Junior Science

Thinking with Evidence

Technical information

Hilary Ferral

NZCER
Rangahau Mātauranga o Aotearoa

New Zealand Council for Educational Research
P O Box 3237
Wellington
New Zealand
www.nzcer.org.nz

ISBN: 978-0-947509-76-7

Contents

Tables

Figures

Technical information

Technical background: Rasch measurement and scales

Why do we need a measurement scale?

This series of *Junior Science: Thinking with Evidence* tests has been designed to assess achievement across year levels 4–6. If we want to measure student progress across time, achievement must be measured with the same metric at different time points if meaningful comparisons are to be made. In other words, the results from each test in the series must be able to be mapped to the *same* underlying scale.

In education, creating such a common scale is achieved by applying the Rasch measurement (RM) model, which was developed by the Danish mathematician Georg Rasch in the 1950s (Rasch, 1980)[1]. RM has been applied to the construction of interval scales in educational and other fields for over 40 years. NZCER's Progressive Achievement Tests in Mathematics and in Reading Comprehension and Vocabulary have undergone major revisions (in 2006 and 2008 respectively) using this methodology.

Measuring progress on a described scale

The development of the skills and knowledge required to "think with evidence" can be thought of as a journey that can be mapped along a continuum. As students progress along the continuum, their knowledge increases and their skills become more sophisticated. Knowledge and/or skill in a particular area of learning is not directly observable but can be inferred from responses to test items designed to probe a student's competency in that subject area. Each test item requires a certain level of subject knowledge and skill to be answered correctly.

This continuum can be mapped to a measurement scale that not only accurately identifies differences between students with different levels of skill, but also indicates differences between the difficulty levels of test items. A student located on the scale at the same place as a test item will have a 50% chance of answering that item correctly. Items located higher on the scale than the student will be more difficult for that student to answer correctly, and items located lower on the scale than the student will be easier to answer correctly.

Separating achievement scores from test items

In Traditional Test Theory (TTT), raw test scores are used to indicate the level of achievement for each student. In TTT we must always interpret student achievement and test difficulty in relation to the test that was used and the particular group of students involved. It is impossible to consider one without the other. The idea of separating or detaching item difficulties from student achievement has advantages and is central to the concept of objective measurement (Thurstone, 1928),[2] which underpins the RM model. Objective measurement requires items' placement on the scale to be independent of the group of students attempting those items. A useful analogy is to imagine measuring the heights of students:

[1] Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (revised and expanded edition). Chicago, IL: The University of Chicago Press [original work published 1960].

[2] Thurstone, L. L. (1928). Attitudes can be measured. *Journal of Abnormal and Social Psychology, 33*, 529–554.

whichever ruler we use, the taller students will always be taller and the shorter students shorter. That is, the scale (the ruler measuring inches or centimetres) is independent of which students we are measuring and how tall or short they are.

Constructing the scale

The construction of an RM scale begins with the assumption that a skill continuum exists and that students' locations on that continuum can be accessed by responses to items written to test (in this case) their skill in "thinking with evidence" in science.

A large number of test items are written, piloted, and assessed by subject matter experts, and finally put to the test by a large sample of students across a range of year levels. Students' responses are then applied to the Rasch model. The Rasch model is a probabilistic model that predicts the probability of success for any student on any of the items calibrated onto the scale. This probability depends *only* on the *difference* between the respective scale locations of the student and item. As a consequence, the location of items on the scale is independent of the distribution and position of student locations on the scale.

First, the test items need to be located on the scale—the calibration phase. Student responses to test items undergo a stringent series of diagnostic tests to assure good fit to the model for all items in the series of tests. Using statistical and graphical fit indicators, each item is examined for suitability for inclusion in the final test series. All items are required to fit independently and also to work in conjunction with other items to produce robust and reliable tests overall. Any items that do not fulfil the requirements of the model are rejected.

Once the items have fixed locations on the scale it becomes possible to place students' levels of achievement on the scale and, subsequently, to define the characteristic distributions of student achievement by year level, which can be used by teachers as reference points.

Practical outcomes of the Rasch model

The primary outcome of applying the Rasch model is a large bank of test items all located on the same scale. This bank can be organised into test forms that best target groups of students at different year levels.

The ability of RM to transform raw scores into scale scores (scale locations) allows us to interpret achievement levels without having to indicate the level of difficulty of the particular test administered to a student. This means that students can be located on the Junior Science (JS) scale regardless of which test in the series they have taken. However, it should be noted that tests need to be chosen at a suitable level for students. Tests that are very easy or very difficult for students will result in large measurement errors when estimating locations on the scale.

