The Reliability Programme

Technical Seminar Report – 7 October 2009

Date: 21 December 2009
Product code: OFQUAL/09/4638

Contents

Introduction

There are two sorts of greetings cards for the exam season: ‘Good luck in your exams’ and ‘Well done in your exams!’ But which one is it? Do candidates get good results because they are lucky or because they have worked hard to reach their potential?

Most people would say it is down to the student but, of course, a bit of luck is involved. Did the ‘right’ questions come up? Was the pollen count low for a hay fever sufferer? There are other factors that can impact on a result that are beyond the candidate’s control, such as the way the paper is set and marked, and how the boundaries between grades are decided. All these factors impact on the reliability of the grade a less exact measurement than we might think.

Stringent processes have been put in place to ensure that the exam boards check and recheck the accuracy of the results they produce, and inspections are carried out by the regulators – formerly the Qualifications and Curriculum Authority and now Ofqual. Despite this, there are judgments being made that are not set in stone, such as the number and spread of questions. Then there is the possibility that someone whose score was very close to a grade boundary could, on another day with a different paper, end up on the other side of it. These small differences – perhaps only one mark – can mean a better grade for some candidates and a worse one for others.

These were issues addressed by leading experts on assessment at a technical seminar held at the University of Warwick on 7 October 2009. The discussion reported below formed part of Ofqual’s ongoing Reliability Programme, a two-year investigation into factors that can undermine the ability of examiners and assessors to produce a grade that truly reflects a candidate’s achievement or mastery of a subject. Kathleen Tattersall, the Chair of Ofqual, has likened the programme to a ‘health check’ on the system.

If it is proved that some assessment results are less secure than others, then should a ‘health warning’ be attached to help the users – such as universities and employers – make decisions about candidates? Or would it undermine public confidence in the qualifications, rendering them less useful? Perhaps we need to change the methods currently used to assess performance in order to achieve greater precision? If we did, how much would this cost in terms of money and hours spent in the exam hall, or on the validity of the assessment?

The seminar considered the different forms of reliability that impinge on the validity of results and discussed what might be done both to improve that reliability and to inform the wider public about it.

Two kinds of studies were presented to inform the debate. First, those carried out with actual children, comparing and contrasting their results with their performance in different national curriculum tests, or comparing and contrasting their performance on national curriculum tests with their performance based on teacher assessments. Secondly, those that used theoretical or empirical models to predict the likelihood of children being awarded the same national curriculum level as that actually awarded.

It is the first time in this country that examiners have been encouraged by an exam regulator to publicly articulate and debate the reliability of the assessments they make so the users – universities, employers, schools, teachers and pupils – can have a better idea of the likelihood that the candidate could be better, or worse, than the grades awarded.

The reliability programme is a work-in-progress and the views expressed by participants in the seminar should be seen in this context, as a contribution to the debate and not a final position. The opinions expressed belong to the individuals and do not necessarily reflect Ofqual's view or the view of the organisations to which they belong.

Eliminating chance

Exams and practical assessments test candidates' performance. This is a reflection of innate ability to a certain degree, but also involves other factors, such as how well they have prepared and their ability to perform in a particular context. Very often, however, the grades that are issued are used as if it were possible to stick a probe into a brain and decide how good a candidate is at English, history or physics. As one delegate suggested, you can have the most accurate thermometer in the world, but how well it reflects the temperature in a room depends on where you put it.

It is the job of examiners to eliminate chance as much as possible from the question paper and subsequent marking. But even the most rigorous quality-control systems to ensure all candidates are marked fairly cannot ensure that a candidate would get the same grade on another day and in a different test. In subjects where there is a larger element of interpretation, such as English or fine art, there may be differences in the way individual markers or chief examiners apply the mark schemes.

To illustrate the variability in exam results, one participant in the seminar suggested that if a syllabus required 100 pieces of information and a candidate knew exactly half of them then it was down to chance that the 'right' questions came up. Those who knew all of the material would have been unaffected, as would those who knew none of it. Those in the middle who knew only part of the syllabus were the ones most exposed to the luck of the draw when questions were randomly drawn from a pool.

This element of chance in the selection of questions could impact on the reliability of an assessment. Generally speaking, the longer the test, the more reliable it will be. ‘It’s a question of risks and costs,’ said Mike Cresswell, director general of AQA, one of the three exam groups in England. 'We could move to a very high level of reliability, but at an immense cost to the student, in what it will take out of them, and to the system, in terms of resources. These are questions about what you want the system to do, the resources you are prepared to put in and the value you think society gets out of it.’

