Norming the Progressive Achievement Tests in Mathematics with the Rasch Model

[a]Charles Darr & [b]Andrew Stephanou

[a]*New Zealand Council for Educational Research;* [b]*Australian Council for Educational Research*

(*Email:* Charles.Darr@NZCER.org.nz)

Abstract: In 1972 the New Zealand Council for Educational Research (NZCER) published a series of mathematics tests, the Progressive Achievement Tests in Mathematics (PAT Mathematics). In 2005 NZCER published a completely revised edition of these tests, centred on a qualitatively described Rasch measurement scale on which student achievement in mathematics from Year 4 to Year 10 can be measured. Reporting of student achievement is done both with reference to year level norms and to the skills that define each level of achievement. This paper describes the construction and description of this Rasch measurement scale and the innovative methodology for norming the published test forms. A distinction is made between trial forms, norming forms and published forms. The national norms for each year level from Year 4 to Year 10 provide a national profile of development in mathematics. The methodology used to construct the scale employed common-item equating to establish links between tests targeted at the different year levels. Common item equating was also used to link the NZCER scale with a similar scale developed by the Australian Council for Educational Research (ACER). The confirmation of a common underlying construct ended in a single NZCER/ACER scale for the measurement of development in mathematics on an interval scale in the same measurement unit, the patm. Results may be reported with reference to Australian or New Zealand norms for any year level for which data were collected. New tests can be calibrated onto this scale and normed with existing data. The transformation of test scores to scale scores makes possible the comparison of results obtained with any of the tests that are calibrated on the scale.

The Progressive Achievement Test of Mathematics

In 1972 the New Zealand Council for Educational Research (NZCER) published a series of mathematics tests for students in Years 4 to 10 (Reid, 1974 and 1993), entitled the Progressive Achievement Test of Mathematics (PAT Mathematics). The tests were part of a range of Progressive Achievement Tests (PATs) developed for New Zealand schools in the late sixties and early seventies that included tests of reading (comprehension and vocabulary), listening, and study skills (Reid, 1969 and 1991, 1971 and 1994, 1978).

The development of the PATs was based on Classical Test Theory (CTT). Achievement on each test was reported in the form of year level and age level norms using percentiles and stanines. For example, students in a particular year group who scored at the fiftieth percentile achieved a test score (raw score) that was greater than or equal to the test scores achieved by 50% of the students who sat the test as part of a reference group carefully chosen to represent the national cohort of students at that year level. Schools were encouraged to use the test results to record student progress and to help teachers group their classes for instruction. Although use of PAT Mathematics was not mandatory, most schools did administer it.

The original PAT Mathematics was made up of two parallel series of test forms. Both series involved seven forms called parts, with each part targeted at a specific year level. Multiple choice items were used exclusively in each test, with half of the items in each form shared with the form immediately below in the series and the other half with the form immediately above. The parallel test forms were developed to be used in alternate years and ensured that students did not repeat any of the overlapping items from one year to the next. Initially, PAT Mathematics was provided free to schools by the New Zealand government. After the education reforms of the early 1990's, schools who wanted to continue using PAT tests had to purchase them directly from NZCER. A large number of schools continued to do this. A revised edition of the test following the same design but with updated norms and some new and revised test items was published in 1992.

In 2004 NZCER embarked on a project to redevelop and update the norms for PAT Mathematics in collaboration with test developers and psychometricians from the Australian Council for Educational Research (ACER). At that time ACER staff were working on a revision of the ACER PATMaths instruments. During the 1980s ACER had also developed a range of PAT tests (e.g. ACER, 1984). Originally based on the New Zealand designs, subsequent revisions of the Australian tests lead to the development of Rasch Measurement (RM) scales (ACER, 1997) and innovative reporting templates (ACER, 2001; Lindsey, 2005). It was decided that the new NZCER version of PAT Mathematics should be developed according to the requirements of RM (Andrich, 1988). In addition, it was agreed that NZCER and ACER would carry out an equating study aimed at the construction of a common NZCER/ACER PAT Mathematics measurement scale.

Rasch Measurement

The development of mathematical knowledge and skills can be mapped along a continuum. As students' knowledge increases and their skills become more sophisticated, they move along the continuum. A student's level of mathematical knowledge and skills is not directly observable, but can be inferred from responses to test items designed to probe mathematical understanding. Each test item requires a certain level of mathematical knowledge and skills in order to be answered correctly.

Progress in mathematics can be measured by constructing a scale that represents this mathematics continuum — a map of mathematical competencies. Such a scale would allow us to locate the achievement levels of different students. It would also allow us to locate the knowledge and skills required to correctly answer test items. Each student achievement corresponds to a fixed location on the scale (student location) and each item (in terms of its difficulty in relation to other items) has a fixed location on the same scale (item location).

Test scores in CTT are indicators of level of student achievement. However, test scores (number of items answered correctly) depend as much on the complexity of the test items as on the skill levels of the students. In a similar way, the facility of each item (the proportion of students answering an item correctly) depends as much on the level of the skills of the particular group of students taking the test as it does on the level of skills that each item requires to be answered correctly. As a result, test scores and item facilities must always be interpreted in terms of the particular test used to generate the test score and the particular sample of students who sat the test containing the items. There is no scale of measurement in CTT that defines a continuum on which both student achievement and item difficulty can be located. In fact there are two scales containing the same information presented in two different ways, one through test scores and the other through item facilities.

Separating the estimation of item locations from student achievement is central to Objective Measurement (OM). OM requires the relative location of pairs of items to be the same for students at any location on the scale (within measurement error and in the absence of misfit), and it requires the location of achievement to be independent of the particular set of items included in the test or tests administered (Thurstone, 1928). Both requirements are identical to those used for measurement in the physical sciences.

OM in education is achieved through the application of the Rasch Measurement model, which was developed by the Danish mathematician Georg Rasch in the 1950s (Rasch, 1960). RM has been applied in the construction of interval scales in education and other fields around the world since the 1960s.

Constructing a Rasch measurement scale

RM assumes that the observed achievements of students and the observed relative difficulties of test items can be represented by fixed locations on an interval scale. Each location on the scale is said to correspond to a certain amount of the attribute being measured. RM then proposes a mathematical model to predict the probability of success for any student on any of the items calibrated onto this scale. According to the model, this probability depends only on the difference between the respective scale locations of student and item. A student who is at the same location as a group of items is expected to answer 50% of these items correctly. The same student is expected to answer correctly more than 50% of items located lower on the scale and fewer than 50% of items located higher on the scale. A consequence of this assumption is that the location of items on the scale is independent of the distribution and position of student locations on the scale. The same item locations are expected, within error and in the absence of misfit, for different distributions of student achievement.

RM produces measures that are recorded on an interval scale in a unit called the logit. This means that an increase of one logit in any part of the scale represents the same size growth in knowledge and skills anywhere else on the scale. Test scores and percentiles do not have this property—they are ranks rather than measures. Identical changes between two test scores may represent different amounts of change in the knowledge and skills represented by the scale score (in logits or patm units). For students achieving around the mean score in a test, a change of one mark or one percentile represents only a very small change in the scale score. However, for students achieving towards the top or bottom of their year level, a change of one test mark represents a much larger change in the scale score.

The construction of an RM scale for mathematics involves writing items designed to assess mathematical knowledge and skills, trialling the items, and collecting norming data by means of test forms assembled using trialled items. By including common items in tests, or administering more than one test to the same selection of students, all items can be linked across the different test forms. The data are then analysed with the Rasch model, initially to estimate the relative location of items on the scale (item calibration). A meticulous fit analysis, examining each item in the light of both statistical fit indicators and graphical displays, shows how well the data fit the measurement model and exposes any items that did not perform as well as expected. Once the scale is finalised, it becomes possible to estimate student locations on the scale and obtain the distribution of student achievement by year level. During the development of PAT Mathematics, RM was applied in the piloting, trialling and norming stages to make sure all items fitted within acceptable tolerances. Items that did not fit the model satisfactorily were excluded from the final tests.

The practical outcome of applying RM is a bank of items distributed along a single measurement scale that can be described qualitatively to show the range of mathematics knowledge and skill levels appropriate to student from Year 4 to Year 10. Items from this bank can then be combined into test forms that best target the location of students in the representative norming sample for each year level.

The ability of RM to transform test scores into scale scores allows us to measure achievement without having to indicate the level of difficulty of the test administered to a student. It is important to note, however, that not all the tests are suitable for administration to any one group of students. To be suitable, the test has to match their range of knowledge and skills. A test that is relatively very easy or very difficult will result in large errors of measurement on the scale.

Once constructed, a scale can be described qualitatively by examining the items in each region of the scale and summarising the knowledge and skills that are characteristic of these items. A student located at any point on the scale is likely to have mastered the skills below that location and less likely to have mastered those above. A student achievement in PAT Mathematics can be reported in terms of both the knowledge and skills exhibited (formative reporting) and the relative position in the distribution of locations of students in a given year level (normative reporting).

