No news is good news?

IAEA conference, Brisbane, 17 September 2009

England has had its share of exam controversies in recent years. New A levels caused headlines back in 2002 and last year we had troubles with our national curriculum tests.

From the ashes of the 2002 troubles came a real focus on making sure we had an agreed way to carry forward grade standards before the next generation of A levels came along. The 2008 national curriculum test problems were about delivery. In fact the controversies we have had have not really been about reliability. However, when Ofqual's Chair, Kathleen Tattersall, spoke at our launch last year one of her major announcements was that we would be launching a review of the reliability of assessments. We felt it was a neglected area and we ought to focus on it before it created headline news rather than after.

We have taken as our starting point for the meaning of "reliability", that it's about quantifying the luck of the draw. What if the candidate took the exam on a different day? Or the exam comprised a different set of questions? Or the script was marked by a different marker? Or cut-scores were set by a different panel? In any of these cases would the same grade have been awarded to the candidate?

More formally we have taken as our definition that reliability is: the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise.

We thought that this programme was needed as, certainly in, England, work on reliability in the context of exams and tests that has been carried out has been relatively isolated. It was certainly not part of routine monitoring. It has been partial, covering only certain facets of a small number of tests and exams. There has been little theoretical work and limited public debate over interpretation. As a result there is little public understanding of reliability.

The programme has been divided into three strands. Strand 1 is about generating evidence on reliability. Strand 2 concerns interpreting and communicating evidence on reliability. Strand 3 is about exploring public understanding of reliability and developing Ofqual policy on reliability.

Four projects have been commissioned in strands 1 and 2 and are due to report very shortly. We are expecting bids for work in five other areas. We have a technical group of six experts who help us steer this work – and the two Scottish representatives are here at this conference. Reports of all these technical projects will be published on our website and we hope to find opportunities at conferences like this one to share the findings more widely.

What I am talking about today are two of the first projects carried out under the public understanding strand. I'm grateful to my colleagues, Andrew Boyle and Annette Kinsella, for writing most of the paper that you have on your conference flash drive.

So is no news good news? Why have we previously shied away from communicating much about reliability in public? Well it is a complex idea that is hard to explain. We have worried in the past that negative news stories about it could damage public confidence. On the other hand shouldn't assessment organisations be transparent and communicate with the public about measurement inaccuracy? It should be possible to have a proactive programme of communication and public understanding of reliability and unreliability that would be beneficial both from the perspective of improving the ethical conduct of assessment organisations and making their job easier by acquainting the public with the truth about inaccuracy.

Since we launched the programme we have attracted a few articles in the press. The headlines haven't been all that bad and we have succeeded in starting to bring some of the issues to the attention of the public.

In the first project to investigate the opinions of several sections of the public in England in relation to reliability or unreliability in examination results, we commissioned the social research company Ipsos MORI to seek the opinions of teachers, students, parents, members of the general public and employers. The research was conducted using a workshop methodology. It involved session facilitators providing more substantive input to participants than would be the case in research methods such as focus groups. This approach was taken because it was felt prior to the field research that participants might well not have developed opinions about the issue under discussion. Therefore information on the topic was provided to participants to help them to develop views on reliability. It was understood that by providing substantial input to participants, the research ran the risk of biasing the participants' views. However, it was felt that this risk was less serious than the risk that participants might not have any view about inaccuracy in exam scores or grades. In this case we would have drawn a blank from the workshops.

The findings suggested a demarcation in the minds of the public between inevitable errors in the assessment process and preventable errors. The research participants appeared to accept that a certain amount of error was inevitable in a large examination system, but they could be intolerant of 'preventable errors'. However, these findings need to be interpreted carefully. It is far from clear that those concepts were the strongest explanators of the variations in respondents' opinions. Rather, it is at least arguable that

differences in opinions can be understood more clearly by referring to the group to which the opinion-holder belongs (teacher, student, parent, employer, examiner), the perceived agent of the error (examiner, exam board, student) and the consequence of the error. Sometimes participants appeared to be making a distinction between inherent and preventable error, but other times not. Also, culpability and assessment error appeared to be entwined issues.