Another aspect that could impact on reliability is the practice of turning raw marks into grades or levels. While it allows easier comparison between candidates, it inevitably involves delicate decisions about where the cut-off point for each grade should fall, participants at the seminar agreed. Grading introduces the concept of misclassification and it is those candidates nearest to the cut-off points who are most vulnerable to being ‘wrongly’ graded, up or down. A higher number of grades increases the distinction between candidates, but also increases the risk of misclassification.

People who were just below the cut-off line on one occasion might have been just above it on another. An education researcher suggested that a solution might be to use a numerical scale – say up to 600 or 700 – using the raw marks and weighting them to give standardised scores that could be compared with those for other subjects. It would then be more obvious to users that a score of 345 was not really much different from 340 or 350. Others warned that the users of the qualification – such as universities or employers – would be likely to impose their own arbitrary cut-off points on a numerical scale as a substitute for grades, and this would move the dilemma, not solve it.

According to Colin Robinson, an independent consultant: ‘All we can do as testers is make the result as accurate as we can within the boundaries of what the student has produced and the structures we have set up to try and make them as accurate as possible. But there will always be a level of imprecision and that is something we have got to educate the outside world about. We must get away from this idea that we are trying to stick the probe into the head of the candidate. The assessment has to be good enough for its purpose. We have to get a handle on what is good enough.’

Openness about reliability

The seminar was divided on the meaning that the word 'reliability' should carry in terms of assessment. Two of the three exam groups in England – Edexcel and AQA – felt the term and the scope of Ofqual's reliability programme should be limited to the factors within the examiners' control. Malcolm Hayes, from the Pearson Group, which owns Edexcel, said: 'If a pupil doesn't get the right mark for what they did they were misclassified. If the thresholds were in the wrong place they were misclassified. If the mark scheme was wrong, then we've got a systematic error and there is a problem. If there are clerical errors in the reporting we have got a problem.'

But what might have been – if the candidate might have done better on a different day or different questions had come up – was not real misclassification and could hardly be called error. 'That is just what happens and when talking about reliability I think we need to separate the real misclassifications from the might-have-beens,' he said.

There was strong agreement, however, on the importance of being more open with the public about the factors that can affect the accuracy of a result. How likely was it that a candidate would have got the same grade on a different paper with different questions? How likely was it that the student would have been awarded different grades if marked by a different examiner? How many candidates would have been affected, up or down a grade or level, by an adjustment to the cut-off point? Did all the questions contribute evenly to the overall purpose of the assessment or were some of them more random and should therefore have been given less importance? Did the test measure the performance of those at the top as accurately as those in the middle? These were all aspects that should be discussed, whether or not they are included in any stricter definition of 'reliability', delegates agreed.

Several speakers drew on analogies from the medical world. Just as people were prepared to acknowledge that a doctor's diagnosis was not straightforward and set in stone, so the users of results could take on board information about the reliability of a test or exam score and adjust their use of it accordingly, it was suggested. People accepted that a doctor's diagnosis was influenced by such things as the patient's accuracy in describing the symptoms and the reliability of the machinery used to detect abnormality.

Dr Cresswell said he had no objection, other than the practical difficulty of routinely churning out statistics, to figures being published about reliability and it would probably 'ginger up' public confidence. However, he agreed with Mr Hayes that the different reliability issues needed to be separated: 'I'm advocating putting some clarity in the debate by having a cut between the immediate procedural matter of turning a performance on an occasion into a grade, and the wider, longer-term debate about how much unreliability or risk is acceptable and how much money you

want to spend to address it. What do you generally think is a fair way of rationing the resources of places in higher education or scarce, well-paid jobs? Because that is what qualifications are there to do, to help people to assign life chances to individuals."

Reliability and national curriculum tests

Professor Dylan Wiliam, the deputy director and professor of educational assessment at the Institute of Education, was not present at the seminar, but his work was keenly discussed. A study he had carried out on the reliability of the national curriculum tests was published in 2001 by the Association of Teachers and Lecturers. The report was picked up by the *Times Educational Supplement*, which used an extract for its quote of the week on 30 November 2001:

> The proportion of students awarded a level higher or lower than they should be is at least 30 per cent at key stage 2 and may be as high as 40 per cent at key stage 3.