The test design and development process for the new NZCER PAT Mathematics

To begin the development process, specifications were drawn up to describe the structure and content of the proposed tests. It was decided that the new PAT Mathematics would be made up of seven separate tests, each targeted at a particular year level from Year 4 to Year 10 and each with its own set of unique items. The use of multiple choice items would be retained, but because there were no plans to use overlapping items in the published test forms, it was decided not to construct parallel tests. A two year development process of item piloting, item review, and national trials was mapped out.

The test specification documents were used as a basis for writing a pool of new items and reviewing the 400 existing items. Of the existing items, 23 were retained virtually unchanged, while a further 52 were modified. A national panel of New Zealand mathematics education experts reviewed all the potential items (new and revised) for their fit to the curriculum, their use of language and the choice of distracters.

Items that were accepted by the review process were then piloted with a small number of students. Item statistics, such as the percentage of students answering the item correctly (the item facility) and the point biserial coefficient (the correlation between scores on an item and scores on the whole test form), were used to help select items for further development. Items that performed poorly in piloting were usually amended and piloted again before a decision was made whether or not to include them in any further work.

The national trials

The next stage of the development process involved two national trials. The first, held in October 2004 became known simply as the national trial. It involved approximately 250 students at each year level and was used to collect data on the performance of the items, including their ability to target particular year groups. The first national trial also helped determine what further item development work was required and presented an opportunity to develop systems to be used in the second and much larger national trial, which was called the norming trial. The norming trial was held in March, 2005 and involved close to 2000 students at each year level.

The schools involved in both trials were chosen as part of stratified random samples. The stratification variables used were school type and decile level. The samples were selected in three parts: a separate sample for Years 4–6, Years 7–8 and Years 9–10. The number of students in the national sample by strata is shown in Table 1. Each school selected as part of the sample was asked to provide one class of students at each year level. Some schools provided more than one class per year level.

Year Level	School Type	Decile1–2	Decile 3–8	Decile 9–10
4-6	Full-Primary	139	1046	523
	Contributing	472	1164	955
	Composite	31	30	246
7-8	Full-Primary	263	1207	725
	Intermediate	158	1004	292
9-10	Secondary 7-15	0	309	256
	Secondary 9-15	243	1146	940
	Composite	0	0	157
		1306	5906	4094

Table 1 Numbers of Students in the Norming Trial by Strata

Both trials involved the construction of seven core test forms from the bank of items, with each form targeted at a particular year level. To ensure the test forms could be linked across year levels, six hybrid forms were also constructed by combining half of the items in each core form with half of the items from the core form for the year level above. This meant for instance, that hybrid form 2A (Test 8 in Figure 1) was prepared by combining half of the items from Core Form 1 (Test 1) with half of the items from Core Form 2 (Test 2).

In addition to the core and hybrid forms, three more forms containing a mixture of NZCER items and items from ACER's version of PAT Mathematics were also prepared. These forms were designed to collect data to equate ACER's tests with their NZCER counterparts. The three equating forms were targeted at Year 4, Year 7 and Year 10 respectively. A similar design was adopted in the collection of Australian norming data for the ACER PATMaths tests. NZCER items were included in three of the ACER norming tests.

In total 16 test forms where used to collect trial data. Figure 1 below shows the data collection design used for the two national trials. The numbers provided relate to the norming trial.

Test	Year	Items																	N
1	4	Y4 A	Y4 B																1800
2	5			Y5 A	Y5 B														900
3	6					Y6 A	Y6 B												900
4	7							Y7 A	Y7 B										900
5	8									Y8 A	Y8 B								900
6	9											Y9 A	Y9 B						900
7	10													Y10 A	Y10 B				900
8	5		Y4 B	Y5 A															900
9	6				Y5 B	Y6 A													900
10	7						Y6 B	Y7 A											900
11	8								Y7 B	Y8 A									900
12	9										Y8 B	Y9 A							900
13	10												Y9 B	Y10 A					900
14	4	Y4 A														ACER Y4			500
15	7							Y7 A									ACER Y7		500
16	10														Y10 B			ACER Y10	500
Number of students per item		2300	2700	1800	1800	1800	1800	2300	1800	1800	1800	1800	1800	1800	1400	500	500	500	

Total number of students per year level

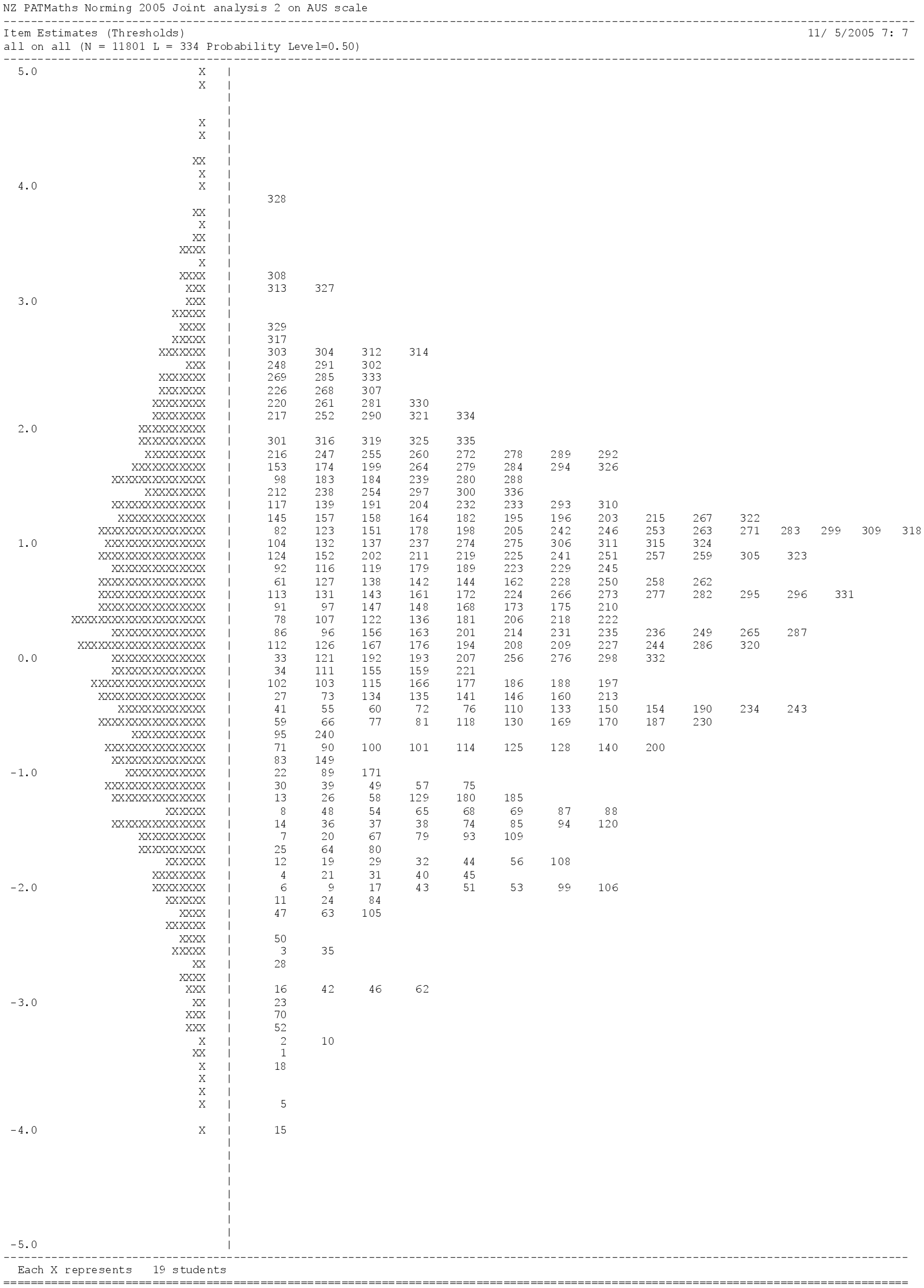
Year	N
4	2300
5	1800
6	1800
7	2300
8	1800
9	1800
10	2300

Figure 1: NZCER PAT Mathematics Norming Design

Item calibration

After each trial, student data from each test form was initially analysed separately. The computer program Quest (Adams, 1996) was used to calibrate the items onto a RM scale. A fit analysis, examining each item in the light of both statistical and graphical indicators showed how well the data fitted the measurement model and exposed any items that did not perform as well as expected. Only a handful of items were excluded from the final tests on the basis of their performance in the trials.

