Living with uncertainty - Reliability in educational assessments

Colin Robinson

Educational Consultant, UK

Presented at the 35th International Association for Educational Assessment (IAEA) Annual Conference in Brisbane, Australia, 13–18 September, 2009

July 2009

Living with uncertainty – reliability in educational assessments

Colin Robinson Educational Consultant, UK

Reporting of results for individual students can take a variety of forms, including numerical scales, ordered categories, competency checklists, brief or detailed descriptions, and demonstrations or portfolios. Numerical scales and ordered categories may or may not be accompanied by descriptive statements for determining or giving meaning to particular scale-points or categories.

Results of assessments are used for a variety of purposes but there is no agreement as to how it is best to present the information for different users nor whether it is necessary or desirable to indicate to users the degree of uncertainty that within the assessment. This is an aspect that Dennis Opposs will be discussing in his presentation later.

In considering the results of assessment we must be as clear as possible what it is we are trying to do and what we mean by valid and reliable assessment.

In Principles for Fair Student Assessment Practices for Education in Canada, validity is defined as follows:

> "Validity refers to the degree to which inferences drawn from assessments results are meaningful. Therefore, development or selection of assessment methods for collecting information should be clearly linked to the purposes for which inferences and decisions are to be made. For example, to monitor the progress of students as proofreaders and editors of their own work, it is better to assign an actual writing task, to allow time and resources for editing (dictionaries, handbooks, etc.), and to observe students for evidence of proofreading and editing skill as they work than to use a test containing discrete items on usage and grammar that are relatively devoid of context."[1]

[1] Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, Alberta: Joint Advisory Committee. (Mailing Address: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, 3-104 Education Building North, University of Alberta, Edmonton, Alberta, T6G 2G5).

Are we attempting to record, as accurately as possible, the candidate's performance as revealed in the evidence presented, or are we attempting to go beyond that evidence and estimate the candidate's underlying ability or competence? For the assessors, the only realistic approach is the former: though the nature of the evidence we accept may vary considerably across different approaches, it is that evidence we are assessing; to go beyond the evidence is to enter the realm of speculation.

Users of the results, however, usually – indeed I would go so far as to say *always* - want to go beyond this narrow assessment. Parents and learners themselves want to know whether they are making good progress, what their strengths are and where their weaknesses lie. With this information they hope to be able to influence their future learning or to choose their future path. Educators want to know whether the learner will be suited to the course they are offering. Prospective employers want to know whether the applicant will do the job satisfactorily. Underlying these decisions is the assumption that the result is indicative of something more fundamental and more enduring – possibly more fundamental and more enduring than the assessment deserves.

Any assessment contains an element of imprecision. A myriad of extraneous factors will have played a part in the assessment and may have caused subtle differences in the resulting grade: the assessment will have taken place in a particular context, on a particular day, and the candidate will have been in a particular state of mind, tiredness and health; the papers will have been written by particular examiners with a particular style; and each of the markers will have their own particular interpretations of the subject. If users do not understand and take account of these factors, and treat the result as an absolutely precise measurement, the decisions they make may well be unsound.

Uncertainty can derive from a number of factors:

- The test paper or task specification
    - Almost all assessments will start with some form of stimulus to which the candidates respond. The choice of questions or tasks will be constrained to ensure as close a match as possible with the syllabus but there will inevitably be some variation. Had a different choice been made a candidate might have given a different (better or worse) response. We cannot know – all we have is the evidence of this particular response.
- The marker or assessor
    - For the majority of assessments, a further source of uncertainty is the marker or assessor. However well-trained and professional they may be, they bring to the assessment their own interpretations of the question or task, of the mark-scheme or success criteria, and of the candidate's response. In any of these aspects a different assessor – or even the same assessor at a different time - may reach a different conclusion. However, assessor judgment is unlikely to be much of an issue with multiple choice (objective) tests. Appropriate checking processes (or machine scoring) should remove any actual mistakes.
- The context
    - The assumption behind the assessment is that all candidates take the assessment in the same, standardised conditions. However in practice the context for the assessment varies and in ways that are difficult to quantify. In e-assessment, for example, one of the issues is the variability of the computer systems that candidates will use for the assessment.

All of these can be brought together under the heading "generalisation". It depends how far we wish to generalise the result beyond individual performance. Consider for a moment the annual sports day at a local primary school. In this case no generalisation is required: the results are the children's performances on the day with the best performance winning. We don't care whether the result this year is quicker or higher than last year's. We don't care whether these children are quicker or higher than those in other schools. In this case, therefore, reliability is hardly an issue. Provided the race is valid and cheating is prevented, the winners are "self-evident".

Let us change the situation. Move to an event where the various schools in the region compete. This time comparison with performances at other events is required: records are at stake. Is this year's winner of the 100 metre sprint faster than previous winners? How does the height achieved by this region's high jump winner compare with those in other regions? In order to be comparable, the contexts in which the events take place must be standardised. If this is not possible, for example, if a following wind might have assisted the athletes, the performance cannot be accepted as a record, even though the trophy will still be presented to the winner.