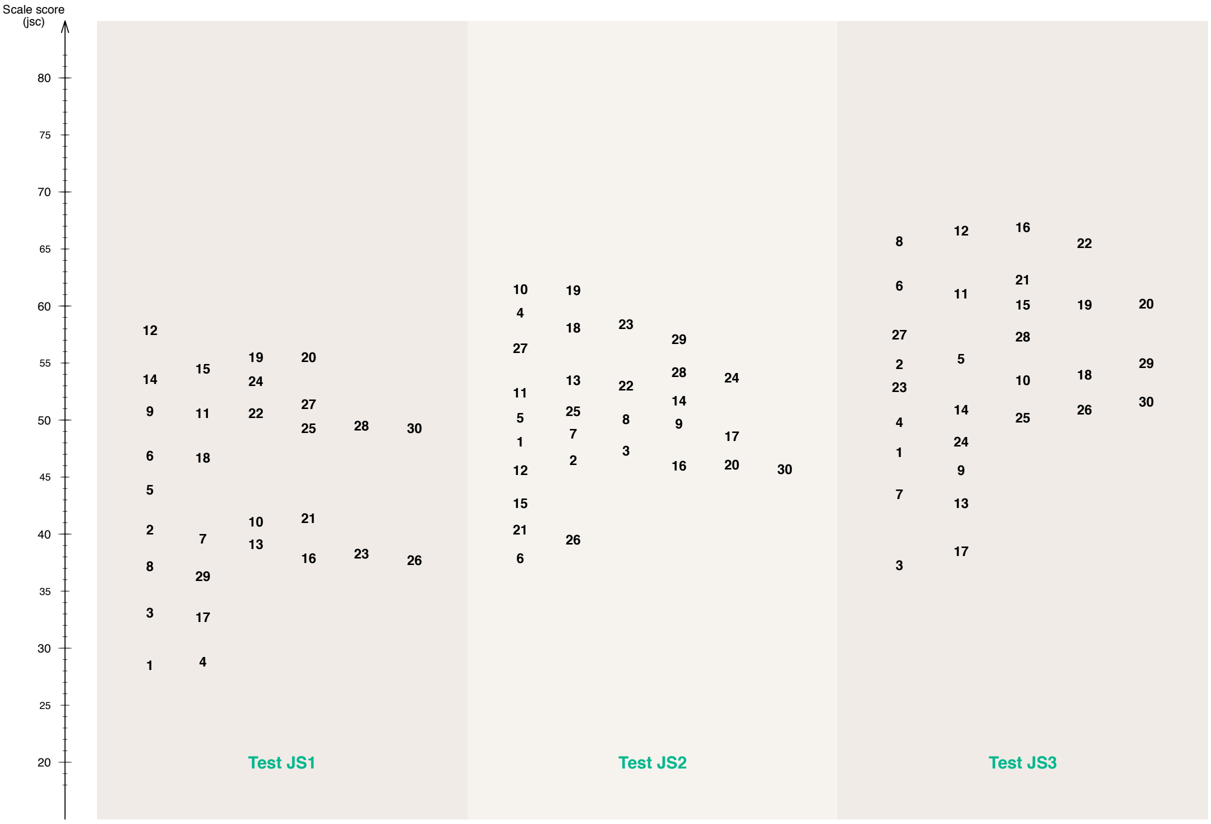
The Junior Science scale

Distribution of test items on the JS scale

Items by test

The scale measures student achievement and item difficulties in terms of a unit, called "jsc". Figure 1 shows the difficulty of all items in the series of tests on the JS scale. Item 1 in Test JS1 is the easiest item, and is located low on the scale at 28.5 jsc units. The most difficult item is item 16 in Test JS3, which is placed on the scale at just under 70 jsc units. The tests have been designed so that the average difficulty of the tests increases gradually with year level.

FIGURE 1 **Difficulty of items by test**



Summary statistics for each of the three tests in the *Junior Science: Thinking with Evidence* series are shown in Table 1. As we have seen, students located on the JS scale at the average difficulty of the test will answer about 50% of the questions correctly.

