA, Amsterdam

This item was located in the *Other Life Science* content area and was assessing the student's ability to *understand simple information*. The item was judged to be not part of the *intended* curriculum for science at standards 2 and 3.

The next example is concerned with *Other Life Science* content and the performance expectation *understanding complex information*. This item forms part of the *intended* curriculum at both class levels.

R4. What is the MOST important reason for people to use sunscreen when they are outside in sunlight?

- A. It protects the skin against dangerous rays from the sun.
- B. It makes the skin more tanned.
- C. It makes the skin smooth.
- D. It makes the skin feel cooler.

Gender	Standard 2 (% correct)	Standard 3 (% correct)
Girls	71	92
Boys	66	69
Difference (%)	+5	+23

Copyright © 1994 by IEA, Amsterdam

Standard 3 girls found this item very much easier than did any of the other groups.

Second, the following two examples are of items that boys performed much better on in comparison to girls:

P7. Which travels fastest?

A. A train

B. An aeroplane

C. Sound

D. Light

Gender	Standard 2 (% correct)	Standard 3 (% correct)
Girls	26	42
Boys	50	67
Difference (%)	-24	-25

Copyright © 1994 by IEA, Amsterdam

This item's content area was *Physical Science* and the performance expectation was *understanding complex information*. The item was also included in the *intended* curriculum for standards 2 and 3.

The second example is also a *Physical Science* item.

P3. Which is NOT used as an energy source?

A. Flowing water

B. Iron ore

C. Sun

D. Oil

Gender	Standard 2 (% correct)	Standard 3 (% correct)
Girls	21	21
Boys	27	41
Difference (%)	-6	-20

Copyright © 1994 by IEA, Amsterdam

The performance expectation being assessed is *understanding simple information*. The item formed part of the *intended* curriculum at both class levels.

Life Science

The content area that girls consistently did better in compared to their male counterparts was *Life Science* (including *Human Biology*), which comprised over 40 percent of the total number of test items at standards 2 and 3. Most of these items focused on the ability of students to demonstrate *understanding simple information* and *understanding complex information*, and to a lesser extent, *theorising, analysing, and solving problems*, and *using tools, routine procedures, and science processes*. Eight of the items were in a free-response format.

At the standard 2 level, girls scored higher than the boys on 18 items out of 45 at a level of practical significance of five percent or higher. There were only 11 items where boys did better. At standard 3, 20 of the 45 items were correctly answered by girls at the same level of significance. The boys again scored higher on 11 items. It was apparent that a large variation existed in the size of the gender differences between the respective items for *Other Life Science* (5–25%) and *Human Biology* (5–9%).

Earth Science

The *Earth Science* (including *Earth Features*) content area was where 9-year-old boys consistently outperformed their female counterparts, which made up almost 20 percent of the test items for science. Most of the items were assessing either the *understanding simple information* and *understanding complex information*, or the *theorising, analysing, and solving problems* performance expectations. Four of the items were multiple-choice in format.

At the standard 2 level, in six out of 17 items boys' scores reached a level of practical significance of at least five percent higher on average than did the girls. There were just two items that the girls performed better on. For standard 3, five items showed significant differences in favour of the boys. The girls achieved higher means on six of the items. The boys recorded the largest gender difference of nine percent for *Earth Science*.

SUMMARY

The results presented in this chapter on the science performance of New Zealand standards 2 and 3 students who participated in TIMSS provided an overview of their achievement by gender and ethnicity, patterns of 'growth', content and performance expectations, and 'easy' and 'difficult' items. Selected international comparisons have also been presented.

Overall, the mean achievement of standard 3 students (61%) was 10 percent higher than the level attained by the standard 2 students (51%). This represents the growth in scientific knowledge and skills from one class level to the next. By content area, New Zealand standards 2 and 3 students performed best in relation to the students of other countries in *Environmental Issues and the Nature of Science*, while the area in which they were weakest internationally was *Physical Science*.

Interesting gender differences of practical significance were reported for some of the content areas. Boys achieved better than the girls on *Other Earth Science* at standards 2 and 3, whereas the girls outperformed the boys on *Other Life Science* again at both class levels. For *Other Science Content*,

a marked difference of 12 percent favouring girls was found at standard 3 — after a minimal difference at standard 2.

Important ethnic analyses were also conducted. It was evident that Pakeha/European and Asian students were doing better, on average, than students from the Maori and Pacific Islands groups. However, the Pakeha/European group achieved higher scores than those students from Asian backgrounds for most of the science content areas at standard 2 and all of the content areas for standard 3. The largest gender differences of significance within the ethnic groupings occurred for Maori students, where the girls outperformed the boys by almost 10 percent, and to a lesser extent Asian boys did better than Asian girls.

The use of two adjacent class levels in TIMSS provided a quasi-measure of ‘growth’ from standard 2 to standard 3. The overall rate of growth was about 10 percent, with the highest improvement occurring in *Other Science Content* (12%) and the lowest in *Other Life Science* (9%).

Nineteen of the 97 science items were identified as being easy for New Zealand standards 2 and 3 students. The majority of these items came from the *Other Life Science* content area, and were, in the main, assessing the *understanding of simple and complex information*. In comparison, there were 17 items that were rated as being difficult for our 9-year-old students. These items were found to be spread across five of the content areas, with a focus on *understanding and theorising, analysing, and solving problems*.

These results from TIMSS provide useful insights into the science achievement of the 9-year-old students in New Zealand. In addition, the data collected will provide a basis for further analysis and comparisons with future studies.