Technical manual

Progressive Achievement Test (PAT): Mathematics *refreshed*

NZCER
Rangahau Mātauranga o Aotearoa

For questions or advice

E-mail assessmentservices@nzcer.org.nz or call Assessment Services on (04) 802 1630

Introduction

This document provides technical information about the eight refreshed PAT: Mathematics tests developed by NZCER in 2021 and 2022, and released for use in 2023. The tests represent comprehensive revisions of PAT: Mathematics Tests 1 to 7 (first published in 2006) and Test 8A (first published in 2012).

The information contained in this document outlines the development process associated with the tests and describes how they were linked to the PAT: Mathematics scale. It is intended that this online document will be updated, as appropriate, to provide technical information associated with any additional refresh activities.

1. Refreshing the assessments

The refresh process began with a decision to update approximately one third of the questions in each of Tests 1 to 7 and Test 8A from the existing range of PAT: Mathematics tests. In addition, it was decided that the graphics for all questions would be updated. The aim was to produce eight tests with refreshed content aimed at specific year levels and linked to the existing PAT scale. To begin, the NZCER PAT: Mathematics refresh team reviewed each of the questions in Tests 1 to 7 and 8A, and made decisions as to whether the item was mathematically relevant, culturally authentic, and accessible. The team was guided by an internally developed NZCER cultural perspectives tool, current national and international research in equity and assessment in mathematics education, and consultation with a teacher advisory panel. Most of the teacher advisory panel were kaiako Māori and Pacific teachers who work in low-decile schools with high Māori and Pacific rolls, and have extensive experience as teachers of mathematics.

After the initial review, and in fitting with the initial intention, approximately one third of items within each of PAT: Mathematics Tests 1 to 7 and 8A were significantly changed. The items then underwent multiple review cycles with the teacher advisory panel and other experts in mathematics education. A pilot study was held in 16 schools, and a larger scale trial (the calibration trial) in 73 schools across Aotearoa New Zealand. Teacher and student feedback on the refreshed assessments was also gathered, analysed, and acted on during the pilot and trial.

As part of the refresh process, all questions used in the Computer Adaptive Test (CAT) version of PAT: Mathematics, including the refreshed content, were updated with new graphics. Careful attention was paid to designing illustrations that supported the mathematical intention of each question and that represented the cultural identities and world views of a diverse range of people and contexts in Aotearoa New Zealand.

2. Linking the refreshed assessments to the PAT: Mathematics scale

Achievement on the refreshed PAT: Mathematics tests can be reported as scale scores on the existing PAT: Mathematics scale.[1] The scale score (in patm units) provides a measure of achievement in mathematics that is comparable across all of the tests that make up the suite of PAT: Mathematics assessments (the refreshed and existing static tests, as well as tests administered as computer adaptive tests). This allows the progress of students to be monitored from late Year 3 to early Year 11, as their understanding of mathematics grows.

The PAT: Mathematics scale was constructed using Item Response Theory (IRT), which is a framework for the development, analysis, and scoring of tests and questionnaires used to measure constructs such as achievement and attitude. More specifically, the scale is based on a mathematical model called the Rasch model (Rasch, 1980)[2] which is a special case of an IRT model.

To link the refreshed tests to the PAT: Mathematics scale, we trialed the questions that would make up the refreshed tests with a sample of students (the calibration trial). Questions in the refreshed tests that had little or no changes made to them were used to locate new and changed items on the PAT: Mathematics scale, based on achievement in the trial.

The calibration trial design

The trial design involved using the refreshed content to construct eight trial test forms. These forms were intended to be as close as possible to what would become Refreshed Tests 1–8.

Most of the questions were drawn from existing Tests 1 to 7 and Test 8A. Some of the repurposed questions had minor changes made, and others had more substantial changes. Some completely new questions were also developed. Due to the addition of updated graphics across the bank of PAT: Mathematics questions, there were very few questions that remained identical to their existing form. In order to identify which questions should be used to link the tests to the existing PAT: Mathematics scale, four levels of change were defined, and questions categorised according to those levels. The levels were as follows.