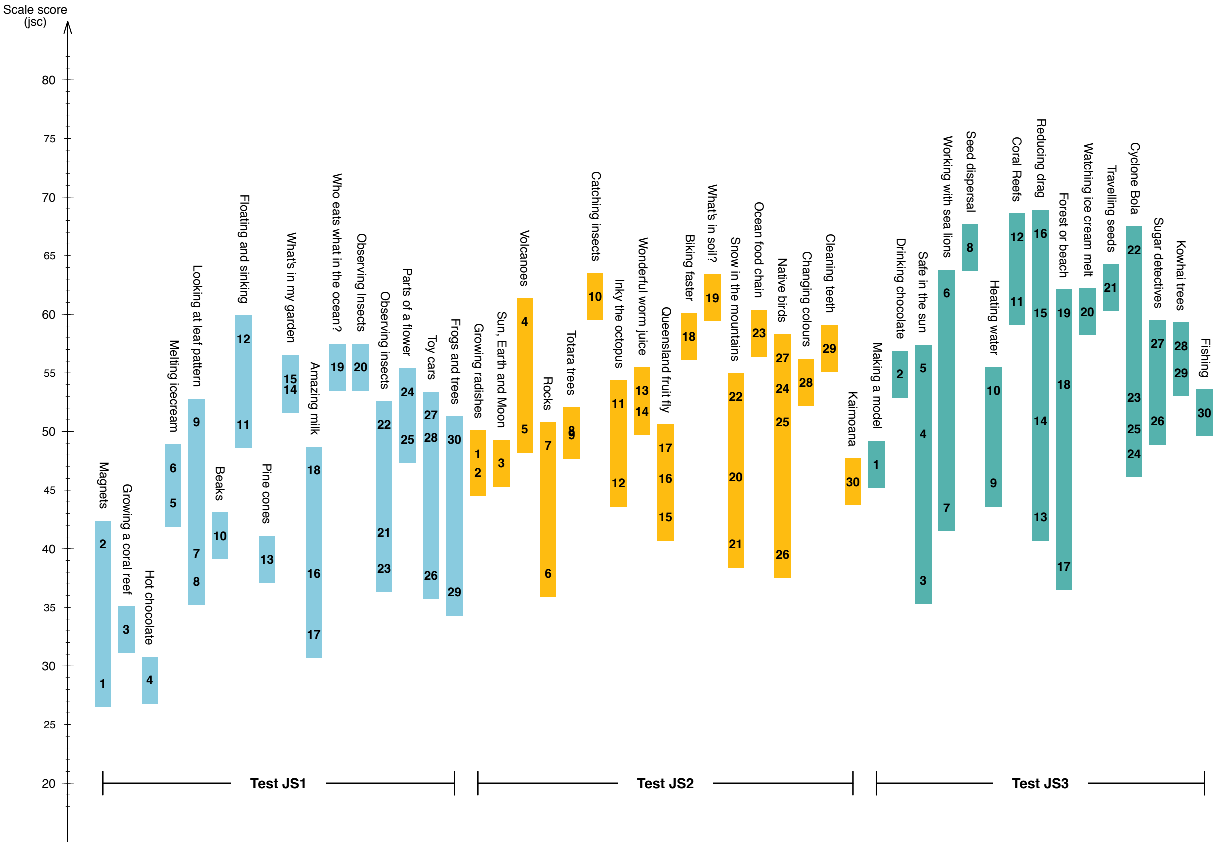
TABLE 1 **Summary statistics for the published *Junior Science: Thinking with Evidence* tests**

Test	Number of items	Average item location
JS1	30	44.4
JS2	30	50.5
JS3	30	54.2

Items by unit

Within each unit, item difficulty can be quite varied, or more clustered. Figure 2 shows all the JS items against the scale, organised into units. We can see how the items become gradually more difficult by test, and also how item difficulties are distributed within the units.

