Dealing with inconsistency and uncertainty in assessment

Graham S. Maxwell
Educational Consultant, Australia

Paper delivered at the 35th Annual Conference of the International Association for Educational Assessment, Brisbane September 2009 as part of a symposium entitled *Reliability in assessment—what role should it play and how should we explain it?*

This symposium is about how we think about and talk about the concept of reliability, both technically and publicly. This paper explores some aspects of what we mean by reliability and poses some issues for further consideration.

Technical and vernacular uses of the term reliability differ. Technically, reliability is seen to be about replicability—if we did the assessment again, would we get the same result? On the other hand for the general public, it is about whether the results are accurate and believable. The technical usage involves estimation—we cannot actually wipe the slate clean and do it all again. The vernacular usage is much more intuitive—can we trust the results? Both deal with the degree of certainty (or uncertainty) we have in the results, but they differ in their focus and scope. Technical usage is specialised and restricted. Vernacular usage is much broader and expansive, incorporating at least something of the concept of validity.

1. *Typically, reliability is distinguished from validity, but is this useful?*

The theories surrounding reliability and validity developed in the context of standardised testing and they still retain the distinguishing marks of their origin. It is not clear that they serve adequately in that form for the broader range of assessment practices that we engage in today such as performance assessment, portfolio assessment and competency assessment. Maybe we need different concepts and language for different types of assessment.

Teachers typically do not approach their assessments through this classical reliability/validity lens. The average teacher, let alone the average member of the public, is unlikely to use the term ‘reliability’ and to distinguish it from ‘validity’. They are more likely to use terms such as accurate, consistent, comparable, believable, dependable and fair. The distinction between reliability and validity does not seem to be a natural one.

The classical definitions of *validity* focus on interpretation of the assessment (and are traditionally split into face, content, construct and predictive validity, each of which may be relevant in particular circumstances). Sometimes, validity is reduced, trivially, to whether ‘the test measures what we want it to measure’. That is unhelpful. Even if the test measures what we want it to measure, that does not guarantee it measures something worthwhile and is appropriate or adequate for the use that is made of it.

Messick (1989), in his extended treatment of theories of validity, made two important

and enduring suggestions: first, that all the previous distinctions of types of validity are really about the same thing (construct validity—basically the issue just considered—what interpretation can we place on the test results?); and second, that the use to which the results are put and the social effects that they have are important (consequential validity). This changes the focus of validity to its being about how we interpret and use the outcomes of assessment rather than its being a property of the assessment *per se*. In designing assessments, we need to be concerned about whether we are in the right target area for the intended interpretation—that is, whether we are likely to be able to draw appropriate conclusions from the results—and whether the use we are making of the results is beneficial—that is, whether the positive effects outweigh any negative effects.

An unfortunate aspect of the classical theories of reliability and validity is their claim that validity is reliant on reliability (though both are, in any case, matters of degree, not perfection).[1] This has had the effect of privileging reliability over validity, and encourages the adoption of forms of assessment with lower validity in the interests of higher reliability. Multiple-choice tests typically have higher reliability than extended tasks (mainly because the marking can be standardised). But this does not ensure higher validity, both construct and consequential, if their sampling of the field of interest is limited in terms of the learning outcomes we want to encourage.

Written examinations necessarily represent a restricted (and therefore biased) sample of the things we would like students to be learning in school. Standardisation of the set tasks and the conditions under which they are sat allows control of some of the potential sources of unreliability but this is bought at the expense of considerably reduced validity. On the other hand, school-based in-situ performance tasks potentially allow more comprehensive coverage of a wider range of learning outcomes. The question is does this seriously compromise reliability or can reliability at an acceptable level still be delivered. The Queensland experience is that it can, provided adequate processes for quality assurance are adopted.

This tradeoff between reliability and validity suggests that they need to be considered together. Uncertainty applies to both validity (what is assessed) and reliability (how well it is assessed). The central issue is whether we have discovered something worthwhile and dependable about each student's capabilities. This is about meaning and interpretation.