After the initial separate analysis, a joint analysis was carried out to locate all the items from the tests on a common scale. The common item equating design used in the collection of data made this possible. The use of hybrid forms meant that students from each year level were administered some items that had also been administered at adjacent year levels. All items used in the trial could therefore be calibrated in relation to each other on the same scale. Figure 2 shows a graphical representation of the item calibration provided by the Quest program. The locations of the items are displayed on the right, with each item identified by an analysis number. The scale locations of the students are shown on the left of the display. In this display an ‘X’ represents 19 students.



```
NZ PATMaths Norming 2005 Joint analysis 2 on AUS scale
---------------------------------------------------------------------------------------------------------------------------------------
Item Estimates (Thresholds)                                                                                            11/ 5/2005 7: 7
all on all (N = 11801 L = 334 Probability Level=0.50)
---------------------------------------------------------------------------------------------------------------------------------------
  5.0                        X   |
                             X   |
                                 |
                                 |
                             X   |
                             X   |
                                 |
                            XX   |
                             X   |
  4.0                        X   |
                                 |     328
                            XX   |
                             X   |
                            XX   |
                          XXXX   |
                             X   |
                          XXXX   |     308
                           XXX   |     313    327
  3.0                      XXX   |
                         XXXXX   |
                          XXXX   |     329
                         XXXXX   |     317
                       XXXXXXX   |     303    304    312    314
                           XXX   |     248    291    302
                       XXXXXXX   |     269    285    333
                       XXXXXXX   |     226    268    307
                      XXXXXXXX   |     220    261    281    330
                      XXXXXXXX   |     217    252    290    321    334
  2.0               XXXXXXXXXX   |
                    XXXXXXXXXX   |     301    316    319    325    335
                     XXXXXXXXX   |     216    247    255    260    272    278    289    292
                   XXXXXXXXXXX   |     153    174    199    264    279    284    294    326
                XXXXXXXXXXXXXX   |      98    183    184    239    280    288
                     XXXXXXXXX   |     212    238    254    297    300    336
                XXXXXXXXXXXXXX   |     117    139    191    204    232    233    293    310
                 XXXXXXXXXXXXX   |     145    157    158    164    182    195    196    203    215    267    322
              XXXXXXXXXXXXXXXX   |      82    123    151    178    198    205    242    246    253    263    271   283   299   309   318
  1.0          XXXXXXXXXXXXXXX   |     104    132    137    237    274    275    306    311    315    324
              XXXXXXXXXXXXXXXX   |     124    152    202    211    219    225    241    251    257    259    305    323
                XXXXXXXXXXXXXX   |      92    116    119    179    189    223    229    245
              XXXXXXXXXXXXXXXX   |      61    127    138    142    144    162    228    250    258    262
              XXXXXXXXXXXXXXXX   |     113    131    143    161    172    224    266    273    277    282    295    296    331
              XXXXXXXXXXXXXXXX   |      91     97    147    148    168    173    175    210
          XXXXXXXXXXXXXXXXXXXX   |      78    107    122    136    181    206    218    222
                XXXXXXXXXXXXXX   |      86     96    156    163    201    214    231    235    236    249    265    287
           XXXXXXXXXXXXXXXXXXX   |     112    126    167    176    194    208    209    227    244    286    320
  0.0           XXXXXXXXXXXXXX   |      33    121    192    193    207    256    276    298    332
                XXXXXXXXXXXXXX   |      34    111    155    159    221
             XXXXXXXXXXXXXXXXX   |     102    103    115    166    177    186    188    197
              XXXXXXXXXXXXXXXX   |      27     73    134    135    141    146    160    213
                 XXXXXXXXXXXXX   |      41     55     60     72     76    110    133    150    154    190    234    243
              XXXXXXXXXXXXXXXX   |      59     66     77     81    118    130    169    170    187    230
                   XXXXXXXXXXX   |      95    240
               XXXXXXXXXXXXXXX   |      71     90    100    101    114    125    128    140    200
                XXXXXXXXXXXXXX   |      83    149
 -1.0             XXXXXXXXXXXX   |      22     89    171
               XXXXXXXXXXXXXXX   |      30     39     49     57     75
                XXXXXXXXXXXXXX   |      13     26     58    129    180    185
                        XXXXXX   |       8     48     54     65     68     69     87     88
                XXXXXXXXXXXXXX   |      14     36     37     38     74     85     94    120
                    XXXXXXXXXX   |       7     20     67     79     93    109
                    XXXXXXXXXX   |      25     64     80
                        XXXXXX   |      12     19     29     32     44     56    108
                      XXXXXXXX   |       4     21     31     40     45
 -2.0                 XXXXXXXX   |       6      9     17     43     51     53     99    106
                        XXXXXX   |      11     24     84
                          XXXX   |      47     63    105
                        XXXXXX   |
                          XXXX   |      50
                         XXXXX   |       3     35
                            XX   |      28
                          XXXX   |
                           XXX   |      16     42     46     62
 -3.0                       XX   |      23
                           XXX   |      70
                           XXX   |      52
                             X   |       2     10
                            XX   |       1
                             X   |      18
                             X   |
                             X   |
                             X   |       5
                                 |
 -4.0                        X   |      15
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
 -5.0                            |
--------------------------------------------------------------------------------------------------------------------------------------
  Each X represents   19 students
======================================================================================================================================
```

Figure 2: Calibration of the items on the NZCER/ACER PAT Mathematics scale

Equating ACER and NZCER's PAT Mathematics scales

The inclusion of ACER items in the collection of norming data in New Zealand allowed the calibration of ACER items on the newly constructed NZCER scale. The equating of the NZCER scale with the ACER scale was based on a comparison between the two sets of item calibrations. A direct and consistent relationship was established, allowing a simple additive adjustment to be made to NZCER's new scale to equate it with ACER's existing scale. In effect, there was now a single NZCER/ACER continuum for the measurement of development in mathematics on an interval scale with the same measurement unit, the patm. A consequence of the single scale is that results on both the NZCER and ACER versions of PAT Mathematics may be reported with reference to Australian or New Zealand norms for any year level for which data are available.

The PAT Mathematics scale

Figure 3 shows another representation of the items calibrated onto the PAT Mathematics scale. This time the items are categorised according to their position in the final published test forms and the scale is shown in patm units. For these tests the scale shows achievement from 5 patm to 105 patm. Item 15 in Test 1 is the easiest question in the tests, and is located low on the scale at 8.6 patm. The hardest question is Item 36 in Test 7, and is accordingly located higher on the scale at 83.5 patm.

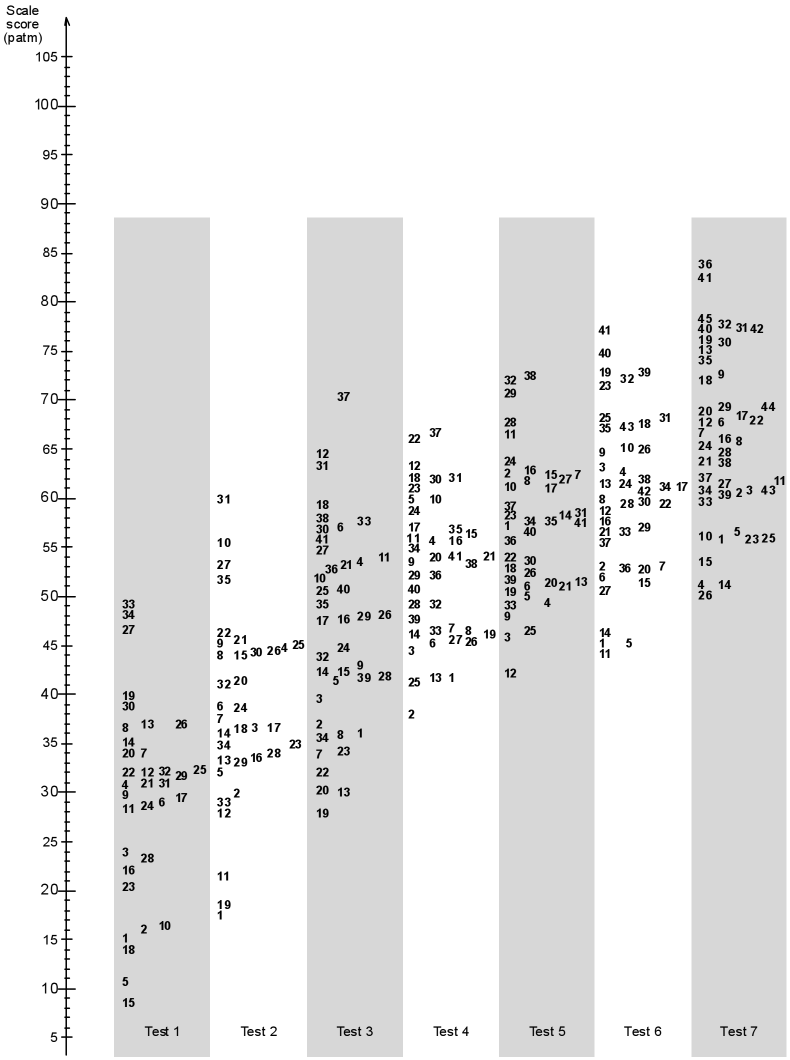


Figure 3: NZCER PAT Mathematics items by test

Properties of the PAT Mathematics scale

The choice of numerical values and the names of units to assign to an RM scale is arbitrary. This is similar to measuring temperature, where degrees Celsius and degrees Fahrenheit can both be used to mark the scale. The unit of measurement used to express locations on the PAT Mathematics scale, patm, is the result of a linear transformation of the probabilistic unit used in the Rasch model to estimate locations on the scale. This unit is called the logit. The transformation from logit to patm is:

PAT Mathematics scale score (patm) = 10 × logit + 50

It follows that:

- 1 patm unit is equivalent to 0.1 logit;
- the mean location of all the items calibrated onto the scale is set to 50 patm;
- a student located at any point on the scale (*b* patm units, for example) is expected to answer correctly 50% of the items at that location, just under 30% of the items 10 patm units higher, and just over 10% of the items 20 patm units higher. Similarly, the same student located at *b* patm is expected to answer correctly just over 70% of the items 10 patm units lower, and just under 90% of the items located 20 patm units lower than *b*.