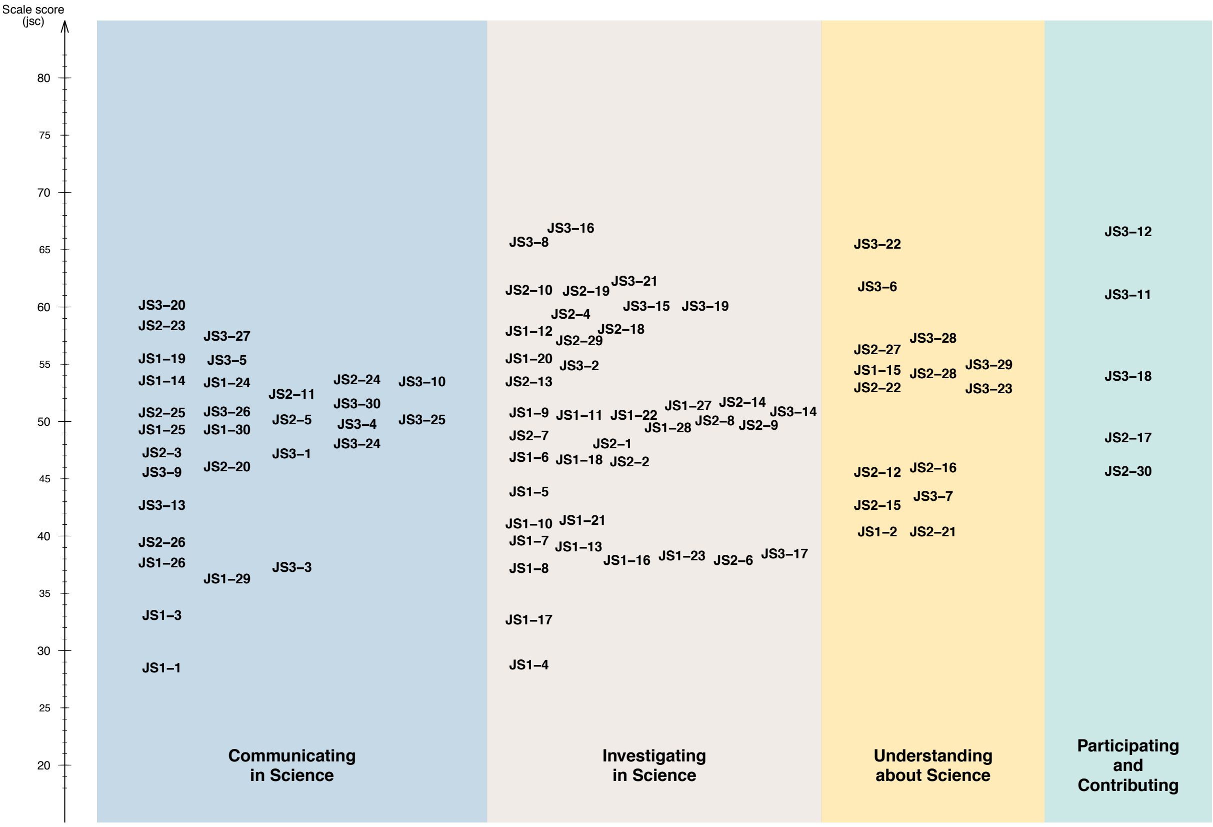
FIGURE 2 **Difficulty of items with in unit**



Items by Nature of Science sub-strand

Figure 3 shows all test items according to their Nature of Science sub-strand. Each sub-strand is represented by a wide range of difficulty, and each of the published tests contains a wide selection of items from each sub-strand. Each item is pre-fixed with its test number. For example JS2–16 represents item 16 in Test JS2.

FIGURE 3 **Items by Nature of Science strand**



Properties of the JS scale

RM produces measures that are recorded on an interval scale. This means, for example, that an increase of 1 jsc unit in a particular part of the JS scale represents the same 1 jsc unit increase in knowledge and skills anywhere else on the scale. Raw test scores do not have this property: they are ranks rather than measures. Each raw mark change generally represents a different amount of change in knowledge and skill, depending exactly where on the continuum the change is taking place. Around the average test score a one mark change represents much less change than a one mark change towards the top or bottom test score. The use of an interval scale means we can make meaningful comparisons between students who have completed different tests and between students' results from year to year.

Choice of scale units

The choice of numerical values and the names to assign to an RM scale are arbitrary. The idea is similar to measuring temperature, where degrees Celsius or degrees Fahrenheit may be used to mark the scale. The units of measurement used to express locations on the JS scale (jsc units) result from a transformation of the unit used by the Rasch model to locate items and students. The unit used by the Rasch model is called a logit.

The transformation from logits to jsc units is:
JS scale score (jsc units) = (logit + 5) * 10.

This means that:

- 1 jsc unit is equivalent to 0.1 logits
- the mean location of all JS items in this series of tests is 50 jsc units.

The transformation from logits to jsc units also means that a student placed at a particular point on the scale (say *b* jsc units) is expected to answer correctly:

- about 50% of items at that location
- about 30% of items at *b* + 10 jsc units
- about 10% of items at *b* + 20 jsc units.

Similarly, the same student (located at *b* jsc units) would be expected to be able to answer correctly:

- about 70% of items at *b* – 10 jsc units
- about 90% of items at *b* – 20 jsc units.

Construction of the *Junior Science: Thinking with Evidence* tests

The construction of the *Junior Science: Thinking with Evidence* tests followed a three-phase process.

Pilot phase

The first phase involved writing test items and reviewing them with the test development team at NZCER and a panel of external subject-matter experts. Items that were accepted from this process were piloted with small groups of students. Information from the pilot phase was used to form a collection of about 120 items suitable for including in the national item trial.

National item trial

Students from each of year levels 4, 5, and 6 were invited from a variety of schools across the country to do the trial tests—Year 4: 276 students, Year 5: 517 students, and Year 6: 344 students. Ten trial tests were constructed for the trial with each test containing some items that were common across the tests. Linking items across tests and year levels in this way allowed all items from all the tests to be calibrated onto a single scale. Items were assessed for fit to the Rasch model, their ability to discriminate between students who achieved well overall, and those who did not, and suitable targeting of suitable year levels. Some items were discarded due to poor fit and/or discrimination. Using the results from the national item trial, seven modified tests were constructed for the norming trial.

