

# **Reliability: A Practitioner's Perspective**

**Gordon Stanley**

**Pearson Professor of Educational Assessment**

**The University of Oxford**

**Director**

**Oxford University Centre for Educational Assessment**

Oxford University Centre for Educational Assessment

15 Norham Gardens, Oxford OX2 6PY, United Kingdom

Phone +44 (0) 1865 274002 Fax +44 (0) 1865 274027

email [gordon.stanley@education.ox.ac.uk](mailto:gordon.stanley@education.ox.ac.uk)

[www.education.ox.ac.uk/assessment/](http://www.education.ox.ac.uk/assessment/)

Let me start by making three points. Firstly I believe OfQual has been very sensible in identifying reliability as a key topic for further study and discussion. It is a topic of basic importance and one which needs better debate and understanding between all the participants in the process of education. The commissioned work programme will enable many issues which I wish to raise in my brief presentation to be debated and clarified.

Secondly, the opinions I express are personal ones and are raised to provide some stimulus for further discussion and debate. They represent some initial musing as I and some of my colleagues at the Oxford Centre for Educational Assessment have begun to think about these issues in the context of free inquiry.

I should say that we are only just starting the dialogue and discussion and have not yet thoroughly resolved our positions. We are asking questions about the most appropriate models to use in educational assessment with particular interest in the definition and measurement of error.

Thirdly the perspective I wish to take as the starting point for commenting on reliability is that which comes from a practitioner perspective.

Now for my personal perspective. As a practitioner recently returned to academia I am somewhat disturbed about the reported size of apparent misclassification in assessment regimes arising from simulations and arguments derived from classical test theory. I am particularly concerned if it results in a view that these assessments are essentially misleading.

In a few moments I will suggest that we are too often captured by this particular psychometric model and its variants when we approach the issue of reliability. But let me digress for a moment to indicate why my initial reaction has been to question the assertion that large misclassification commonly occurs.

For a decade I was responsible for the public defence of a relatively high stakes testing regime where some 140,000 Australian students undertook external

examinations each year. Given the high stakes associated with these examinations, and the fact that school assessment was weighted equally with the external test results, I would not have been able to sleep comfortably if I thought there was significant misclassification on the external examination.

After all, my signature was on their certificate of results and that meant the buck stopped with me. Certainly we would have had to face considerable criticism if the two sets of scores were classifying students very differently.

Maybe I should have had sleepless nights! Rarely however did we have schools raise concerns about differences between their examination and their school results. Why?

The answer is that the examination results fell within the range expected on the basis of school assessments. Indeed most schools in the system I was responsible for had their own model for predicting the final result which produced few surprises. There was even a commercial website which provided predictions for students.

On the basis of such experience, it is my belief that we need to approach the issue of error and reliability within a context of realism, i.e. in terms of what a careful analysis of the various and vigorous debates around assessment can tell us. We need to be confident that we are not misleading either in our criticism of current practices or in terms of the alternatives.

Current examination procedures have evolved over time and quality assurance standards have been adopted by most testing bodies. Such standards require that professional processes are in place in the development of examination questions, in marking procedures and in data handling.

Are such processes error free? Generally such processes are audited by professionals to ensure best practice is in place. Something which does not always occur in the classroom or at school level.

Ultimately all assessment is based on professional judgment about student performance. This is true independently of whether or not it occurs in formal external tests or in classroom assessment.

Professional judgment is required in the design of the assessment as well as in the marking and interpretation of student performance. As with all human judgments, professional judgments can be contested on many grounds. However, not all contesting is necessarily well-founded.

A feature of professional judgment is that there will be many occasions on which differences between judgments should not be considered error. Such differences are likely to represent some genuine and legitimate differences of opinion within that field of knowledge. It is not always the case that one professional judgment is correct and that the other is in error.

It is common knowledge that professional consensus in some areas is greater than in others. For example, from a perspective of public examinations (and for that matter in classroom assessment) judgments about essays in the humanities are more commonly going to demonstrate such variation. Generally this is accepted and provides a source of variation that is intrinsic to all assessment regimes.

Clearly we hope that in assessment in the school sector of education the extent of professional disagreement is relatively small in relation to the material being examined. With respect to other sectors my experience is that academics who are very demanding and critical of assessment regimes in the school sector are surprisingly loose in the technical design and operation of their own assessment process at university, and differences in point-of-view often dominate the assessment outcome.

At this point I am moving from familiar territory to new. Since arriving in England last year I have become aware of issues around the National Testing programme. National testing has provided an avenue of mapping student progression in terms of levels of achievement. Level descriptors indicate the knowledge and skills

expected to be achieved at certain stages of schooling. Levels are milestones on what is assumed to be a continuum of development. It is to be expected that students are not 'standing still' in their development with respect to levels, but improving over time. It is also important to acknowledge that levels represent arbitrary points (cut-scores) on a distribution of performance (scores).