All this suggests that our initial focus needs to be on validity to ensure that we are on the right target. Then, we could consider what steps need to be taken to ensure that the assessments are as reliable as possible. At least this has the virtue of seeing reliability as a problem to be solved rather than something to be estimated (powerlessly) after the fact.

This is the way school-based assessment is thought of in Queensland. Written examinations assess a limited range of the knowledge and skills we want students to acquire and suffer from being single-occasion, time-constrained and resource-restricted. Continuous assessment allows more comprehensive coverage of intended learning outcomes, involving a variety of forms of assessment over time (including, for example, projects, designs, artefacts, presentations, performances, and maybe tests), all tailored to the aims of the curriculum and the learning context, with a

[1] Pamela Moss was one who challenged this assumption in her classic article *Can there be validity without reliability?* (Moss, 1994). She has continued to write about these issues (e.g., Moss, 2004).

selection of outputs collected into a portfolio to represent the 'latest-and-fullest' evidence of the student's capabilities. The validity (both construct and consequential) is therefore enhanced. Reliability, in the sense of comparability of judgments of the quality of student achievement is a secondary, but nonetheless seriously important issue, accomplished to professional and public satisfaction through a system of external moderation through panels of expert teachers. The issue in this paper is not how this is done but the orientation involved—a primary focus on what is to be assessed and a secondary focus on how to ensure public confidence in the results.

2. *The classical definition of reliability focuses on errors of measurement*

Classical reliability starts with the question of replicability: if we changed the test, the occasion and the marker, would we get the same results? This list identifies different sources of unreliability. 'Test' can more broadly stand for 'task' or 'exam paper' (including specifications, instructions, mode and items), 'occasion' for 'context' (including physical and temporal circumstances), and 'marker' for 'assessor' or 'judge' (including marking guides and training).

The effects of the first two of these sources of unreliability (test and occasion) can be checked (after the fact) by giving a second (parallel) test to a sample of students and comparing the results. This produces a test-retest correlation coefficient as the measure (actually an estimate) of reliability. Sometimes that is too difficult to arrange and we cheat by treating the test items themselves as if they were individual parallel tests, calculating a measure of internal consistency, thereby not actually controlling for such sources of unreliability as test mode and occasion.

It should be noted, too, that in many test situations we don't expect much internal consistency, since the components deliberately sample different knowledge and skill to make the representation more valid overall. That is usually the point of having several tasks and/or several criteria. We expect inconsistency across the components. With perfect internal consistency, we would not need several criteria or several tasks—one would be sufficient. It can of course be argued that components that are negatively correlated should probably not be added together since they are assessing completely different characteristics that are best kept separate. In practice, it is often the case that we combine components have may very low (even negative) correlations, such as for example, in portfolios.

At best, any such reliability estimations (whether test–retest or internal consistency) are fairly crude, because they consider only some of the possible sources of inconsistency in the results. A more systematic approach would try to address many likely sources of inconsistency together. The generalisability approach suggested by Cronbach et al. (1972) does just that. However, such approaches are rarely manageable in real time with real assessments. Some have suggested that it is the *limits* of generalisation that we need to be concerned about and propose a focus on *transferability* instead (see Gipps, 1994).

Where marking is a source of unreliability, there can be intra-marker and inter-marker inconsistencies, though usually the latter are the primary concern—it is assumed that if a marker is inconsistent within their own marking then they will be inconsistent with other markers, so it is sufficient to check for inter-marker consistency. Even so, marker training needs to attend to both.

Also often forgotten (because it is wrapped up in 'occasion') is the application of the student—did the student really try? And therefore, did we really discover what the student is capable of? Classically, if the student persistently or deliberately does not try, then that is seen as a failure of the test to engage the student, which therefore undermines its validity rather than its reliability. Reliability is only concerned with inconsistent performance (different levels of effort and engagement on different testing occasions).