Norming trial

The norming trial followed a similar format to the national item trial. However, this norming trial involved fewer test forms, and a much larger sample of students. Around 1,000 students from each of Years 4, 5, and 6 were selected as part of a stratified random sample representing students across New Zealand in these year levels. A detailed description of the norming study is given in the next section.

A detailed analysis followed, and year-level distributions of achievement were described. The best performing items that targeted the range of achievement found at Year 4 to Year 6 were combined into the three final published tests described in this manual.

The norming study

The norming sample

To calibrate items on an RM scale it is necessary to test items with a moderate number of people who display the general characteristics of the population for whom the test items are intended. A strictly representative sample is not required by RM because the calibration of the items on the scale is statistically independent of the distribution of people attempting the items. However, it is informative for teachers to be able to make comparisons between their own class and a nationally representative group of students at the same year level. To achieve this sort of reference, it is essential to construct a large and nationally representative sample.

Stratification of the sample

The sample of schools was selected from a large group of schools that has indicated an interest in being involved in the development of the test. An important prerequisite for being part of this group was an assurance that the school had appropriate infrastructure to support an online assessment.

The sampling frame of schools was stratified by decile. Deciles were grouped 1–3, 4–7, and 8–10. Schools were selected with a sampling probability proportional to number of students attending schools in the respective decile groups. Each selected school was asked to provide a class of Year 4, Year 5, and Year 6 students to participate in the study. A "class" consisted of 25–30 students. We accepted smaller groups of students in smaller schools where there were not 25–30 students in the year levels of interest.

Once the sample had been selected it was inspected to make sure that it had reasonable representation across other variables such as school type, school location, and ethnic proportions.

In total, 54 schools participated in the JS norming study.

TABLE 2 **Number of schools participating in the norming study, by decile group and year level**

	Year 4	Year 5	Year 6
Deciles 1–3	8	7	8
Deciles 4–7	13	15	16
Deciles 8–10	15	15	16

TABLE 3 **Number of students participating in the norming study, by decile group and year level**

	Year 4	Year 5	Year 6
Deciles 1–3	176	140	175
Deciles 4–7	375	424	432
Deciles 8–10	375	398	486
Total	926	962	1,093

The norming study assessments were completed towards the end of Term 1, 2017.

Representation of the sample

In practice it is often the case that an achieved sample does not quite meet the requirements of the sample design. In this norming study, high-decile schools tended to be over-represented. A post-stratification procedure was carried out to examine the extent to which the achieved sample may have contained bias, and to make any corrections necessary.

Post-stratification

A repeated sub-sampling procedure was employed to remove any "decile bias" from the sample. The procedure involved making 200 random sub-selections from the achieved sample. Within each sampling stratum (decile group), each sub-sample contained student records in the correct proportions according to the most recent Ministry of Education school rolls database. Each sub-sample contained around 63%

of the achieved sample. The means and standard deviations from each sub-sample were calculated and averaged over the 200 readings to obtain final means and standard deviations to describe the distributions of achievement for each year level.

The sub-sample distributions of student achievement on the assessments at each year level closely approximated a normal distribution. This allows us to base the description of national year-level distributions on the means and standard deviations calculated in the post-stratification process.

Table 4 shows sample statistics for the distributions of Years 4, 5, and 6 students on the JS scale.

TABLE 4 **Summary statistics for the JS norming sample measured in Term 1, 2017**

	Number of students	Mean (jsc)	Standard deviation (jsc)
Year 4	908	41.3	9.9
Year 5	951	46.4	9.8
Year 6	1,089	50.1	9.4

Normative information

Normative information in this manual is provided to give teachers using these tests a reference with which to assess their students' achievement compared to a national group at the same year level. The normative information should not be used as a level or goal to strive for, but more as a guide to answer the question, "So how do my class results compare with similar classes nationally?" The formative feedback from the tests described in other parts of the manual will be more useful for guidance on future directions in teaching and learning.

Three achievement bands (Low, Medium, and High) are used to report the normative information. The "Low" band indicates that a student's score fell in the range of scores for students in the respective reference sample who achieved the lowest 23% of scores. The "Medium" band indicates a student's score fell in a range that matched the middle 54% of scores from the reference sample. Finally, the "High" band indicates that a student's score fell in the highest 23% of scores achieved by students in the reference sample.