In educational assessment, we need to consider how far the results will be generalised beyond the individual performance. If the performance itself is the focus of interest, then rigorous analysis may be unnecessary, we can just report what the student did in the circumstances. There may be validity issues related to such assessments, but it is unlikely that reliability studies would be relevant or useful.

In the vast majority of educational assessments, whether these are in the classroom, homework assignments, formal examinations, part of on-the-job training or even self-assessed private study, we want to generalise beyond the individual performance and provide a report to the student or to other users of something deeper and potentially more useful. Graham Maxwell's paper looks at different approaches to reliability. Dennis Opposs focuses on the ways in which the uncertainty (what I would prefer to call imprecision) of the results are reported in the media.

Many systems convert marks (numerical scales) into grades before reporting results. This may partly be to make results comparable across different year groups or across different subjects, but it is also to simplify the report. Grades (or their equivalents such as the Uniform Mark Scale, or the National Curriculum levels) are normally an ordinal scale. In other words, all we should infer from them is that a higher grade represents a

better performance. We should not assume that the scale is linear – i.e. that the distance between grades is constant. And most certainly they cannot be regarded as ratio scales – even the lowest possible grade cannot be taken to mean a complete absence of the characteristic we are attempting to assess.

Grade scales are used in many contexts around the world. Any scale is subject to some level of imprecision, even if the underlying performance can be measured with total accuracy. The larger the units we use to report (and usually the smaller number of them), the greater the degree of uncertainty. Even the most accurate clock, if reported only in complete minutes, will be inaccurate for most of the time – only for the split second when the time happens to hit a whole minute will the clock record the exact time.

Educational assessment is not measuring something physical like length or time, where it is possible to make reference to a standard invariable unit. It is therefore at best only an indicator. In a paper to a previous conference, Alastair Pollitt pointed out that interactions between the task, the candidate and the assessor can lead to differences of interpretation that are not necessarily a reflection of the candidate's performance.

> "Consider what happens when a Marker, M, awards a mark to a response given by a Candidate, C, to a question set by a Setter, S, (this is the heart of the assessment process):
> M evaluates
> M's interpretation of
> C's expression of
> C's answer to
> C's interpretation of
> S's expression of
> S's task, using
> M's interpretation of
> S's expression of
> S's demands."[2]

[2] Pollitt, A & Ahmed, A. 1999 *A New Model of the Question Answering Process* A paper presented at the IAEA Conference in Bled, Slovenia, May 1999

These are not mistakes. The setter must design a task which aims to provoke a particular response from the candidate that will demonstrate what the candidate knows or can do. The candidate has to interpret the task in order to decide how to respond. The candidate's response is interpreted by the marker (who is quite probably not the same person as the setter) in the light of the marker's own interpretation of the task and the mark scheme.

Even if each of these interpretations is perfectly valid, they can contribute to variations in the results. If the candidate interprets the task in a way that was not expected by the setter and therefore included in the mark scheme, there is a reasonable chance that the assessors will not value the response as highly as one that conforms to their interpretations.

In another paper, Alastair Pollitt highlights the fact that there is no absolute imperative to use marks or scores in assessments.

> "The requirement is that we find some way to ***judge*** the students' performances in order to create the scale we need, and marking items to add up their scores is just the way we have chosen to do this." (original emphasis)[3]

Assessments can be made by directly judging the quality relative either to some notion of an "ideal" performance or to another, actual performance. In such a context, classical definitions of reliability are inappropriate, but that does not mean that such judgements cannot be carried out reliably.

One of the problems with the classical definition of reliability is that it refers to "error". In this context error refers to the imprecision of the assessment result and contains all the factors that are irrelevant to the

[3] Pollitt, A (June 2004) *Let's stop marking exams* Paper presented to the International Association for Educational Assessment Annual Conference 2004, Philadelphia USA

measurement what is being assessed. Unfortunately for us, however, the term "error" has a more general meaning of "mistake". This leads to the misunderstanding that it is something that the awarding body (or marker or awarder) did wrong and is therefore seen by many as something that it should be possible to eradicate.

There are many aspects of the assessment process that can be improved to remove real mistakes such as incorrect application of the assessment criteria, incorrect addition of marks, or incorrect combination across components, but we cannot remove some element of variation that will lead to uncertainty in the results obtained.

It is essential that our assessments are as valid and as reliable as possible. However, we must not become obsessed by reliability – to do so places validity in jeopardy.

I end with two quotes from Albert Einstein.

"Everything should be made as simple as possible, but not simpler."

"Not everything that counts can be counted, and not everything that can be counted counts."[4]

Colin Robinson

July 2009

[4] downloaded from http://rescomp.stanford.edu/~cheshire/EinsteinQuotes.html