Whether the student's application is relevant depends on what we think we are trying to find out. Are we assessing underlying capability (for monitoring progress) or is it a 'contest' (for certification). In the latter, application is a key component of what is being assessed.

Classical test theory (and its derivatives such as IRT) propose the existence of a 'latent trait' on which we are attempting to locate the student's 'true score'. This leads to the famous equation: $o = t + e$. That is, the observed score is equal to the true score plus error. Of course, we can't estimate the individual student's error, only the standard deviation of the distribution of errors for a group of students (the standard error, SE). That then leads to probabilistic statements of the kind: $t = o \pm 1.96SE$ for a 95 per cent confidence interval—that is, that a student's true score is estimated to lie within a range of the observed score at a certain level of probability (here, a probability of 0.95).

We will not go here into the assumptions involved in these estimations—they are legion; nor into methods of calculation. The point here is merely to remind us of the kind of thinking involved, especially the notion of a 'true score' and the underlying assumption that we are attempting to locate this true score along a continuum together with some indication of its inaccuracy. The best estimate of the individual student's true score is the observed score but the true score lies is a region of uncertainty (the confidence interval) around it.

Explaining standard errors to the public is difficult. In fact, the very word 'error' itself is fraught. Never admit you've made an error—it indicates a 'mistake', whether deliberate or accidental. Mistakes are preventable, whereas fuzziness in assessment is not. We need better language to capture this.

The other difficulty for public explanations is the probabilistic thinking involved in the standard error. Probability as a formal concept is difficult. Frequently, the probability level itself is not stated, leaving an undefined confidence interval (that is, undefined in the sense that no level of confidence is indicated). Also, whether or not the confidence level is indicated, the possibility that the true score could lie outside these limits (either way) is difficult to grasp (and accept). The possibility, however small, that in some cases the true score could lie way outside these limits is largely unimaginable to most people. Yet for individual students, we could be way off target with their standardised test results, just in terms of what the test asked them to do, let alone in terms of its validity.

3. *A different set of considerations informs performance assessments*

The notion of a true score has been a useful fiction for some purposes, especially for large-scale standardised testing for accountability and system monitoring, where we

might be justified in thinking that we are assessing a stable 'ability' (well, stable at that point in time, and also ignoring the biasing effects of the context). It seems less useful for other forms of assessment, such as public examinations, school-based assessments, portfolio assessments, and work-based assessments, where the focus is on what the student did with the task in hand. That is, we are less interested in some transcendent 'ability' than in the performance itself, what the student did. The performance may or may not be conditioned by (or indicative of) an underlying 'ability', but there are at least other factors involved, such as study and effort. In these assessments, the personal and situational factors are not sources of unreliability but part of what is being assessed. We don't play games concerning 'what if'; we are interested in how the student performed in the circumstances in which they were placed. What we report is how well they 'did what they did'. Perhaps, their performance is indicative of a more generalised ability, but we cannot really know about that.

However, what comes to the fore in such assessments is the judgment of the 'marker' or 'assessor' and the consistency of that judgment (as compared with other judgments). That is, the reliability question becomes 'have we given the performance a mark/result that is what it is truly worth and that is consistent with the marks/results given to other performances?'. That means that we are focusing on the judgments of the markers or assessors and their degree of consistency. Another way of saying this is 'how certain are we that we've given each student a mark/result that is appropriate for their performance?'. In other words, reliability in this situation is about the consistency of markers/assessors (or alternatively about the consistency of the marks/results themselves).

In this case, there is not a 'true score' in sight, nor any 'standard errors'. If we want to report the degree of consistency, we use a measure of agreement between two or more markers of the same student performances. However, we should note that a correlation coefficient is no use here because that takes no account of 'level', since it is based on standardised scores and therefore uses essentially only rank order—two sets of marks can be perfectly correlated but be systematically displaced along the scale and therefore represent completely different score-points (levels). Instead, we can use the percentage of (exact) agreements, or the average number of scale-points of difference between markers/assessors.