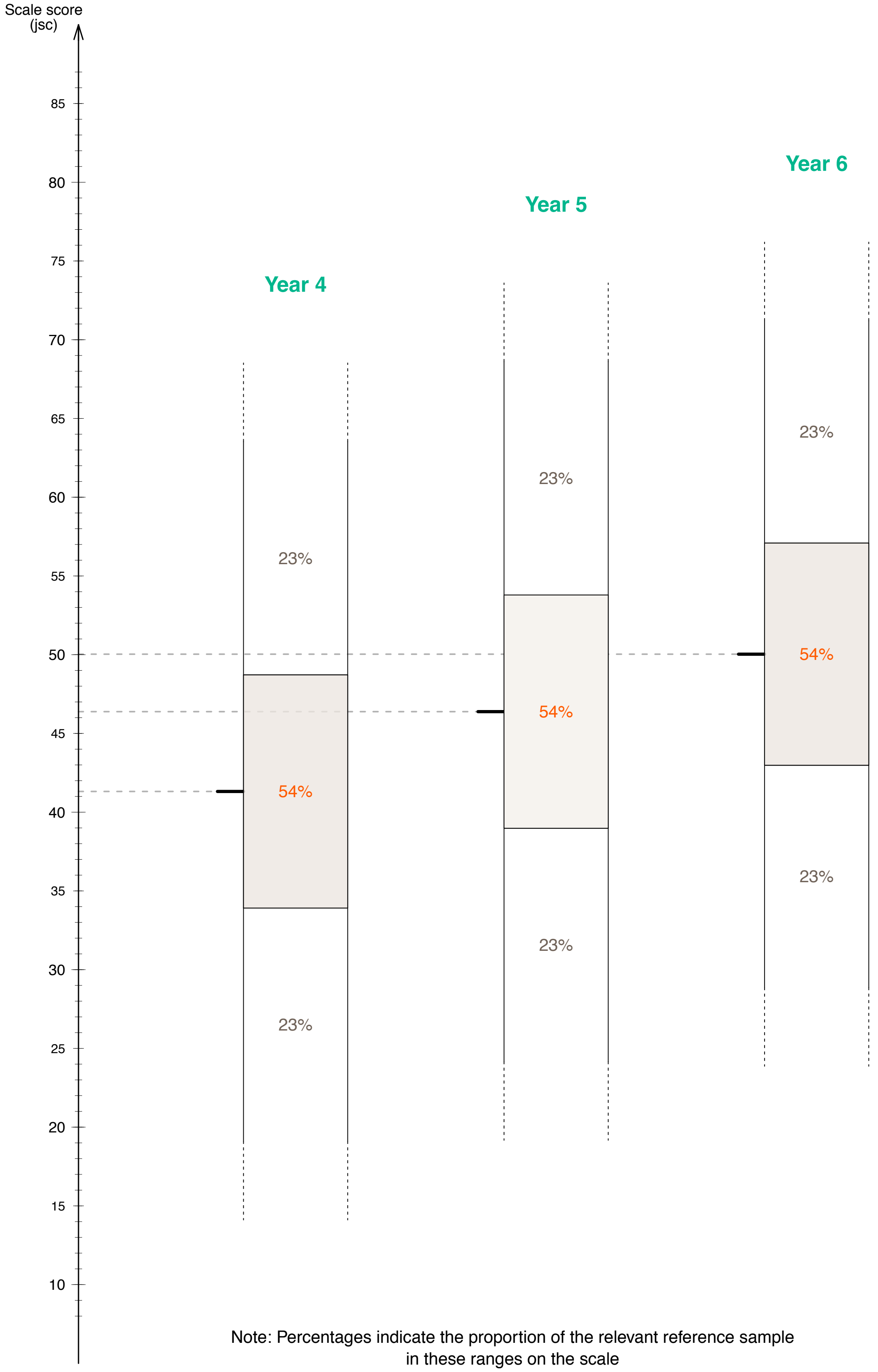
Figure 4 shows the distribution of student achievement for the reference samples, by year level, on the JS scale. For each year level, the Low, Medium, and High scoring bands are clearly indicated, along with the mean score.

Tracking progress on the JS scale

The JS scale provides the ability to track the progress of individuals, or groups of individuals, in a very simple way. Because the scale is an interval scale, and all the tests in the series have been mapped to the same scale, any of the tests will estimate a student's location on the scale, and the result can be directly compared to any previous results. This is easier, more accurate, and more informative than using norms, which are always dependent on the sample that provided the information.

For example, a class of students who in Year 5 score an average of 43 jsc units, and in Year 6 score an average of 50 jsc units, have made substantial progress—more than expected. From the normative information in Table 4 we can see that the expected growth on the scale is about 3.6 jsc units between Year 5 and Year 6, but this class has increased its average by 7 jsc units on the JS scale. See "Precision of scale scores" in the next section for more information about comparing scale scores.

