A hitchhiker's guide to reliability

Charles Darr

Validity and reliability are two key ideas in assessment. In the last issue of ***set*** I looked at the concept of validity and how it might inform the assessment decisions we make as classroom practitioners and school leaders. In this article I address the issue of reliability, and how it too can help inform our assessment strategies and practices.

What is reliability?

Reliability refers to the consistency of the results we obtain from an assessment. This may mean:

- Consistency across time—would the results have been the same if the test or assessment had taken place on another day, or at another time?
- Consistency across tasks—would the result have been the same if other tasks had been chosen to assess the learning?
- Consistency across markers—would the results have been similar if another marker had scored the assessment?

The higher the level of consistency, the more reliable are the results. No results, however, can be completely reliable. There is always some random variation that affects the assessment.

Chase (1974) helps us understand reliability by using an analogy from everyday life. When we measure the length of a room, the consistency of the results we get will vary depending on what instrument we use to take the measurement. For instance, a conventional metre ruler will give us much more consistent results than an elastic tape measure. The ruler is rigid and stable; however many times we use it to measure the room, there will be a high degree of agreement from one measurement to the next. The elastic tape measure, on the other hand, will need to be stretched just the right amount to show an exact metre, and so will produce much less consistent measurements. Sometimes we will stretch the tape too far and end up underestimating the length; sometimes we will not stretch the tape far enough, and overestimate it. The elasticity of the tape measure introduces an element of random variation into our measuring process that will make our measurements less consistent, and so less reliable.

When the results of an assessment are reliable, we can be confident that repeated or equivalent assessments will provide consistent results. This puts us in a better position to make generalised statements about a student's level of achievement, which is especially important when we are using the results of an assessment to make decisions about teaching and learning, or when we are reporting back to students and their parents or caregivers.[1]

Determining reliability

Determining reliability has traditionally been seen as a statistical exercise. It usually involves calculating a reliability coefficient to indicate how well assessment results agree over repeated uses of the assessment tool. Reliability coefficients vary between zero and one, with zero indicating no agreement and one, total agreement (a result that is never actually obtained in educational assessment). Test developers use several methodologies to calculate reliability coefficients, depending on the type of consistency they are interested in. Some of these are briefly described below.

Reliability coefficients on standardised tests are often greater than 0.9, indicating a high degree of reliability. The question of how high the reliability for a set of assessment results should be, however, depends very much on what level of decision making will be based on our assessment results. When we are dealing with assessments that are highly significant we need high reliability. When the decisions do not have lasting consequences, a high reliability measure is not as important.

The Standard Error of Measurement

The reliability coefficient for a set of test results is sometimes used to calculate what is called the standard error of measurement (SE_m). This can be used to describe a band of achievement within which we can be reasonably sure a student's true level of achievement actually lies. A true level of achievement can be thought of as the level that would be achieved by the student if the test was perfectly reliable. For example, when a score is reported as 30 with a SE_m of 3, we can be reasonably confident that the true achievement level of the individual would fall somewhere in the range 27 to 33. In educational testing it is considered good practice to report the SE_m. In the Progressive Achievement Test of Reading, for instance, the manual writers provide an estimate of the SE_m for each individual test (which is usually around 3 marks). Being aware of the SE_m helps us regard a student's test score as describing a range rather than a precise point.

Reliability in the classroom

It is unlikely that teachers will spend time calculating reliability coefficients or reporting standard errors of measurement for the assessment tasks or tests they develop

Table 1 METHODOLOGIES FOR DETERMINING RELIABILITY

Method	Process
Test/retest	The same test, or an equivalent version, is administered at two different times to the same group of students. The two sets of results are then compared to calculate the reliability coefficient. This method provides an indication of how consistent the results are over time or between equivalent forms of the same test.
Internal consistency	The results on different tasks or sections of an assessment are compared to see how well they relate. Several different methods can be used, including dividing the test into two halves and comparing the results on each half (split-half method). Other instances of this type of reliability coefficient involve more sophisticated statistical methods. In test manuals it is common to see what is called Cronbach's alpha, or the use of Kuder Richardson formula 20 or 21. Calculating this type of reliability coefficient provides an indication of how consistently the items or tasks within an assessment promote the same result.
Inter-rater reliability	Results from different markers can be compared to ascertain the level of agreement. This method used to show how consistently two or more assessors are scoring the same tasks, is called *moderation* when it is used in the context of assessment for qualifications.