Both of these measures are used by the Queensland Studies Authority in reporting the outcomes of its post-hoc random sampling exercise for evaluating the quality of school-based assessments for the Queensland Certificate of Education (see QSA, 2000–2008).

4. *Different types of assessment demand different levels of reliability.*

The degree of certainty we demand of assessments depends on the use to which the results are put. Where life decisions and destinies are at stake, we may want to have higher certainty. On the other hand, where the results are interim or situated in the context of other assessments, lower certainty might be tolerated because there is opportunity to check with other assessments whether the results are believable and for other assessments to replace them if they are not.

One issue is the type of scale we are using, especially the 'grain-size' or number of categories of differentiation. While scales are sometimes thought of as continuous and therefore having an infinite number of points, in practice there are limits to the number of points it is reasonable to use. At one extreme, perhaps, we could choose to have, perhaps, 10 000 divisions along the scale. But, since it is unreasonable to expect exact precision in placing students along such a scale—assessment is not an exact science—we have to admit that adjacent scale points do not represent clearly differentiable performance. In fact, the measurement confidence interval is likely to be so great that we cannot properly distinguish students over a very wide range. Such scales would seem to offer more than they can reasonably deliver. Their assumed precision masks their real imprecision. As an illustration see Table 1, which shows how wide average confidence intervals can be, even for very high levels of reliability (here on an assumed 100 point scale with a standard deviation of 16). It should be noted that typical marker reliability for examinations rarely reaches .80.

Table 1: Confidence intervals (CI) for different reliability levels (for sd=16)

rel	$\sqrt{(1 - rel)}$	Sem*	68% CI	95% CI
.99	.1	1.6	±1.6	±3.1
.96	.2	3.2	±3.2	±6.3
.91	.3	4.8	±4.8	±9.4
.86	.4	6.4	±6.4	±12.5
.75	.5	8.0	±8.0	±15.7

*Sem = Standard error of measurement

At the other extreme is a two-category scale such as pass / fail or competent / not-yet-competent. Categorisation can be approached in two ways: either by applying a cut-score to a continuous scale or by placement through judgment. If we adopt a latent trait approach, cut-scores have measurement confidence intervals of the kind just described. Necessarily, there will be some classification uncertainties ('errors'), which become increasingly problematic the closer a score is to the cut-score. Some students will necessarily be 'misclassified' in the sense that, if we could repeat the assessment, they would switch categories.[2]

Alternatively, abandoning the latent trait approach and just focusing on the performance demonstrated on the occasion, we could treat assessment scores as 'accumulated points' or 'bankable credits', with the total score simply indicating a 'grand total of points awarded' or an 'account balance'. In that case, failure to exceed the cut-score is simply failure to reach a specified (though somewhat arbitrary) threshold of bankable credits. This seems to be the logic behind a lot of assessment, such as when a specified percentage of marks (say 40% or 50% or 65% defines a pass mark). Getting over the threshold is all that matters. It is unusual on such occasions to consider any underlying unreliability of the marks as an issue. Rather, it is expected that the 'umpire's decision' will be accepted as final. We might want to look at the replays but only post hoc and out of curiosity (or possibly to help future

[2] We will not consider here the issue of uncertainty associated with the cut-score itself. This uncertainty results from the process of determining the cut-score, usually calculated as the mean judgment of a sample of expert assessors.

improvement in conducting the assessment processes).

When classification is judgment-based, a different logic applies. In this case, it is not necessary to posit an underlying dimension, simply to define the characteristics of performance that satisfy the positive category, such as, the performance requirements for being considered competent. In the Australian competency-based assessment for vocational education and training, competent performance is recorded but performance that is not-yet-competent is unrecorded and the student can try again. To some extent, this reduces the pressures on failure. How certain we need to be in making a judgment of 'competent' depends on its consequential validity. Where a licence-to-practice results and incompetence can be life-threatening (such as for pilots, nurses and chefs), we might want to be very certain indeed. The degree of certainty is unlikely to be evaluated by means of test theory or evaluated at all. Instead, there are likely to be various checks and balances in the assessment so that it is a collective and cumulative decision and therefore more robust.