FIGURE 4 **Distribution of student achievement on the JS scale, by year level**



Note: Percentages indicate the proportion of the relevant reference sample in these ranges on the scale

Reliability and validity

Test reliability

The reliability of a test is the degree to which it will provide consistent scores in repeated testing. Reliability coefficients are based on statistical calculations. A perfectly reliable test would have a reliability coefficient of 1, and a completely unreliable test (random estimates) would have a reliability coefficient of 0. A more reliable test leads to smaller errors associated with the scale score estimates based on responses to that test. A less reliable test leads to larger errors on the estimates.

The WINSTEPs software package that was used to construct the *Junior Science: Thinking with Evidence* scale was used to calculate a reliability coefficient for students' achievement locations on the scale. The coefficient was 0.80. This is a little lower than for some other assessments (for example, the Progressive Achievement Test: Mathematics) and is due in part to the length of the test forms being relatively short. It means that 80% of the variation in students' scores is attributable to real differences between their achievement levels and 20% is due to measurement error. A longer test (more than 30 items) would render more reliable scores. However, a longer test, with an unavoidably higher reading load, would not necessarily constitute an appropriate assessment of science for these younger students.

Precision of the scale scores

Each raw score on a test is associated with a scale score on the JS scale. Associated with every estimated scale score is a margin of error, within which we can be reasonably sure a student's true scale score lies. For example, a raw score of 14 in Test JS1 gives a scale score of 42.9 jsc units, and carries with it an error of 3.9 jsc units. This means that the true scale score will lie within +/– 3.9 jcs units of 42.9 (39.0 jcs units to 46.8 jcs units) in approximately 70% of cases.

It is important to notice that scale score errors increase dramatically for those students who do very well or very poorly on any particular test. A test will provide the most information about students who score around 50% on a test—thus scale scores are estimated more precisely around this percentage score. If a test is too hard or too easy for a student, a different test in the series should be selected which better suits the student and which can locate them on the JS scale more precisely.

A scale score resulting from test responses should be considered as a "band" on the scale rather than a precise point. This is especially important when comparing individual students' scores. Proper comparisons between students must take the errors surrounding the scale scores into consideration.

Validity

Validity can be defined as the extent to which a test measures what it was intended to measure.

The JS scale has been planned and constructed so that items on the scale assess the knowledge and skills advocated in the science education literature and reflected in the revised science learning area of *The New Zealand Curriculum* (Ministry of Education, 2007).[3] The items have been subjected to careful scrutiny by practising teachers and science teaching specialists and examined by NZCER test development staff.

Furthermore, the items that constitute this series of tests have been individually analysed and deemed to have satisfactory fit to the RM model. This means that, collectively, the items can be thought of as belonging to a single underlying construct representing a developmental continuum that can be qualitatively described *and* measured on an interval scale.

[3] Ministry of Education. (2007). *The New Zealand Curriculum.* Wellington: Learning Media.

Teachers intending to use one of the tests should prepare by working through the items to evaluate its suitability for the assessment objective (its validity), and to check for compatibility with the general aptitude of their class.

Gender differences

With respect to the national norms, there is a small difference at Year 4 between the average expected score for females compared with that for males. It is likely that this is strongly linked to the differences in outcomes for general literacy between males and females at this age. On this series of assessments the difference between average scores for males compared with females decreases in Year 5, and all but disappears by Year 6. This information is shown in Table 5.

TABLE 5 **Difference in average scores between males and females in the JS norming study**

	Females		**Males**	
	Mean (jsc)	**SD (jsc)**	**Mean (jsc)**	**SD (jsc)**
Year 4	42.9	9.7	40.0	9.6
Year 5	47.0	9.7	46.1	9.8
Year 6	50.1	9.8	50.1	9.1

Differential item functioning

Differential item functioning (DIF) occurs when an item functions differently for different sub-groups within the population. Typically we test for DIF between males and females, and between ethnic sub-groups. DIF was observed for a small number of items. These items were reviewed to examine whether they contained any features that might give a group an unfair advantage and any necessary action was taken to remove bias from the published tests.

Gender DIF

A gender DIF analysis revealed three items showing DIF; two where boys rated the item as relatively easier than girls and one where girls rated the item as relatively easier than boys. One item belonged to a three-item unit where the item in question was much more difficult than the other two. The decision was made to split the unit in order to target the items to more suitable year levels. The amount of DIF associated with the other two items did not have a noticeable effect on the overall test performance and it was decided to leave them in the tests as is.

DIF related to ethnic sub-groups

It is easy to misinterpret DIF analyses where one sub-group is large and the other rather small. The calibrations from the small sub-group will be comparatively imprecise, rendering what often looks like large differences. In this project we conducted DIF analysis with respect to Māori and non-Māori students. This was the only sub-grouping that made any sense to pursue.