- No change—these questions were identical to their existing form.
- Insignificant change—these questions had changes that we were confident would not affect their function.
- Minor change—these questions had changes that were not expected to affect how they functioned. However, we were not confident to label them insignificant without comparing the question's behaviour before and after the changes.
- Significant change—these questions had changes that were expected to affect their function.

[1] The PAT: Mathematics scale is due for a full recalibration in late 2023. At that time the intention is that the existing static tests (Tests 1 to 7, and 1A to 8A) will be retired. The questions in the refreshed tests will be recalibrated, along with all the other questions in the PAT: Mathematics assessment bank that are available for selection in computer adaptive testing. In the interim, the refreshed tests have been linked to the existing PAT: Mathematics scale.

[2] Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests (revised). The University of Chicago Press. (Original work published 1960.)

Table 1 shows the distribution of questions in the eight trial tests by relationship with existing questions. For ease of reference, “no change” and “insignificant change” will hereafter be referred to together as “unchanged”, as we acted on the assumption that the function of these questions was not affected by the refresh process.

Table 1 Amount of change to test questions by test form

Trial test	No change	Insignificant change	Minor change	Significant change	New	Total
1	3	9	12	6	5	35
2	6	10	13	4	2	35
3	8	9	11	11	2	41
4	2	17	14	6	2	41
5	12	9	11	9	3	44
6	8	9	17	6	6	46
7	10	18	11	5	3	47
8	10	9	12	10	6	47
Total	59	90	101	57	29	336

The calibration trial sample

While all of the PAT: Mathematics static tests can be used across multiple year levels, they are designed with a target year level in mind. Trial Tests 1 to 6 were administered to a single year level, being the intended target for the resulting refreshed test. Trial Test 7 was administered at both Years 9 and 10, with Refreshed Test 7 intended for Year 10, but also for Year 9 when testing is done in Term 4. Trial Test 8 was administered to Year 10 students, with Refreshed Test 8 being intended primarily for the end of Year 10. Table 2 shows the number and year level of students who trialed each test.

Table 2 Number of students sitting each trial test

Trial test	Year level(s)	Number of students
1	4	617
2	5	624
3	6	718
4	7	595
5	8	619
6	9	560
7	9 and 10	429
8	10	519

Seventy-three schools representing the range of deciles participated in the trial. Most of those schools trialed multiple test forms. The trial tests were administered online in August and September 2022.

The calibration process

Calibration involves locating the difficulty of each question on a measurement scale—in this case the PAT: Mathematics scale. Calibration was carried out using Rasch modeling. Each test form was calibrated separately and linked to the PAT: Mathematics scale using common item equating.

The calibration process was multi-step. First, a preliminary analysis based on item statistics and graphical indicators looked at how well the questions functioned within the context of their respective tests. Most performed well.

Next, we assessed how closely the question difficulties of the "unchanged" questions in the trial reflected their existing PAT: Mathematics scale difficulties. We identified the questions that seemed to be functioning consistently and designated these as "link" questions. The average trial difficulty of the link questions was aligned with their average difficulty on the PAT: Mathematics scale by calculating the difference between the two sets of averages and applying that "shift" to the trial difficulties. The shifted difficulties of questions with minor changes were compared with the existing difficulties for those questions to help determine whether the change had affected question performance. If the pair of question difficulty estimates were within 0.2 logits (an approximate 95% confidence interval for a question's difficulty estimate is +/- 0.2 logits), the question was considered to be functioning similarly enough to retain its existing difficulty estimate. If the estimates were more than 0.2 logits apart, the question was considered to be functioning differently, and needing a new difficulty estimate.