Similarly, ratings or grades can be allocated by direct judgment based on specifications of the characteristics typical of each rating or grade. Five category systems are common. There is a practical reason for this. There are severe limits to human cognitive processing capacity, most famously suggested Miller (1956) to be 7±2 categories. This number of categories can be extended by nested-judgment, for example, placement into five categories followed by subsequent judgment of high, middle and low within each category yields 15 categories. These kinds of judgments can be made holistically with respect to complex performances or collections of evidence such as portfolios. There is no need for 'marks'. Such grades are ordered categories and can be labeled descriptively rather than numerically.

As with judgments of competence, judgments of grades can be made against specified performance standards for each grade. With careful consideration, 'misplacements' can be minimised though not completely avoided. Again, the consequences influence how careful the judgment needs to be or how many checks and balances are used. The focus is usually on *judgment of the evidence* and therefore on differences among judges (assessors). We are less interested in replicability over different tasks and different occasions. The focus is on what the student did, that is, the evidence produced. It might be expected that judgment by another assessor would produce a limited number of differences by one grade level and essentially none by two grade levels. It is important to note that the aim is for concurrence among assessors *irrespective of their reasons they give* for placement in a particular grade.

Allocating grades by cut-scores along a continuum of marks is a different process, a generalisation of the single cut-score mentioned previously. The focus is on determining cut-scores that represent the boundaries between adjacent grades. The care with which this is done depends again on the consequences—how high the stakes.

Judgment of cut-scores may be guided by defined performance standards for each grade. However, clearly, these standards are not applied directly to each student's performance. It is assumed that the student's total score places them in the appropriate grade category. This is actually more like a 'bankable credits' approach than a 'latent trait' approach, since for an individual student the extra points that push a student over the boundary between one grade and the next may have nothing to do with the characteristics of performance at that grade level (unless the points were to

form a perfect Guttman scale).

Apart from the differences in underlying assumptions and processes, the important point here is that the allocation of grades—whether by direct judgment or by cut-scores—recognises the imprecision of assessment. A small number of categories is sufficient for differentiating different levels of quality in performance. Fine-scaled differentiation introduces an unsustainable level of precision where differences between adjacent score-points invite over-interpretation and are essentially 'noise'. Over-precision can undermine confidence in the assessment by being unreasonable and indefensible.

5. *What steps can be taken to improve marker/assessor consistency?*

It is nearly 100 years since Daniel Starch and Edward Elliott conducted their classic studies on reliability and gave a bad name to subjective rating of 'essays' by showing enormous variability among assessors. We have moved on since then and found ways to prevent the inconsistencies that result from letting assessor subjective judgment loose unguided and unchecked. The key is to adopt overall quality assurance/management, including clear assessment frameworks, assessor/marker training, explicit criteria and standards (rubrics), exemplars (for different grade levels), openness and transparency, and moderation processes (especially for school-based assessments). How much is invested in each of these depends on the stakes involved, public tolerance, public scrutiny and political commitment. In other words, there is no magical formula. It all depends on the local socio-economic-political context in which the assessment occurs.

Quality assurance has a variety of meanings. Some definitions from other fields but with relevance for educational assessment include:

- planned and systematic production processes that provide confidence in a product's suitability for its intended purpose
- all those planned or systematic actions necessary to provide adequate confidence that a product or service is of the type and quality needed
- monitoring and controlling procedures and results to insure the reliability of results.

Two of these definitions mention 'confidence' and the third 'reliability'. A key factor in all of them is taking action (planned and systematic) to produce a desired level of confidence or reliability. Also, there is recognition that the outcomes may not be perfect, but nevertheless suitable for intended purpose or of desired quality.

By considering assessment reliability as part of assessment quality assurance, we:

- ensure attention to all threats to confidence in the results
- situate reliability alongside validity so that they are not considered in isolation
- emphasise that confidence in assessment results can be developed not just left to chance.