Two items showed some (> 0.5 logits) DIF: one where Māori students rated the item as relatively more difficult, and one where non-Māori students rated the item as relatively more difficult. The two items are in different tests, and it was decided that the DIF effect on each of the published tests overall would be negligible, and therefore they were left in the published tests as is.

Score conversion tables

Score Conversion Table: Test JS1

JUNIOR SCIENCE TEST

JS1

Test score (number correct)	Scale score (stwe units)	Error (stwe units)	Year 4 reference	Year 5 reference	Year 6 reference
30	93.3	>10.3	23% of Year 4 students expected to score at this level	23% of Year 5 students expected to score at this level	23% of Year 6 students expected to score at this level
29	80.9	10.3			
28	73.5	7.4			
27	68.9	6.2			
26	65.4	5.5			
25	62.6	5.1			
24	60.2	4.8			
23	58.0	4.5			
22	56.0	4.4			54% of Year 6 students expected to score at this level
21	54.2	4.2			
20	52.5	4.1		54% of Year 5 students expected to score at this level	
19	50.8	4.1			
18	49.2	4.0			
17	47.6	4.0	54% of Year 4 students expected to score at this level		
16	46.0	3.9			
15	44.5	3.9			
14	42.9	3.9			23% of Year 6 students expected to score at this level
13	41.3	4.0			
12	39.8	4.0			
11	38.1	4.1		23% of Year 5 students expected to score at this level	
10	36.4	4.2			
9	34.7	4.3			
8	32.8	4.4	23% of Year 4 students expected to score at this level		
7	30.8	4.6			
6	28.6	4.8			
5	26.1	5.1			
4	23.2	5.6			
3	19.8	6.3			
2	15.1	7.5			
1	7.6	10.3			
0	–4.9	>10.3			

JUNIOR SCIENCE TEST

JS2

Score Conversion Table: Test JS2

Test score (number correct)	Scale score (stwe units)	Error (stwe units)	Year 4 reference	Year 5 reference	Year 6 reference
30	98.2	>10.2	23% of Year 4 students expected to score at this level	23% of Year 5 students expected to score at this level	23% of Year 6 students expected to score at this level
29	85.9	10.2			
28	78.5	7.4			
27	73.9	6.2			
26	70.5	5.5			
25	67.8	5.0			
24	65.4	4.7			
23	63.4	4.5			
22	61.4	4.3			
21	59.7	4.1			
20	58.0	4.0			
19	56.4	3.9			54% of Year 6 students expected to score at this level
18	54.9	3.9			
17	53.4	3.8		54% of Year 5 students expected to score at this level	
16	52.0	3.8			
15	50.5	3.8			
14	49.0	3.8			
13	47.6	3.8	54% of Year 4 students expected to score at this level		
12	46.1	3.9			
11	44.6	3.9			
10	43.0	4.0			23% of Year 6 students expected to score at this level
9	41.3	4.1			
8	39.5	4.3			
7	37.6	4.5		23% of Year 5 students expected to score at this level	
6	35.5	4.7			
5	33.2	5.0	23% of Year 4 students expected to score at this level		
4	30.4	5.5			
3	27.0	6.2			
2	22.5	7.4			
1	15.1	10.2			
0	2.7	>10.2			

JUNIOR SCIENCE TEST

JS3

Score Conversion Table: Test JS3

Test score (number correct)	Scale score (stwe units)	Error (stwe units)	Year 4 reference	Year 5 reference	Year 6 reference
30	103.0	>10.3	23% of Year 4 students expected to score at this level	23% of Year 5 students expected to score at this level	23% of Year 6 students expected to score at this level
29	90.6	10.3			
28	83.1	7.4			
27	78.5	6.2			
26	75.1	5.5			
25	72.3	5.1			
24	69.9	4.8			
23	67.7	4.5			
22	65.8	4.3			
21	63.9	4.2			
20	62.2	4.1			
19	60.5	4.0			
18	59.0	4.0			
17	57.4	3.9			
16	55.9	3.9			54% of Year 6 students expected to score at this level
15	54.3	3.9			
14	52.8	3.9		54% of Year 5 students expected to score at this level	
13	51.3	3.9			
12	49.7	4.0			
11	48.1	4.0	54% of Year 4 students expected to score at this level		
10	46.4	4.1			
9	44.7	4.2			
8	42.8	4.4			23% of Year 6 students expected to score at this level
7	40.8	4.6			
6	38.6	4.8		23% of Year 5 students expected to score at this level	
5	36.1	5.1			
4	33.3	5.6	23% of Year 4 students expected to score at this level		
3	29.8	6.3			
2	25.1	7.5			
1	17.6	10.3			
0	5.1	>10.3			