The last step in the calibration process involved a further application of the Rasch model to determine the final question difficulty estimates. For this model, the questions that were designated as link items were "anchored" (their difficulties were fixed using their existing

PAT: Mathematics scale locations). The other questions were allowed to "float" (the Rasch model calculated the difficulty estimates for these on the basis of the anchors and the trial data). In total, 11 of the trialed questions were considered inappropriate for inclusion in this model. One of those questions was unchanged and excluded from the model because it had an adverse effect on linking. The other 10 questions were new or had been changed, and exhibited fit issues. Of these, four were deleted from the refreshed tests, five were altered and assigned difficulty estimates based on similar items, and one reverted to its original form and difficulty. After the difficulty estimates were finalised, 15 questions were deemed unnecessary and removed from the tests. A further eight questions were moved from one test form to another to ensure each test included questions that appropriately covered curriculum objectives and were of appropriate difficulty for the test. The resulting set of eight test forms comprise Refreshed Tests 1 to 8. Table 3 shows the average difficulty of the questions in each of the refreshed tests. Figures 1 and 2 show the distributions of question difficulties by test (Figure 1) and curriculum strand (Figure 2). The shaded bars to the right of each figure provide some indication of how the questions relate to achievement associated with curriculum levels 1 to 5 of *The New Zealand Curriculum*.

Table 3: **Number of questions and average question difficulty by test form**

Refreshed test	Number of questions	Average question difficulty (patm units)
1	35	31
2	36	40
3	38	47
4	38	53
5	41	57
6	43	60
7	43	66
8	43	71