Elsewhere, I have argued that it may be preferable to use the terms quality management or quality control when such processes deliberately monitor and adjust the assessment processes and outcomes through to their conclusion (Maxwell, 2006, 2007). Quality assurance is sometimes seen as a front-end process that sets up the

appropriate conditions and assumes that everything will then turn out all right. But since so much depends on assessor judgment this is rarely sufficient to produce confidence in the outcomes, especially where outcomes are public and high-stakes.

The Queensland system of externally moderated school-based assessments for the Queensland Certificate of Education at the end of Year 12 is a case in point, where the moderation processes are seen as contiguous with quality assurance. Descriptions of the principles involved are discussed elsewhere (e.g., Maxwell, 2007). These include the factors already mentioned for successful delivery of confidence in the assessment outcomes:

- an overall quality assurance/management system
- clear assessment frameworks (through syllabus specifications of desired learning outcomes and through approval of school assessment plans)
- assessor/marker training (through teacher workshops, extensive participation by teachers in moderation panels with consequent return of expertise to schools, and panel advice to schools on the quality of their assessments)
- explicit criteria and standards for judging the quality of student performance
- openness and transparency about these criteria and standards, both through sharing with students and sharing with other teachers—that is, it is expected that teachers be able to defend their judgments objectively when challenged
- panel moderation processes through which schools receive feedback on their judgments of the quality of student portfolios and are held to account for adjusting their judgments to achieve comparability with other schools.

This system is built on processes that emphasise consultation and consensus. There is no guarantee that such processes will work. That depends on teacher goodwill, trust and professionalism. However, the annual follow-up random sampling process, which is directed at uncovering where the system may be working least successfully, persistently shows high levels of comparability (QSA 2000-2008). A research study on the processes involved in judgments of the student portfolios produced in this system showed higher levels of reliability than are typical for public examinations (Masters and McBryde, 1994) (with the added advantage of higher validity).

Moderation is a broader concept than its application just in school-based assessment. The principles are applicable wherever there are multiple judges of standards and comparability is an issue (Maxwell, forthcoming). Ironically, there is a ‘chicken and egg’ issue here: comparability is unlikely without assessor proficiency yet moderation builds that proficiency through participation. Participation in the moderation system in Queensland is seen as a powerful means of professional development of teachers.

References

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cronbach. L.J., Linn, R.L., Brennan, R.L. & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school

effectiveness**,** *Educational and Psychological Measurement*, *57*(3), 373–399.

Gipps, C.V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.

Masters, G.N. & McBryde, B. (1994) *An investigation of the comparability of teachers' assessments of student folios*, Brisbane: Tertiary Entrance Procedures Authority (now Queensland Studies Authority).

Maxwell, G.S. (2006). *Quality management of school-based assessments: Moderation of teacher judgments*. Paper presented at the 32nd IAEA Conference, Singapore.

Maxwell, G.S. (2007). *Implications for moderation of proposed changes to senior secondary school syllabuses*. Brisbane: Queensland Studies Authority. <www.qsa.qld.edu.au/downloads/learning/snr_syll_rv_paper_imp_mod.pdf>

Maxwell, G.S. (forthcoming). Moderation of student assessments by teachers. In Baker, E., McGaw, B. & Peterson, P. (eds), *International Encyclopedia of Education* (3rd ed.). Oxford, UK: Elsevier.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.), 13–103. New York: Macmillan.

Miller, G.A. (1956). The magical number seven, plus or minus two. *The Psychological Review*, *63*(2), 81–89.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229–258.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12.

Moss, P.A. (2004). The meaning and consequences of reliability. *Journal of Educational and Behavioral Statistics*, *29*(2), 241–245.

Queensland Studies Authority (QSA) (2000–2008). *Random Sampling of Assessments in Authority Subjects: Annual Reports.* <www.qsa.qld.edu.au/assessment/2135.html>