Concern has arisen that such tests might at best result in a serious degree of misclassification of the levels of student attainment. A figure of 30% has filtered into popular discussion to the point where I have heard it said that all tests yield such misclassifications.

Such a large figure, if correct, is somewhat disturbing given the importance to students and teachers of having appropriate feedback about the level of their performance. However before we get too excited about this being the case we need to consider the context in which such claims are made and why further research and debate such as that initiated by OfQual is needed.

Classical test theory was developed in psychometrics when the major interest in psychological testing was on aspects of intelligence or ability which were meant to be relatively stable in character and to determine an individual's performance on different tasks. The observed score on any particular task was assumed to be composed of two elements: the stable or true score plus the unstable or error component. By making assumptions about the distribution of the error and by repeated observations estimates of the 'true score' on what is described as a latent trait could be obtained.

I think it is worth asking the question as to the appropriateness of this classical model and its variants to measures of educational achievement, when our focus is on developmental change rather than fixed positions on an unobserved or latent trait. The question needs to be asked because the language and modelling implicit in classical test theory and some of its variants leads straight away to seeking 'truth' and seeing variability as something to be considered as error.

I'm not trying to play a post-modernist trick here and reject the quest for truth, but rather to point out that models make assumptions which if not appropriate can lead to consequential misunderstandings.

So let us consider the issue of 'misclassification' and what we mean by it in the context of classifying human performance. Misclassification implies that there is error in the classification. How can we establish whether or not this is the case and how much should we expect?

All human performance is subject to variability, even in the absence of illness or trauma. On any one occasion we may be performing at personal best or personal worst. Or for that matter at any point on a range between these two positions. Such variability should not be considered error but a natural property of human performance.

Another aspect of human performance is related to the fact that at any point in time we may be changing and consolidating the depth of our knowing something or in our ability to do something. In classrooms this should be occurring at different rates in individuals at different times. Again this source of change means that observations at one point in time will be different at another not because of error of observation but because what we are observing is a product of a process of change.

The distribution of scores or individual positions on a continuum of development will of necessity reflect the sources of variability just mentioned. This of course has led to the problem that we cannot rely on repeated measures over time to give us an indication of the degree of stability of a currently observed performance and this has led to an over-reliance on internal consistency measures of reliability such as Cronbach's alpha being used in technical reports.

It is important to recognise that we can only estimate 'misclassification' on the basis of assumptions about what causes the distribution of observed scores and these assumptions made by psychometric models are unrealisable precisely in the real world of testing and examinations.

Another important point about classifying student performance on national tests into levels arises from the fact that whenever there is classification based on cut-scores on a distribution there will be students for whom the cut-point represents the area of their transition from not being at that level to being at that level. Their own performance will be varying around the cut point not because of error in measurement but because that is the nature of their evolving knowledge and skill at this point in time.

In addition the number of students classified as achieving a level will also be influenced by the process used to determine the cut-score. Cut-scores can be made on the basis of test-equating across years and then using the same point on the distribution or by making judgments about the content of student work at a given score to align with level descriptors. Across education and testing systems there are a variety of standards-setting procedures in use. Each procedure has advocates and critics. The consensus is that it is preferable to be consistent in the use of a given approach, especially when year-on-year comparisons are likely to be needed.

As I said at the beginning of this short presentation, it is important that we bring together practical experience and theoretical developments to provide better agreement about the nature of the reporting of assessments. The Ofqual research and communication programme will enable better understanding of good practice and provide for all stakeholders to know the strengths as well as the limits of examination and testing regimes.

In his lecture on examinations delivered in Lent Term at the University of Cambridge in 1880, Dr J. G. Fitch made some important admonitions which are still relevant:

This whole problem of examinations and the right way of conducting them and preparing for them touches very nearly the morality of the school life. Look well to the influence which examinations you use are having on the ideal of work and duty which your scholar is forming... Determine that whatever happens you will not pay too heavy a price for success in

examinations. Discountenance resolutely all tricks, all special study of past papers and of the idiosyncrasies of examiners, and all speculations as to what will and will not 'pay' to learn. It is because sufficient regard is not paid to these considerations, that many thoughtful persons now are fain to denounce examinations altogether, as the bane of all true learning, and as utterly antagonistic to the highest aims of a teacher. There ought to be no such antagonism. In their proper place examinations have done great service to education, and are capable of doing yet more. But they can only do this on one condition. Let us make sure that for us, and for our pupils, success in examinations shall not be regarded as an end, but as a means towards the higher end of real culture, self-knowledge and thoughtfulness.