Can test results help us make Overall Teacher Judgements?

CHARLES DARR

From time to time the New Zealand Council for Educational Research is asked how test results can contribute to making an Overall Teacher Judgement (OTJ). This is an important and complex question and one that we continue to grapple with. In this Assessment News article I begin to explore what we need to consider when using test results to support and promote teachers' professional judgements.

Making an OTJ involves using a range of evidence to decide how a student is achieving relative to a National Standard. If, as teachers, we want a test result to contribute to this decision-making process, we can make a start by clarifying three things:

1. the extent to which the test represents the knowledge and skills described by the standard
2. the level of performance we can reasonably expect on the test from students who are achieving at a standard
3. additional factors besides those being assessed that may have influenced group and individual performance on the test.

The extent to which the test links to the standards

When we want to use a test to help us understand whether students have reached a year-level National Standard, we first need to evaluate how well and to what extent the test reflects the range of competencies (knowledge and skill) described by the standard. This is always a matter of degree. No test can ever fully reflect the intent or scope of a standard; tests are limited, for instance, by the kinds of responses they can elicit, how authentic the tasks are and the amount of testing time available. Some tests may be useful for examining performance on a focused portion of what a standard describes; others might allow a more general, but less detailed, interpretation of achievement against the whole standard. In some cases tests might be very poorly linked to the outcomes described by the standard and results may be difficult to interpret. In general, the stronger the link between what the test assesses and the kinds of competencies the standard describes, the more potential the test has to inform OTJs about student performance.

Judging the performance standard on the test

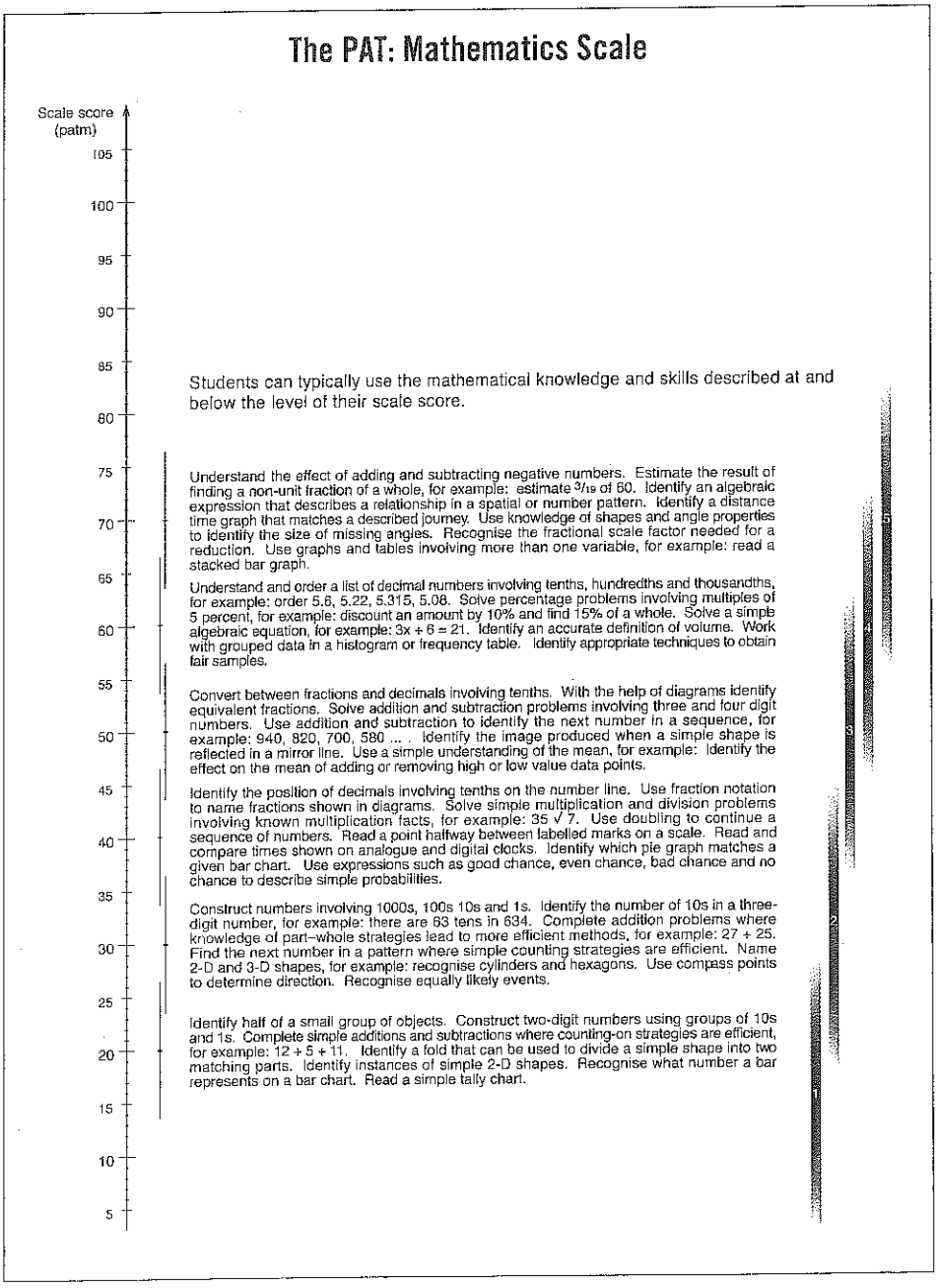
Once we understand how the test relates to the standard, we need to evaluate what kind of performance on the test can be reasonably expected from students who have the skills and knowledge the standard describes. This is a criterion-related interpretation and is different from working out whether a test result is high, average or poor for a student in a year level (a normative interpretation). On some tests, even a score that is relatively low when compared to other students' scores could still indicate that the student has the competencies described by a year-level standard.