Examples of items calibrated on the PAT Mathematics scale

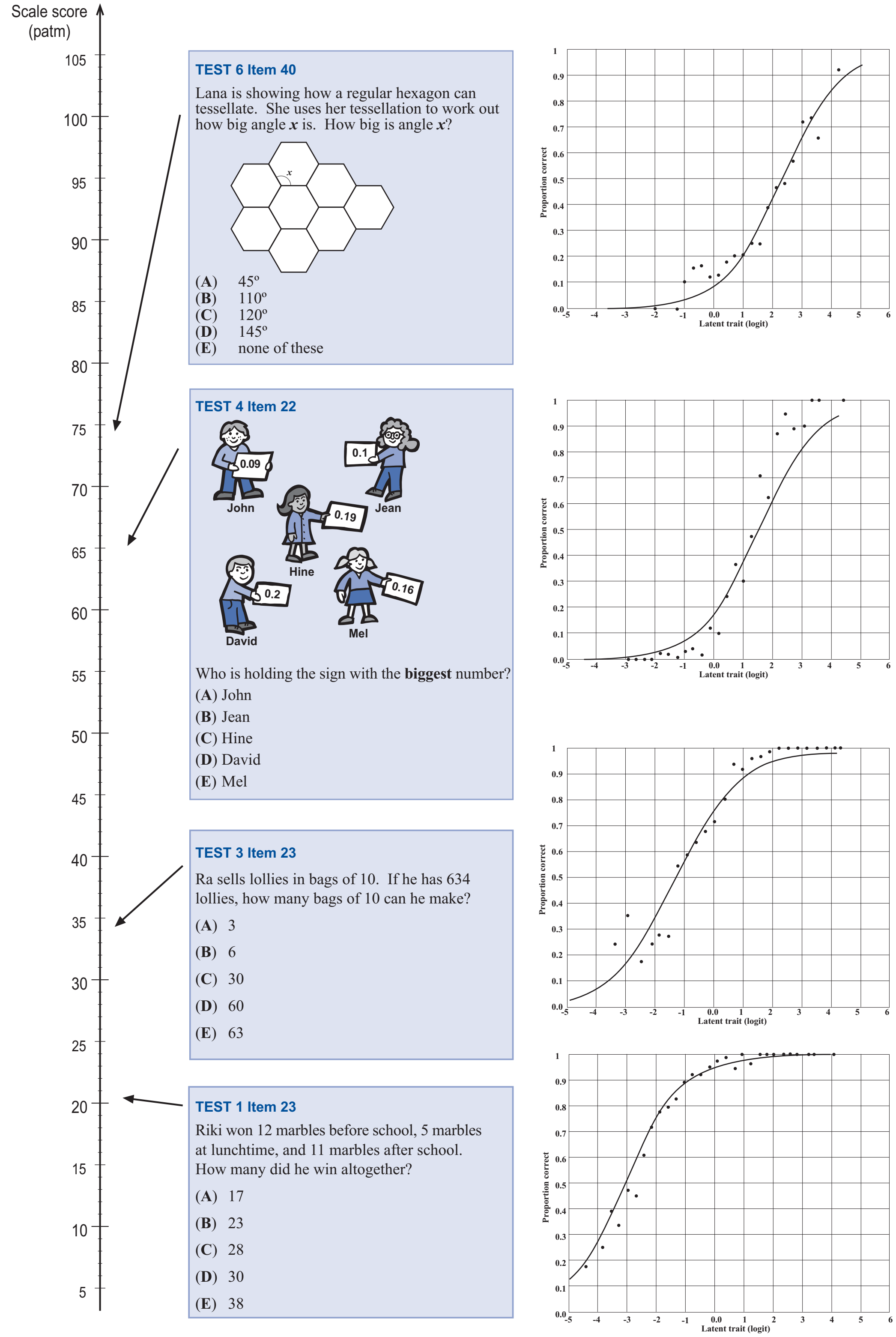
Figure 4 shows four examples of PAT Mathematics items and their location on the PAT Mathematics scale. As a group the items illustrate the increasing sophistication in mathematical knowledge and skill required to successfully answer items at increasing scale locations. Item 23 from Test 1 for instance, is the example item shown with the lowest location on the scale (20 patm units). This item requires a basic understanding of addition and can be answered successfully by using a "counting on" strategy. If a part–whole strategy or an addition algorithm is used, the number of ones accumulated in the addition does not have to be renamed as a ten. The next highest item, Item 23 from Test 3 at 34 patm units requires a more sophisticated understanding of number. Here a student must know how to decompose a three digit number, including how to rename the six hundreds as sixty tens. More sophistication again is required to answer Item 22, which is located at 64 patm units. For this item students need to be able to order decimal numbers and in particular appreciate that whole number thinking is not appropriate when dealing with decimal representations. Finally, the example item shown with the highest scale location (Item 40 from Test 6 at 74 patm units) not only involves a developed understanding of number, but also the ability to recognize and exploit geometric relationships. For instance, a student might recognise that the angle shown is one third of a full three hundred and sixty degree turn.

By analyzing the items in terms of their content and their locations on the PAT Mathematics scale, test users can begin to appreciate how the mathematical construct underlying PAT mathematics progresses.

Item characteristic curves

Each of the example items in Figure 4 is displayed next to its item characteristic curve (ICC). Produced from the norming trial data using the program Conquest (Wu, 1998) ICCs such as these were one of the indicators used to analyze the fit of each item to the RM model.

The smooth curve of each ICC indicates the *expected* proportions of students to answer the item correctly at various locations on the scale. The dots represent the *observed* proportions of students who answered the item correctly at various locations on the scale. Differences between expected and observed proportions indicate misfit. All of these items displayed in Figure 4 show acceptable fit. Item 22 from Test 4 does show some misfit. This item is discriminating more than expected, with greater proportions of high achieving students and lesser proportions of low achieving students answering this item correctly than was expected.



TEST 6 Item 40

Lana is showing how a regular hexagon can tessellate. She uses her tessellation to work out how big angle ***x*** is. How big is angle ***x***?

(**A**) 45º
(**B**) 110º
(**C**) 120º
(**D**) 145º
(**E**) none of these

TEST 4 Item 22

Who is holding the sign with the **biggest** number?

(**A**) John
(**B**) Jean
(**C**) Hine
(**D**) David
(**E**) Mel

TEST 3 Item 23

Ra sells lollies in bags of 10. If he has 634 lollies, how many bags of 10 can he make?

(**A**) 3
(**B**) 6
(**C**) 30
(**D**) 60
(**E**) 63

TEST 1 Item 23

Riki won 12 marbles before school, 5 marbles at lunchtime, and 11 marbles after school. How many did he win altogether?

(**A**) 17
(**B**) 23
(**C**) 28
(**D**) 30
(**E**) 38

Figure 4: Example PAT Mathematics Items with Scale Locations and Item Characteristic Curves

Developing year level profiles of achievement

Once the items were calibrated on a single scale it became possible to locate student achievement at each year level on the same scale. Again, the computer program Quest was used, this time to calculate the best estimate of student locations on the scale given the calibrations of the items used in each test form. The random sampling used in the trials meant that the resulting distributions of student achievement at each year level represented a national profile of achievement in mathematics from Year 4 to Year 10.

This methodology used to develop student profiles of achievement or norms represents a radical departure from CTT. As discussed previously, student norms in CTT can only be understood in terms of the actual test that is used to collect the norming data. As a result, a trial form used in CTT must be the same or nearly the same as the one that is ultimately published. In RM the location of students on the scale does not depend on which test forms were used to collect the data or which forms are finally published. Norms can be developed using data from multiple test forms and then applied to any other form constructed from items calibrated onto the scale. As noted, for PAT Mathematics a minimum of two different test trial forms were used to collect student data at each year level (a core form, a hybrid form, and in the case of Years 4, 7 and 8 a joint NZCER/ACER equating form). As will be discussed later, changes were made to the core forms, before final test forms targeted at particular year levels were constructed for publishing.