Factors affecting reliability (and reliability coefficients)

- The number of tasks in the test or assessment—more tasks will generally lead to higher reliability.
- The suitability of the questions or tasks for the students being assessed—questions that are too hard or too easy for the students will not increase reliability.
- The spread of scores produced by the assessment—the larger the spread of results, the higher the reliability.
- The training of the assessors.
- The clearness of marking guides and checking of marking procedures.
- The wording of the rubric—carefully worded rubrics make it easier to decide on achievement levels.
- How closely standardised procedures and conditions for assessment are followed.
- How well questions and tasks are phrased.
- The anxiety or readiness of the students for assessment—assessing students when they are tired or after an exciting event is less likely to produce reliable results.

themselves. This does not mean, however, that the issue of reliability is irrelevant to classroom assessment—it is very important to base judgements and decisions about students on assessment results that are dependable. So how can classroom teachers ensure that their own assessments are reliable? Jeffrey Smith (2003) argues for an alternative definition of reliability that he believes is more appropriate for classroom-based assessment. He proposes the idea of "sufficiency of information". For Smith, judgements about the reliability of classroom assessments can be built on a question such as: "Does this assessment provide me with enough information to make a judgement of each student's level of accomplishment with regard to this learning?" (Smith, 2003, p. 26). Taylor and Nolan (1996) also promote this point of view. According to them, "A wide range of assessments can serve the purpose of a long test—the more sources of information, with demonstrable evidence for validity, the more likely dependable decisions can be made" (p. 11).

This kind of definition begs the question: how much information is enough? There is a fine line between "enough information" and too much assessment. There is of course no easy answer to this; in the end, it comes down to a professional judgement. Some practical advice on this issue, however, is provided by Anne Davies (2000). She suggests that teachers use the concept of triangulation as a way of increasing the reliability and validity of classroom assessments. Triangulation involves using three different sources of assessment evidence as the basis for any decision making. For Davies, these areas are observations of learning, products students create (including test results), and learning conversations. When teachers collect and consider evidence from each of these sources, they are far more likely to reach dependable and meaningful conclusions about students' progress than when they rely on one single area or result alone.

The issue of reliability alerts us to the fact that random variation does occur in assessment. It is something that should concern us when we assess students, particularly when the results of assessments are used to make decisions about individual students and/or the teaching and learning programmes they are involved in. As in my previous article on validity, this is only a "hitchhiker's guide" to what is a very important assessment concept—there is a lot more we could say about reliability. This may, however, serve as a reminder that it is important to take issues such as reliability into account if we are to make informed decisions about both the process of assessment and the results.

References

Chase, C.I. (1974). *Measurement for educational evaluation*. Reading, MA: Addison-Wesley.

Davies, A. (2000). *Making classroom assessment work*. Courtenay, BC: Connections Publishing.

Smith, J.K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, *22*(4), 26–33.

Taylor, C.S., & Nolen, S.B. (1996). What does the psychometrician's classroom look like?: Reframing assessment concepts in the context of learning. *Education Policy Analysis Archives*, *4*(17). Retrieved from http://epaa.asu.edu/epaa/v4n17.html

Note

1 It is important to note that unreliable results will not lead to valid inferences about student achievement. However, just because assessment results are reliable does not mean that we are assessing what counts. Reliability is therefore a necessary but not sufficient condition for validity.

Charles Darr is a senior researcher at the New Zealand Council for Educational Research.
Email: charles.darr@nzcer.org.nz