Figure 1: The distribution of question difficulties by refreshed test

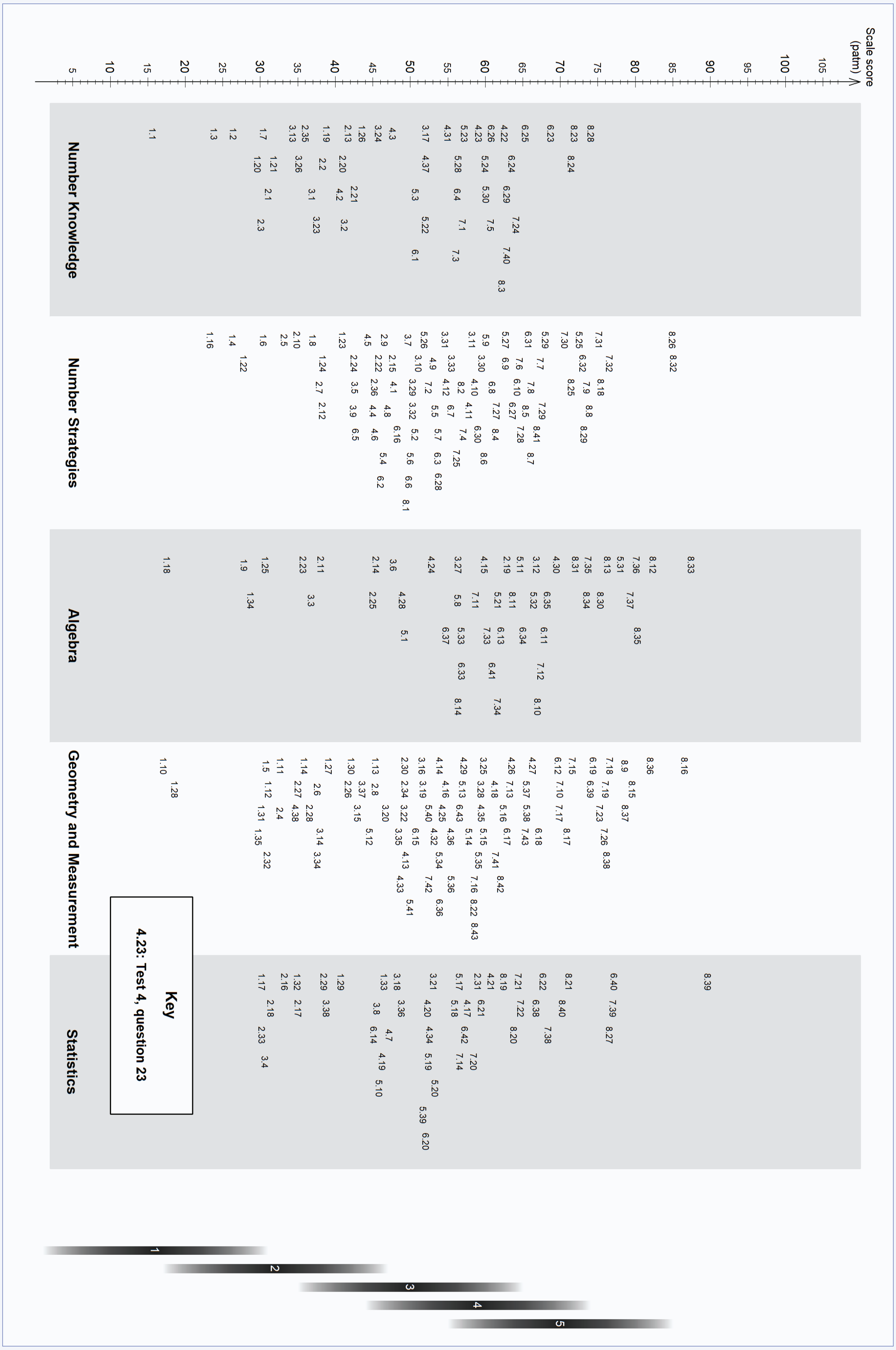
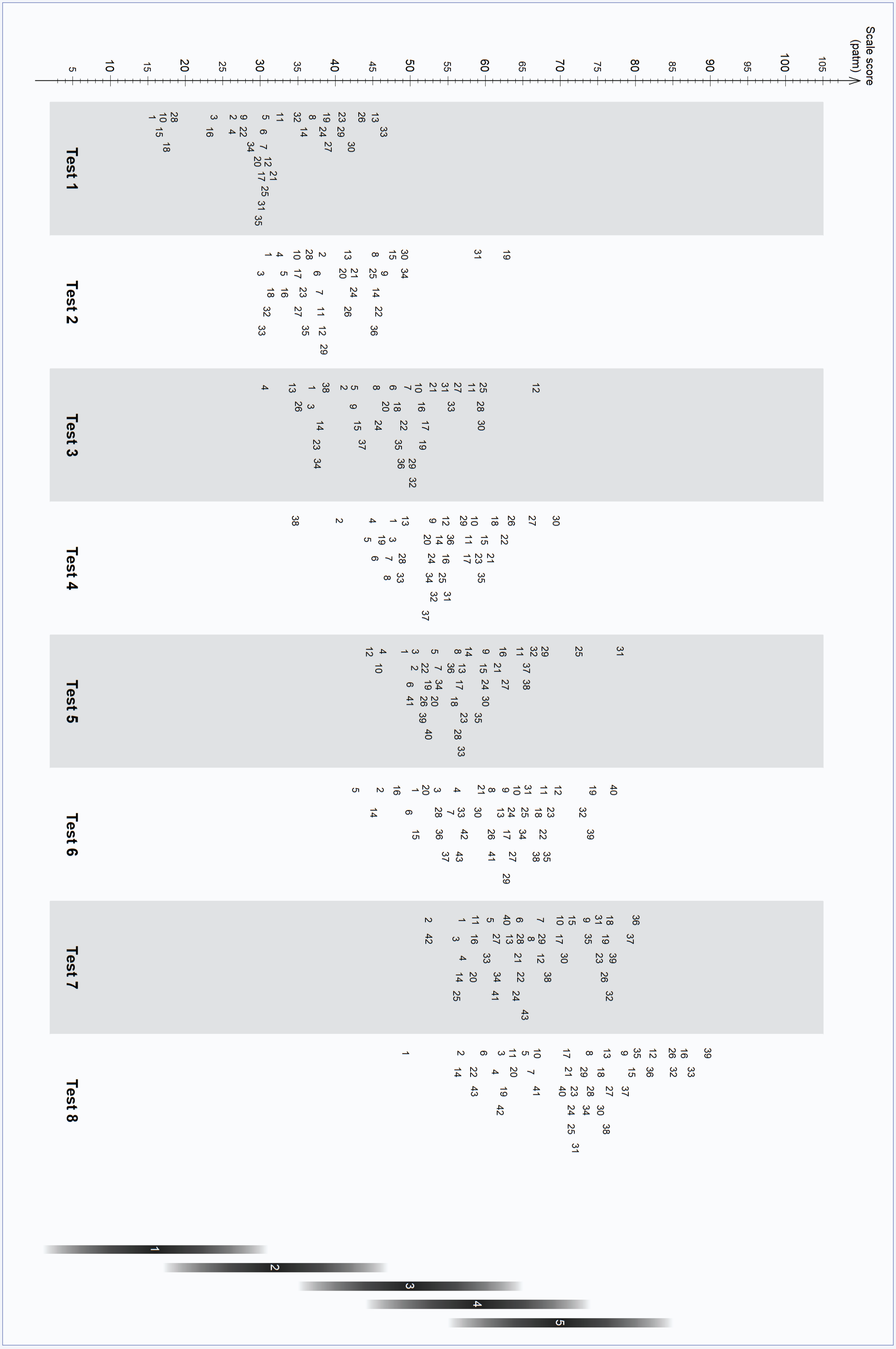