Figure 5 shows the distribution of student achievement by year level on the PAT Mathematics scale. As can be seen, the mean scale score for each year level increases at a fairly constant rate.

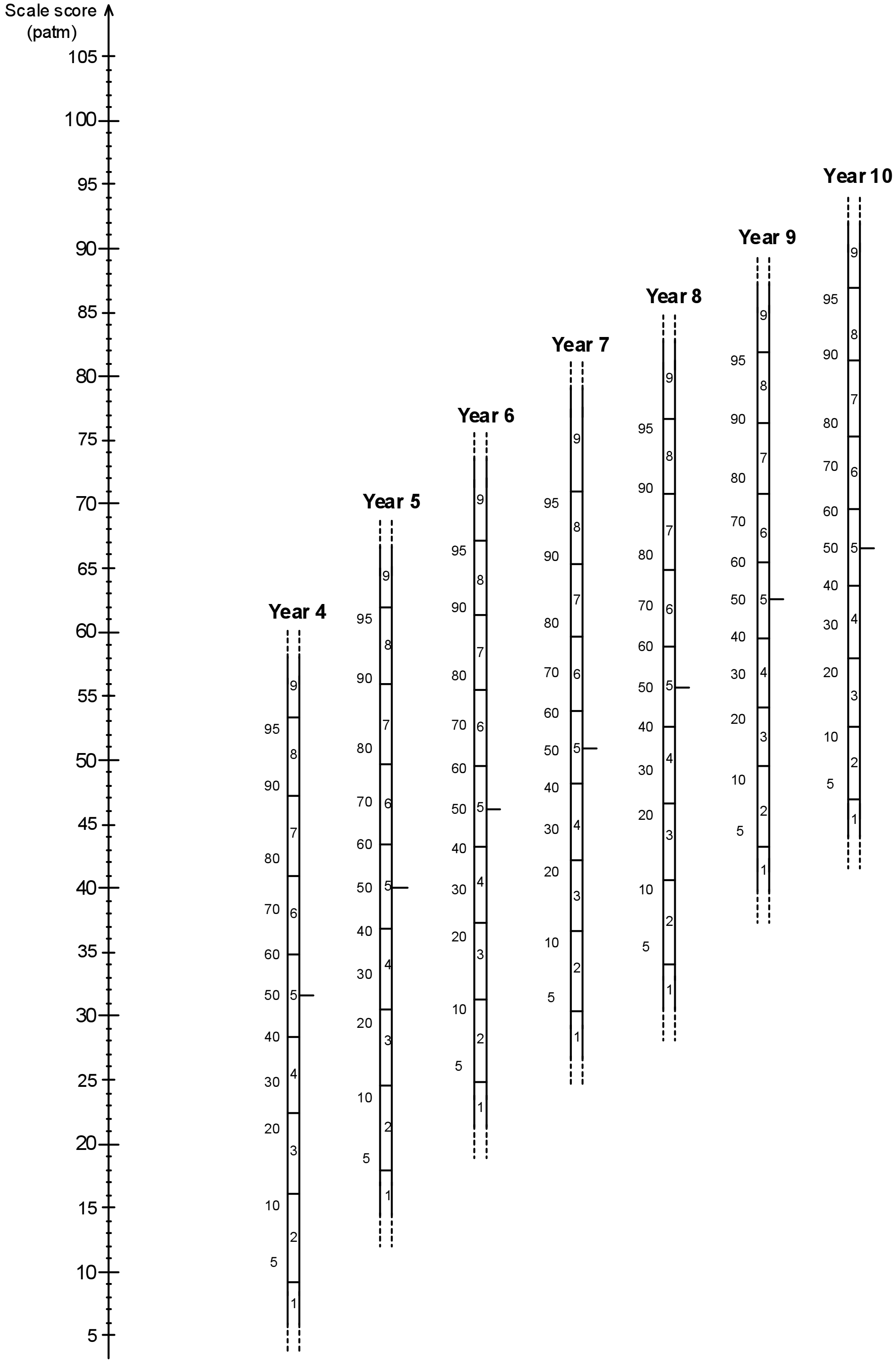


Figure 5: Distribution of student achievement by year level

Differential item functioning

During the revision of PAT Mathematics careful attention was given to gender and ethnic bias.

First in what is essentially an exercise in judgement, the test development team and the national review panel reviewed each item to evaluate the extent of possible bias in language or presentation. A member of NZCER's Te Wananga Kura Kaupapa Maori also examined test forms for evidence of cultural bias. Any suspect items were either modified or excluded.

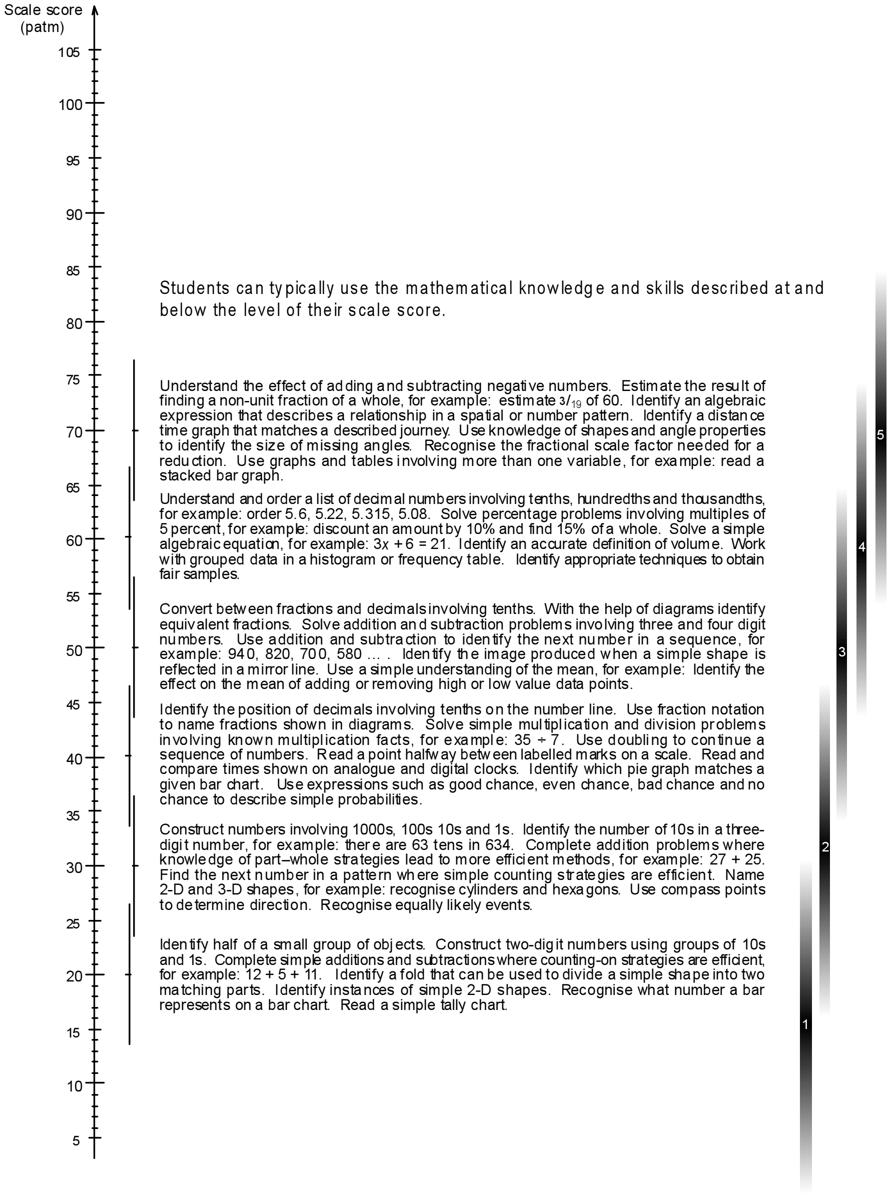
The second approach was statistical; it exposes or detects differential item functioning (DIF) for two groups of students. Individual items were examined to see whether performance on any of the items was different for any particular subgroups within the national samples. DIF analysis was carried out according to both ethnic group and gender. These analyses identified a handful of items that appeared to function slightly differently for different genders or ethnic groups. However, these differences at the item level did not have a noticeable affect on the overall test performance. Therefore no items were excluded on the basis of DIF.