If at least three in 10 children could be given the wrong level for the national curriculum tests they sit at the age of 11, then how much should the results be relied on to judge the performance of schools? If right, does it mean the results for GCSEs and A levels are similarly flawed (in which case, universities and employers might as well switch their selection procedures to a lottery)? If wrong, do studies relying on statistical models instead of individual candidates do anything more than scare the public?

Dr Paul Newton, the director of the Cambridge Assessment Network, a keynote speaker at the seminar, sought to put Professor Wiliam's work in context of the current debate. In doing so, he said he would be using the word 'error' in the same way as it is used within educational measurement and not in its more popular meaning to denote a mistake. In educational measurement, an error occurs whenever a student is awarded a result, such as a grade or a level, that differs from the one they ought to have been awarded given their true level of attainment. This kind of error does not necessarily imply that someone has made a mistake. A student might have performed below his or her ability on the day of the test, for example. Or it might have been a borderline case where there was ambiguity as to whether the performance was worthy of a particular grade.

Several studies had already been carried out on reliability and they did not all address the same aspects. He warned of the importance of being clear about the factors explored in each piece of research to avoid confusion over the findings. A study looking at agreement between examiners marking the same performance is quite different from one investigating the chance of results from a particular test composed of a specific set of questions. Outcomes from different studies cannot,

therefore, be compared directly. He identified two types of research looking at reliability in terms of:

- consistency – how likely is it that students would be awarded different grades or levels if they were marked by a different examiner? Or how likely is it that they would be awarded different grades or levels if they sat a different test?

- correctness – how likely is it that students are awarded incorrect levels or grades given their true performance scores or their true test scores?

‘Marking may be perfectly consistent but still inaccurate, just as an extremely good watch set to run 10 minutes fast is both 100 per cent consistent and 100 per cent inaccurate at the same time,’ he said.

He identified a third kind of study that is looking beyond the reliability of a given assessment to its ‘validity’ – how well a test or assessment seeks out the truth about a candidate’s ability. To capture all aspects of test error you need to go beyond consistency and correctness to consider the real accuracy of a set of scores.

‘None of the studies of reliability are going to address the real question the public wants answering, which is how likely it is that students are going to be awarded incorrect levels given their true construct scores. By “construct score”, I mean what the student would get if you stuck an attainment probe into their brain.’

Unless researchers are clear about what their studies can show there is a danger that their work will be misunderstood. For example, some studies define the true score as that set by the chief examiner while others say the true score is defined as the mean of the scores that an individual would get if he or she took the test an infinite number of times.

Professor Wiliam had not set out to judge the absolute validity of the tests – how well the levels awarded matched the children’s underlying levels of attainment – but, simply their reliability. This is classification ‘correctness’ in the sense used by Dr Newton – how likely it is that students are awarded incorrect levels or grades given their true performance scores or their true test scores.

Professor Wiliam’s work was based on a statistical simulation that began with an imaginary set of marks with a distribution similar to national curriculum test results, which he called the ‘true’ scores. He then used a statistical tool to generate ‘observed’ scores, which were his original true scores altered semi-randomly so that the true and observed scores would differ to an extent that might be expected from a test of a certain level of reliability.

He decided that, based on statistics and information about similar tests around the world, the reliability of the UK's national curriculum tests at that time was likely to be between 0.8 and 0.85.

His next step was to apply the cut-off points for each of the national curriculum levels covered by the test to both his true scores and his observed scores. Finally, he compared the proportion of children put into each level on both sets of data.

Having assumed a level of reliability of between 0.8 and 0.85 – which Dr Newton said was not an unreasonable estimate for tests such as the key stage 2 English tests – Professor Wiliam estimated that the proportion of candidates put into the wrong level for the new, observed scores was between 27 per cent and 32 per cent. 'It is quite a crude simulation, as I am sure he would accept, but to me it doesn't seem to be a completely unreasonable simulation. You'd think it ought to be in the right ballpark,' said Dr Newton.

On the other hand, work done by the National Foundation for Educational Research had suggested the likelihood of error to be half that described by Professor Wiliam's study.

'We are trying to get some good answers, but I think what we have so far are crude approximations and there is a lot more work to be done,' concluded Dr Newton.

Statistical issues

Malcolm Hayes, the second speaker, said that talk of large numbers of children being misclassified was very uncomfortable for the staff at Edexcel, who had been involved in setting and marking some of the tests. 'The figure seemed very surprising to me and I decided to check it out,' he said.