Figure 2: The distribution of refreshed question difficulties by curriculum strand



Norming

The normative information for the refreshed tests is based on the norming study described in the 2009 version of the PAT: Mathematics teacher manual (Darr et al., 2009).3 The only exception to this is the normative information for Year 10 students in Term 4, which was collected in 2011.

Table 4 shows the estimated average and standard deviation of PAT: Mathematics scores for students in each year level as well as the number of students used to estimate these statistics. It is reproduced from Darr et al. (2009), but supplemented with the Year 10, Term 4 distribution information.

Table 4 Average PAT: Mathematics scores, by year level

Year level*	Number of students	Average achievement score (patm)	SD in achievement score (patm)	Average progress score since previous year (patm)
3	2,744	21.4	12.5	na
4	10,609	30.6	12.8	9.2
5	11,505	38.9	12.4	8.3
6	11,353	45.1	11.7	6.2
7	6,900	49.6	11.5	4.5
8	6,751	55.0	11.4	5.4
9	2,156	60.6	11.6	5.6
10	1,991	65.4	11.8	4.8
10, Term 4	1,280	66.6	12.5	1.2

* Unless stated, all norming data were collected in Term 1.

[3] Darr, C., Neill, A., Stephanou, A., & Ferral, H. (2009). *Progressive Achievement Test: Mathematics, Teacher manual* (2nd Ed. Revised). NZCER.

Validity and reliability

Validity

The validity of a test is the degree to which the test measures what it was intended to measure. We believe the PAT: Mathematics refreshed tests provide a valid measure of student proficiency in mathematics because:

- they were planned and constructed to fit the New Zealand curriculum
- a range of experts and stakeholders were involved in the development process, and the tests underwent multiple stages of trialling and review
- the response data generated from the tests fitted the Rasch measurement model well
- they exhibited strong psychometric links to the existing PAT: Mathematics tests.

Reliability

The reliability of a test is the degree to which it will provide consistent scores in repeated testing. In IRT, test reliability is measured using a reliability coefficient.

The reliability coefficients for the refreshed tests, based on the trial data, were between 0.86 and 0.9. A reliability coefficient of 0.86 indicates that 86% of the variance in scores can be attributed to actual differences in student achievement, while 14% of variance is attributable to measurement error. The more reliable, or precise, a test, the smaller the errors associated with estimated scale scores based on responses to that test.

Precision of scale scores

The Rasch measurement model provides an estimate of the student's PAT: Mathematics scale score. This estimate comes with an error range that can be used to determine the precision of the estimate (the measurement error). An error range of 3.6 patm, for instance, indicates that the reported scale score for a student is likely to be within plus or minus 3.6 patm units of the student's true scale score in about 70% of cases.

It is important to note that the error range associated with a scale score is higher for students who do extremely well or very poorly on a particular test. This is because in these cases the test provides less information about what these particular students can actually achieve than it would for students who obtained 50% of the maximum test score. A test better targeted to high- or low-performing students would provide a more precise indication of their scale score. We recommend that teachers take the scale score error into account when interpreting scores for individual students. A scale score should be understood as a *range* on the scale instead of a precise point; for example, a scale score of 48 patm with an associated error range of 3.6 patm should be thought of as a range from 44.4 patm to 51.6 patm.

NZCER
Rangahau Mātauranga o Aotearoa

Rangahau Mātauranga o Aotearoa |
New Zealand Council for Educational Research

facebook.com/nzcer

@NZCER

www.linkedin.com/company/new-zealand-council-for-educational-research