Describing the scale

The ability of RM to locate both test items and student achievement on the same scale made it possible to describe student achievement at different locations on the PAT Mathematics scale in terms of the types of mathematical knowledge and skills tested by items situated at the same locations on the scale. To develop this description, items were first divided into their different content categories, for instance Number Knowledge and Number Strategies. They were then further divided into groups of items involving similar conceptual material, for instance items dealing with fractions. Each of these subgroups was then ordered according to their locations on the Rasch scale and items with similar scale locations examined for common features. The features were described and the descriptions attached to the appropriate location on the scale. This process resulted in six levels of descriptions on the scale for each major content area. A sample of these descriptions is shown in Figure 6.

Locating curriculum levels on the PAT Mathematics scale

The use of RM also made it possible to show how items representing different levels of the New Zealand Mathematics Curriculum were distributed on the scale. First, each item was allocated to the curriculum level it was judged to represent. The curriculum levels of the items were then plotted against their scale locations, making it possible to show where items from different curriculum levels appeared on the scale. As expected, there was no strict delineation between items at different curriculum levels. Instead the curriculum levels overlapped, with the mean item difficulty for each curriculum level rising steadily. The pattern of curriculum levels can be seen to the right of the descriptors in Figure 6.



Students can typically use the mathematical knowledge and skills described at and below the level of their scale score.

Understand the effect of adding and subtracting negative numbers. Estimate the result of finding a non-unit fraction of a whole, for example: estimate $3/_{19}$ of 60. Identify an algebraic expression that describes a relationship in a spatial or number pattern. Identify a distance time graph that matches a described journey. Use knowledge of shapes and angle properties to identify the size of missing angles. Recognise the fractional scale factor needed for a reduction. Use graphs and tables involving more than one variable, for example: read a stacked bar graph.

Understand and order a list of decimal numbers involving tenths, hundredths and thousandths, for example: order 5.6, 5.22, 5.315, 5.08. Solve percentage problems involving multiples of 5 percent, for example: discount an amount by 10% and find 15% of a whole. Solve a simple algebraic equation, for example: $3x + 6 = 21$. Identify an accurate definition of volume. Work with grouped data in a histogram or frequency table. Identify appropriate techniques to obtain fair samples.

Convert between fractions and decimals involving tenths. With the help of diagrams identify equivalent fractions. Solve addition and subtraction problems involving three and four digit numbers. Use addition and subtraction to identify the next number in a sequence, for example: 940, 820, 700, 580 … . Identify the image produced when a simple shape is reflected in a mirror line. Use a simple understanding of the mean, for example: Identify the effect on the mean of adding or removing high or low value data points.

Identify the position of decimals involving tenths on the number line. Use fraction notation to name fractions shown in diagrams. Solve simple multiplication and division problems involving known multiplication facts, for example: 35 ÷ 7. Use doubling to continue a sequence of numbers. Read a point halfway between labelled marks on a scale. Read and compare times shown on analogue and digital clocks. Identify which pie graph matches a given bar chart. Use expressions such as good chance, even chance, bad chance and no chance to describe simple probabilities.

Construct numbers involving 1000s, 100s 10s and 1s. Identify the number of 10s in a three-digit number, for example: there are 63 tens in 634. Complete addition problems where knowledge of part–whole strategies lead to more efficient methods, for example: 27 + 25. Find the next number in a pattern where simple counting strategies are efficient. Name 2-D and 3-D shapes, for example: recognise cylinders and hexagons. Use compass points to determine direction. Recognise equally likely events.

Identify half of a small group of objects. Construct two-digit numbers using groups of 10s and 1s. Complete simple additions and subtractions where counting-on strategies are efficient, for example: 12 + 5 + 11. Identify a fold that can be used to divide a simple shape into two matching parts. Identify instances of simple 2-D shapes. Recognise what number a bar represents on a bar chart. Read a simple tally chart.

Figure 6 The PAT Mathematics scale, with sample descriptors

Preparing the published test forms

The make up of the final test forms for publishing was decided after the collection of the norming data. At each year level, items were selected to represent the content categories of number knowledge, number strategies, geometry and measurement, statistics and algebra. To ensure the final test forms targeted the distribution of student achievement at each year level, the match between an item’s location on the scale and the distribution of achievement for the year level being tested was carefully considered before the item was included in a form. Once a final selection of items had been made, a score conversion table was constructed for each test form to allow test scores (raw scores) to be converted to scale scores in patm units.

Figure 7 uses percentage test scores to show the relative difficulty of the seven published tests. Percentage scores for each test are shown, along with their corresponding scale scores. As the number of items in each test differs, percentage test scores provide a useful method of comparison.

Figure 7 shows that Test 1 is the easiest test, with a 50% test score converting to 29.5 patm. Test 2 is slightly more difficult, with a 50% score that is 9 patm units higher than Test 1. As can be seen, the tests become progressively more difficult. The most difficult test is Test 7, with a 50% score at approximately 66 patm.

The difference in difficulty between the easiest and most difficult of the tests is considerable. A percentage test score of 10 on Test 7, for instance, is the same level of achievement on the PAT Mathematics scale as a percentage score of almost 72 on Test 1.