Some teacher and employer participants in the research stated their differential attitude to error depending upon whether the error changed a student's grade or mark. They considered grade-related error to be more consequential than mark-related error. Participants' views about error could vary by group, and by the perceived cause of the error. For example, students and teachers could be intolerant of typos in papers while examiners could be more relaxed – taking the view that what was important was that any mistakes that did occur were rectified.

The findings on 'examiner-related error' show how the various strands are intertwined. For example, there is evidence that students are aware that some inconsistency between human markers is inherent when assessing subjects such as English. However, there are also statements that such inherent error should be minimised or even eliminated. Some participants suggested practical measures such as the double marking of papers or making markers do their work in marking centres rather than at home.

The final finding from the Ipsos MORI research concerns the word 'error'. The researchers reported that this term had some negative impacts when used with the public. In particular, the common meaning of that word, in contrast to its technical meaning, reinforced an inclination to treat unreliability as necessarily implying culpability. Further, the word grammar of 'error' tends to cause the issue of inherency, agency and culpability to be further muddied. For example, to speak of 'an error' seems to imply a single event, for which some person or thing must be responsible. In contrast, the slightly less common in public parlance, more 'technical' use of the word 'error' lessens the necessary connection with culpability. This degree of syntactical subtlety and potential for ambiguity suggests that this is not an ideal word to use centrally in an important public communication campaign.

The second piece of work was developed after Ofqual staff reflected on the experience of running the first opinion-gathering exercise and particularly the issues around the use of the word "error". It consisted of a session with the communications messaging consultant Blue Rubicon, which was used to produce a narrative for Ofqual staff to use when speaking about reliability and unreliability. The spur for this work came from the observation that it was not

easy to express ideas around reliability in terms that were informative yet consistent, concise and comprehensible. This was felt particularly to be an issue when different members of staff, for instance communicators, researchers and policy makers, would need to speak about reliability, or when third parties, for instance consultants or contractors, would need to do so.

Narratives of this type are often used as part of campaigns – for instance by companies promoting a product or service, or by political parties or other campaigning organisations. However, in this instance no campaign was being undertaken except – perhaps stretching the term – a public information campaign; trying to help the public to become more informed about reliability and unreliability. Indeed perhaps it's not such a bad idea to think about us using professional communicators to help us carry out such a campaign – explaining reliability to the public.

The document also provides an agreed set of terms with which to refer to key concepts in the programme. In particular, it settles on the term 'variation' (in scores, assessment procedures, etc) to describe the thing the reliability programme is talking about.

It was necessary to choose an alternative term to 'error' as this was too closely associated with culpability, and because it had an unhelpfully subtle word grammar. 'Variation' was felt at the session to be the best candidate to use, ahead of alternatives such as variance, uncertainty, discrepancy, inconsistency or clash. It is possible that some of these terms could also be used, while others would not be useful. 'Variance' would probably not be a good candidate, since it is confusingly associated with a statistical concept (which has a certain, but not complete, relationship to reliability). 'Clash' is probably not close enough in meaning to unreliability and also has the potential to provide incendiary headlines. However, members of the reliability programme may try out some or all of the other terms when speaking in public about reliability.

So where have we got to so far? Can non-specialists understand the complex area of reliability? Do you have to educate them? The Ipsos MORI research does – in places – support the view that the public can formulate and debate sophisticated validity arguments. This may mean that, in seeking the public's views about inaccuracy, researchers will have to address a dilution from the purist notion of reliability and use a notion of measurement inaccuracy covering reliability, validity and comparability. This seems preferable to seeking to 'educate' respondents in reliability theory.

What's next for the Ofqual programme? Well, we have made a start but probably only scratched the surface of this complex area. We are hoping that

on the back of today we may get to hear how others around the world have tried to tackle this communications challenge. Do you report results as a range or with a standard error of measurement? If so, does anyone take any notice of it? What do you do to explain reliability publicly? Ofqual's next major step on the communications side of the programme is to use a written questionnaire so that we can access a larger sample of people than the workshops allowed. But of course, being a sample, we will have to ask when we see the results, how reliable are they? Thanks for your attention.