Interpreting what a performance level on a test means in terms of the level of knowledge and skills it represents is a judgement call. It requires a strong understanding of the curriculum and the standards, coupled with an ability to unpack what the questions in a test are probing. The objective is to develop a "feel" for what different levels of performance (score ranges) on the test indicate in terms of the knowledge and skills we can attribute to the student. When we develop this feel we can use a test result as another piece of evidence regarding the student's ability to work productively with at least some of the skills and knowledge the standards describe. This piece of evidence can then be considered alongside other information we have about the student when making an OTJ.

Reaching an understanding of what a test performance might indicate in regards to the National Standards is best done with colleagues in the school and, if possible, in other schools. In doing this teachers share their beliefs and understandings about what knowledge and skills are being described by each standard and how responses to the test items demonstrate this. When this shared understanding can be extended to teachers in other schools we introduce the potential for system-wide learning.

One possible idea for a shared exercise could involve teachers working through a test question by question to divide the questions into two groups. The first group would contain the questions that call on the skills and knowledge the teachers believe students who are working at the standard should have. The second group would contain the remaining questions; for instance, some of these might call on skills and knowledge that might be more reasonably expected from a student working above the standard (say the next standard or higher). Teachers would individually evaluate how many questions in each group a student working at the level of the standard should be able to answer correctly. Discussion could then be used to reach a consensus. If needed, an average number correct could be calculated across the teachers. This information would then be converted to an indicative total score on the test.

FIGURE 1 TAKEN FROM THE PAT: MATHEMATICS MANUAL



As an example, a group of teachers might look at a mathematics test with 35 questions that the school has developed to assess their Year 6 students' algebra knowledge. After some discussion, they might decide that the content lines up well with objectives related to algebra in the curriculum and, particularly, with the kinds of knowledge and skill described by the Year 6 standards. After working through the test they decide that 25 of the questions should be able to be answered correctly by a student working at the Year 6 standard. After some discussion they are prepared to accept that a student might get 20 percent of these incorrect in a test situation, meaning a score of 20 out of 25 is still in the expected performance range. They might also agree that a student working at the standard in their school should be able to get at least three of the remaining 10 questions correct. Overall, this suggests that a score of 23 or more on the test is an indicator that a student is achieving at the standard when it comes to algebra. Note that we are interested in the total test score, rather than which individual questions a student actually got correct. The total score is the most reliable piece of information we have from the test about the students' overall level of competency. On any one occasion, students can get individual questions correct or incorrect for all sorts of reasons. If tested again, however, their total score is more likely to remain within a fairly constant range.

Subject matter experts and test developers went through these kinds of judgement processes with tools such as asTTLe, PAT: Reading Comprehension and PAT: Mathematics to reference scale scores to curriculum levels—for example, see Figure 1 taken from the PAT: Mathematics manual.

More recently, the Ministry of Education has invested in studies to link scores on these assessments and others to the different year-level standards (see http://assessment.tki.org.nz/Assessment-tools-resources/Alignment-of-assessment-tools-with-National-Standards). Schools can use these studies to help them understand how others have judged the performance demands of these well-known tests against the curriculum levels and standards. Schools might also want to do their own exercises, particularly in relation to how they use the tests within their teaching and learning context.

What else do we need to think about when using test information?

How was the test administered?

Students can be advantaged or disadvantaged depending on how we administer the test. For instance, if the test was administered in a haphazard way (for example, different students received different levels of instructions) or if students were given extra support, then it is going to be difficult to interpret scores.

Do students have the assumed background knowledge and skills to perform on the assessment?

Other factors not related to the competencies being tested can affect a student's score. When a test of mathematics, for instance, requires strong reading skills, a weaker reader may find it difficult to show how mathematically competent they are. If the questions were read to them they might do much better. In this case, we might decide that we can't expect the same level of performance on the test from these students and that the assessment result underestimates what they can do. In this case we should look for other ways to assess these students.

Did the students prepare for the test?

If students have done special preparation for a test they are likely to perform better. We need to be sure that the higher scores achieved after this kind of preparation represent real and long-lasting changes in understanding and knowledge, and are not just the effect of special preparation.

How long ago was the test administered?

Students are changing all the time. An old test result will not necessarily reflect what a student can do now, particularly if they have had opportunities to learn in the meantime. We need to be cautious in validating today's judgement using out-of-date information.

A system that puts teacher judgements at the centre makes use of tools in a professional manner. It never reduces decision making to strict applications of cut scores, nor ties itself slavishly to what one tool says. It does, however, attempt to validate decisions by providing worthwhile evidence. Using tests in a standards environment can add to this evidence when professional and collegial judgements are used to understand what the tests are telling us about student performance.

CHARLES DARR is a senior researcher and manager of the assessment design and reporting team at NZCER.

Email: charles.darr@nzcer.org.nz