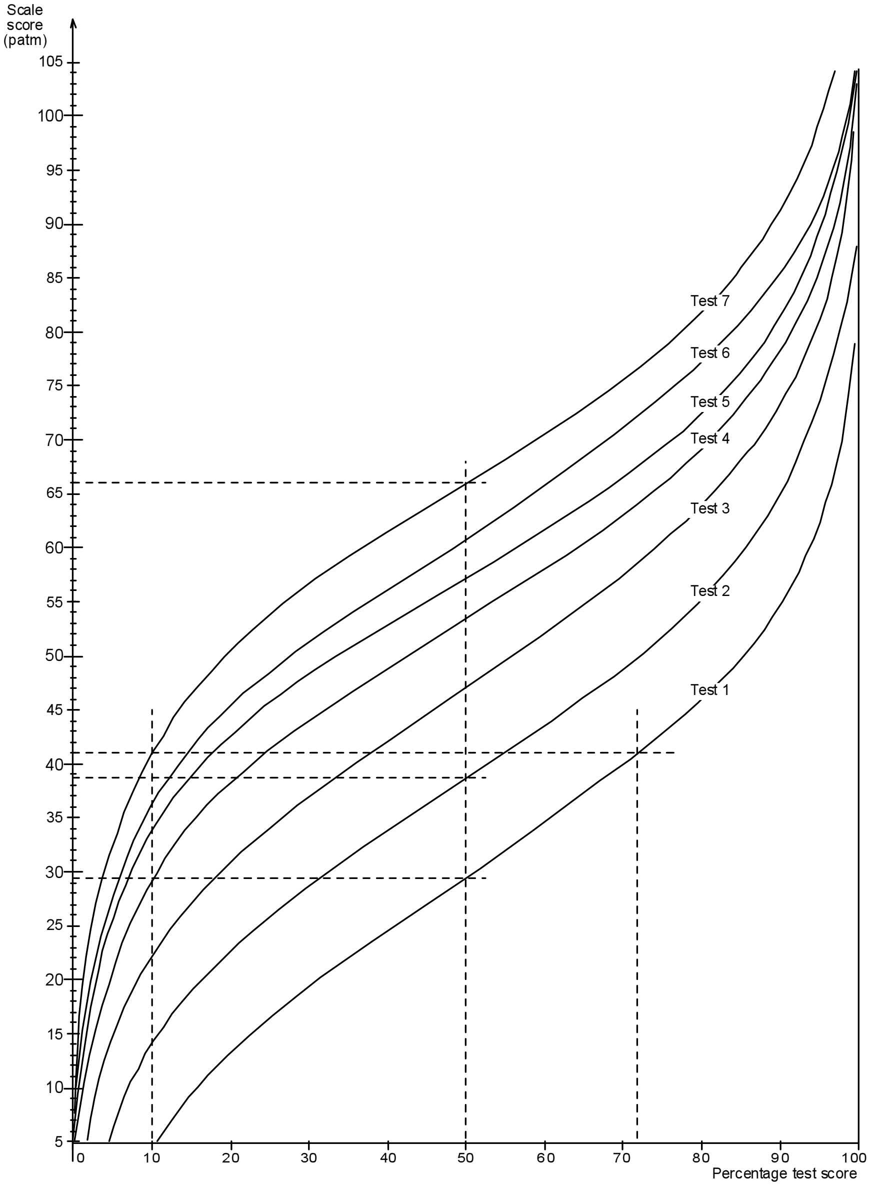


Figure 7: Test Characteristic Curves

Reporting test results

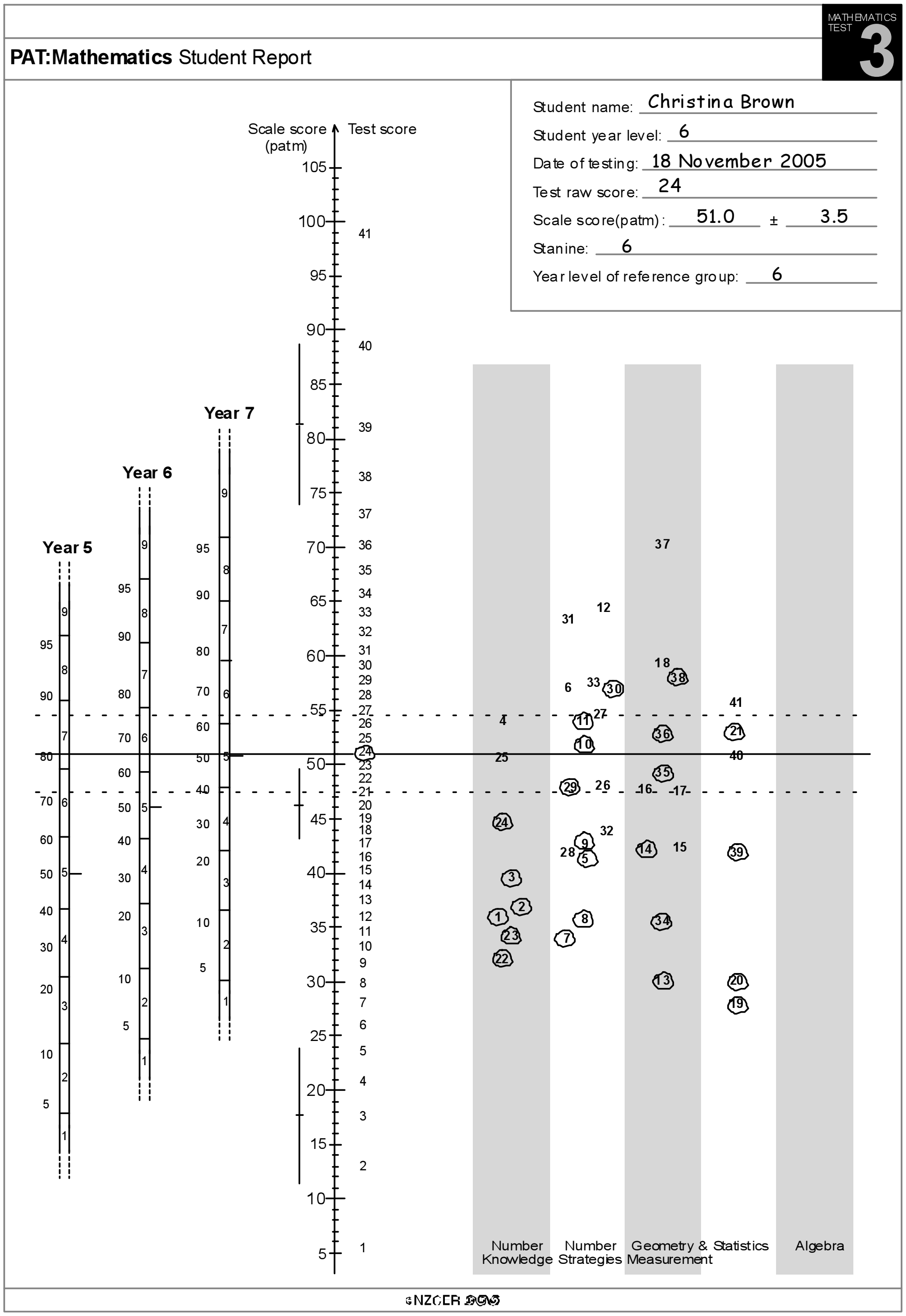
To help teachers analyse student performance, a student report for each final test form was developed. Each report is made up of two parts. The first part allows the teacher to locate the student on the scale and to compare this achievement with the location of the test items and with student norms at three adjacent year levels. The second part allows a student's location on the scale to be compared with a series of scale descriptors for each content area assessed in the test.

Figure 8 shows an example of the first part of the report for a fictitious student Christina Brown. Christina has scored 24 in Test 3 and a solid line has been drawn across the page at the location of her scale score (51.0 patm). At the left of the report the line intersects the three strip graphs that show the national profiles of achievement at Years 5, 6 and 7. This shows that for Year 6 students, Christina's achievement places her in stanine 6 (slightly above average).

On the right of the report the line separates the test items into two groups: those above Christina's location on the scale and those below. All of the 24 items that Christina has answered correctly have been circled. This can be used to show unexpected answering patterns. For instance, if items have been answered correctly that are well above Christina's achievement level, or incorrectly that are well below her achievement level, Christina's pattern of answering is consistent with her scale location. A significant amount of deviation from the expected pattern would indicate that the measurement involves misfit.

The two dotted lines in Figure 8 have been drawn to indicate the precision of the measurement. These lines provide a range within which we can be reasonably sure that Christina's true location on the scale actually lies. The level of precision depends on the test score. A high or low test score results in larger measurement errors than scores in the midrange for a test. The vertical line segments shown on the left of the scale indicate the level of precision possible at different scale locations for this particular test.

Figure 9 shows the second part of Christina's report. Here the descriptors for the Number Knowledge component of the test are displayed. Other descriptors are also available for the other content categories in the test. Again a line has been drawn across the page at the level of Christina's scale score. The blocks of descriptors that are below the line refer to items that given Christina's scale score, we can expect to be answered correctly. She should also be able to give correct answers for about 50% of the items described by the block of descriptors situated at the same level as her scale score. Any blocks of descriptors that are located above Christina's scale score refer to the knowledge and skills involved in questions that she is less likely to be able to answer successfully.



MATHEMATICS TEST 3

PAT:Mathematics Student Report

Student name: Christina Brown

Student year level: 6

Date of testing: 18 November 2005

Test raw score: 24

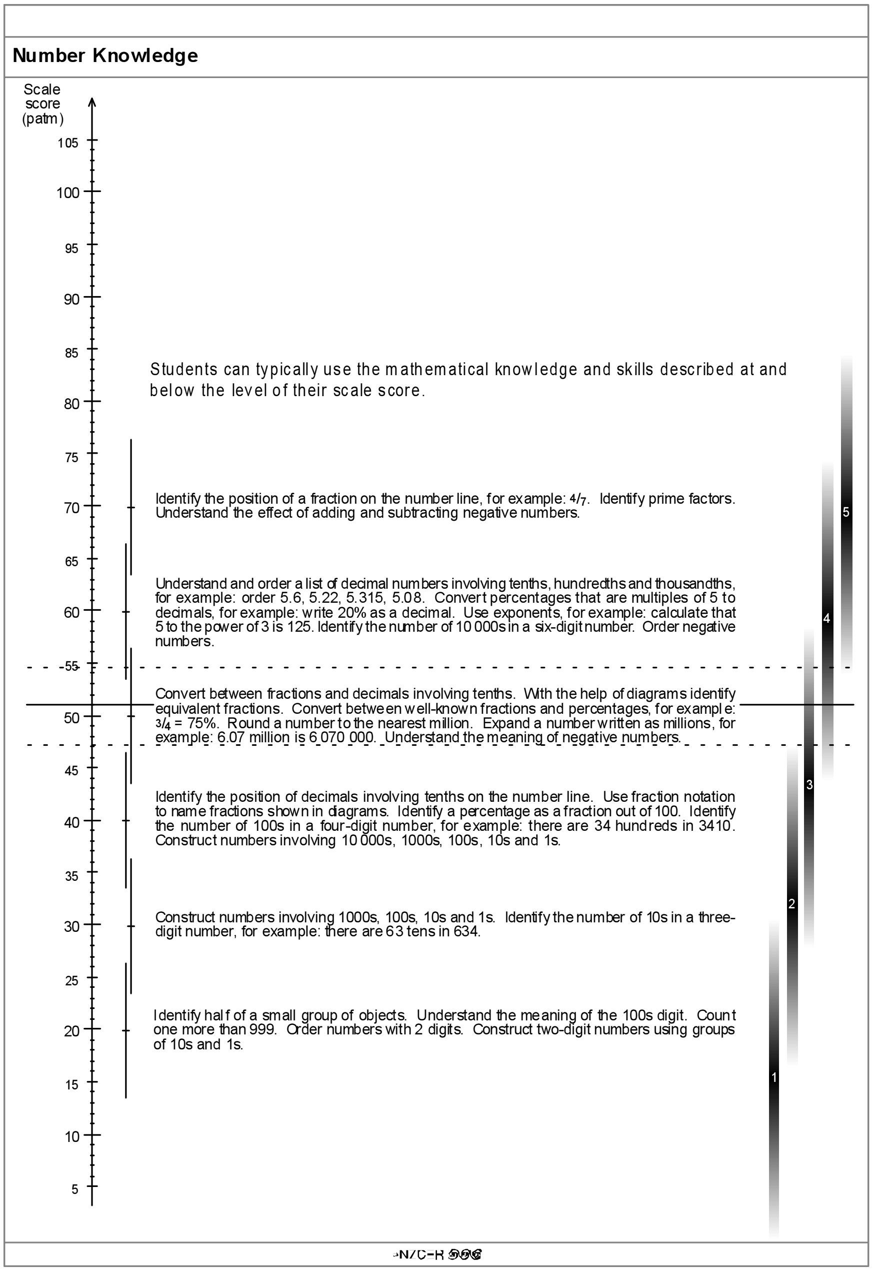
Scale score(patm): 51.0 ± 3.5

Stanine: 6

Year level of reference group: 6

©NZCER 2005

Figure 8: Sample Diagnostic Graphical Student Report Part 1



Students can typically use the mathematical knowledge and skills described at and below the level of their scale score.

Identify the position of a fraction on the number line, for example: 4/7. Identify prime factors. Understand the effect of adding and subtracting negative numbers.

Understand and order a list of decimal numbers involving tenths, hundredths and thousandths, for example: order 5.6, 5.22, 5.315, 5.08. Convert percentages that are multiples of 5 to decimals, for example: write 20% as a decimal. Use exponents, for example: calculate that 5 to the power of 3 is 125. Identify the number of 10 000s in a six-digit number. Order negative numbers.

Convert between fractions and decimals involving tenths. With the help of diagrams identify equivalent fractions. Convert between well-known fractions and percentages, for example: 3/4 = 75%. Round a number to the nearest million. Expand a number written as millions, for example: 6.07 million is 6 070 000. Understand the meaning of negative numbers.

Identify the position of decimals involving tenths on the number line. Use fraction notation to name fractions shown in diagrams. Identify a percentage as a fraction out of 100. Identify the number of 100s in a four-digit number, for example: there are 34 hundreds in 3410. Construct numbers involving 10 000s, 1000s, 100s, 10s and 1s.

Construct numbers involving 1000s, 100s, 10s and 1s. Identify the number of 10s in a three-digit number, for example: there are 63 tens in 634.

Identify half of a small group of objects. Understand the meaning of the 100s digit. Count one more than 999. Order numbers with 2 digits. Construct two-digit numbers using groups of 10s and 1s.

Figure 9: Sample Diagnostic Graphical Student Report Part 2

Discussion

The Rasch measurement methodology adopted in the norming of the PAT Mathematics tests has considerable advantages over the traditional norming (TN) methodology (e.g. de Lemos, 2000) in which there is no calibration of all items in the tests onto a common measurement scale. Instead, each test is normed on separate samples of students and independently from the other tests in the study. Tests that are assumed to be of the same difficulty (parallel forms) and forms containing common items are often included in the published version of these norming studies (e.g. ACER, 1986) to provide some form of comparability of results from various tests.

The calibration of all items on the same Rasch measurement scale has the following advantages:

1. The "sample-independent" property of the estimation of item location on the scale allows all data collected for each item, including data from students in different year levels and data from different tests in which an item was included, to be used to locate it on the scale, thus reducing measurement error. In TN there is no estimation of item location on a measurement scale. Only the sample-dependent item facility is used as an indicator of item difficulty. More than one facility is reported for items included in different forms.

2. The "instrument-independent" property of the estimation of student location on the scale allows all data collected from students in a year level, including data obtained with different tests, to be used in the norming of test forms for each year level, thus reducing sampling errors. All data collected, including the data on equating forms, are used to estimate student location on the scale. Each form can be normed with data from all year levels for which data were collected, including year levels to which a particular form was not administered. In TN each test is normed only on the data collected with that test.

3. The calibration of all items on the same scale allows a thorough analysis of fit of the data to the Rasch model and a qualitative description of the scale showing the construct as a developmental continuum of mathematical skills. Both normative and formative reporting are possible. Both questions "how well has a student achieved compared to some meaningful sample and what does it mean substantively for a student to have achieved a particular score on a test?" can be answered. In TN studies only normative reporting in terms of percentiles, stanines and similar statistical indicators is possible (e.g. de Lemos, 2000).

4. The calibration of all items on the same scale allows distribution of student achievement in a year level to be compared with other year levels on an interval scale showing growth trajectories across year levels. Only distributions of student achievement in terms of test scores that are not comparable are available in TN, thus it is not possible to report growth from year level to year level.

5. The calibration of all items on the same scale allows comparability of the relative difficulty of all tests. The Rasch difficulty of tests, which is independent of the location of student achievement on the scale can be expressed in logits. In TN the difficulty of tests cannot be compared independently of student achievement.

6. New test forms can be calibrated on the constructed scale and the already existing reference distributions of student achievement by year level can be used to norm new test forms or any set of items from the item bank. After a collection of equating data with either common items or common people, the new items can be located on the existing scale. A score equivalence table can be compiled for each new test form and thus normed on the existing distributions of student locations by year level (e.g. Stephanou, 2006). In TN new

norming data must be collected to norm a new test form because none of the existing data can be used for the new norming.

7. The Rasch measurement scale allows the estimation of the expected facility for each item according to each year level sample for which data were collected (Darr, 2006). In TN only the item facility observed for the year levels for which data were collected can be reported.

The norming methodology described in this paper has its precursor in a norming study of reading comprehension tests with cloze type items, the "Tests of Reading Comprehension" (TORCH) which was completed in Western Australia (Mossenson, 1987). The first state-wide numeracy and literacy testing program in Australia (Masters, 1990) that made use of the Rasch methodology was followed by the adoption of the new methodology in the other Australian states and territories, and by developments that are leading to national tests in 2008 for monitoring student achievement at years 3, 5, 7 and 9. The wide acceptance of the methodology that followed is due to improvements in the following areas: analysis of the data and reporting of results (e.g. the W.A. Monitoring Standards in Education public reports that are available on line at http://www.det.wa.edu.au/education/mse/reporting.html), the intuitive explanation of underlying objective measurement ideas, and the understanding of the advantages of methodologies based on scale scores over those based on raw scores. PAT Mathematics is an example of this development in educational measurement.

References

ACER (1984). *Progressive Achievement Tests in Mathematics, Teachers Handbook.* ACER Press, Hawthorn, Australia.

ACER (1986). *Progressive Achievement Tests in Reading, Teachers Handbook Second Edition.* ACER, Hawthorn, Australia.

ACER (1997). *Progressive Achievement Tests in Mathematics (PATMaths). Revised, Teacher's manual.* ACER Press, Camberwell, Australia

ACER (2001). *PAT-R Progressive Achievement Tests in Reading: Comprehension and Vocabulary, Teacher's manual.* ACER Press, Camberwell, Australia.

Adams, R.J., & Khoo, S.T. (1996). *QUEST: The Interactive Test Analysis System.* Version 2.1. ACER, Hawthorn, Australia.

Andrich, D., (1988) *Rasch Models for Measurement*. SAGE University paper. SAGE Publications, CA, USA.

Darr, C., Neill, A., Stephanou, A., (2006), *PAT Mathematics Progressive Achievement Tests in Mathematics, Teacher Manual Third Edition.* Wellington: NZCER

de Lemos, M., (2000) *Reading Progress Test Stage 1 and Stage 2 Australian Norms Supplement*, ACER Press, Camberwell, Australia.

Lindsey, J., Stephanou, A., Urbach, D., Sadler, A., (2005), *PATMaths Progressive Achievement Tests in Mathematics, Teacher Manual Third Edition*, ACER Press, Camberwell, Australia.

Masters, G., et. al., (1990), *Profiles of Learning: The Basic Skills Testing Program in New South Wales 1989*, ACER, Hawthorn, Australia.

Mossenson, L., Hill, P., Masters, G., (1987), *TORCH Tests of Reading Comprehension*, ACER, Hawthorn, Australia.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Reid, N. A. & Hughes, C. D. (1974). *Progressive Achievement Tests in Mathematics, Teacher's Manual.* Wellington: NZCER.

Reid, N. A., Croft, C. A., Elley, W. B. (1978). *Progressive Achievement Tests: Study Skills: Teachers Manual.* Wellington, NZCER.

Reid, N. A., & Elley, W. A. (h1991). *Progressive Achievement Tests of Reading, Teacher's Manual.* Wellington, NZCER.

Reid, N. A. (1993). *Progressive Achievement Tests in Mathematics, Teacher's Manual.* Wellington: NZCER.

Reid, N. A., Johnston, I. C., Elley, W. B. (1994). *Progressive Achievement Test of Listening Comprehension, Teacher's Manual.* Wellington: NZCER.

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

Stephanou, A., (2006), *PAT Maths Third Edition and I Can do Maths revised reports*, ACER Press, Camberwell Australia.

Stephanou, A., (2006), *Reading Progress Tests on the PAT-R Comprehension Scale*, ACER Press, Camberwell Australia.

Wu, M.L., Adams, R.J., Wilson, M. (1998). *ACER ConQuest*: *Generalised Item Response Modelling Software*. ACER, Camberwell Australia.