Professor Wiliam had calculated reliability on the basis of all the people who took the test – the population – and then expressed it in terms of the individual, he claimed. 'Population parameters don't say anything about the individual. If you use statistics to decide the reliability of the test for the whole of the population it tells you nothing about what happens to the individual. You are on dodgy ground saying we can go from a population correlation to work out what happens to an individual. The reliability coefficient of .8 or .9 is based on a data set. You don't actually care who took the test. It is the same status as a weather forecast. It might happen and it might not happen.'

Studies claiming misclassifications of 30 or 40 per cent of pupils were based on unknown factors, such as what might have happened if different questions had come up or the candidates had done it on a different day. 'I don't think that is a valid misclassification. That is just what happens,' Mr Hayes said.

He had carried out his own statistical simulation, taking into account the various variables such as test mean score, standard deviation, reliability coefficient, number of thresholds and position of thresholds that affect the level of misclassification, and got a possible misclassification of 10 per cent, which showed how you could play around with figures. Reliability statistics did not apply to each candidate equally, which meant a lot depended on the spread of candidates around each grade or level cut-off point. ‘We have to be very careful about the assumptions that we make if we are going to do this kind of modelling,’ he said.

It would be difficult to calculate the true score in terms of the average score that someone would get if they took the test repeatedly because taking it more than once was bound to affect their performance. Lessons learned from looking at the accuracy of teacher assessment showed that, though there was quite a lot of agreement, there were a significant number of cases where the same teacher had made a very different assessment on the same child on two occasions. The teachers had not necessarily misclassified the child because things could change in the week or month between the assessments. ‘Perhaps the child did some revision or something happened on the first occasion that stopped them doing something on the second, or they produced a new work which influenced the teacher,’ he said.

‘We have a duty of care and we have to make sure we construct fair tests and have equality of access and make sure we have accurate marking and reporting. But we also have to maintain faith and make sure that the wider public believes that what we are doing is the best that we can do. If we go around telling them that we misclassify 30 per cent of people based on what I consider to be some flimsy logic then we don’t just do ourselves a disservice, but we do the wider public a disservice. If they lose faith in the testing systems that we have got then they take away a lot less from them.’

Factors affecting reliability

Noticeably absent was any data on the reliability of the more complex A levels and GCSEs. Sandra Johnson, from Assessment Europe, said that while examiners here were working within a clearly defined syllabus, unlike assessors for the national curriculum tests where teachers provided different schemes of work and children experienced varied curricula, reliability remains an issue, and should be properly researched.

One way forward would be to list all the possible factors that could affect reliability and see how they interacted for each assessment. She said: ‘A candidate’s ability will be the main determinant of an examination grade. But there will be other influences, too, including the questions set and the markers who mark the responses. We need to quality assure our tests and examinations, so that they are as fair as possible to the highest number of candidates.’

Sarah Maughan and Ben Styles from the National Foundation for Educational Research (NFER) reported on a study that had been undertaken for Ofqual. The study compared the results of a group of pupils who took a live key stage 2 science test with their results when they trialled the following year’s test in advance for the QCA.

‘The levels that the pupils were awarded on each of these two versions of the tests were compared using a variety of statistical methods and the internal reliability of the tests was also calculated. Results from the analyses indicate that the tests have reasonably high levels of reliability on the different measure for tests of this type.'

The analysis done over five years showed that between 83 per cent and 88 per cent of pupils were likely to be given the ‘correct’ level and that the tests had become more reliable over the last three years, perhaps reflecting changes to the way the pre-testing was carried out.

Tom Benton and Dougal Hutchison, also of the NFER, emphasised that even the best tests can never be completely precise. 'The topics you have revised may not come up or you may be feeling a bit below par, so that you don’t do as well as you might have done. Alternatively you may strike lucky on both counts and do really well,' said Dr Hutchison, the NFER’s chief statistician.

Reporting on a study into the reliability of a key stage 2 English pre-test - which groups of pupils sit after their actual test to help examiners set the standard for the next year’s live test - they stressed that any given result is a combination of the effects of time, test material, marking procedures, and scoring methods. It is not possible, therefore, to reproduce an assessment procedure completely.

Their study asked the question: 'If we did ‘it’ again, how similar a result would we get?' They compared a range of ‘internal’ methods, which tried to pin down the variability of the pre-test due to only being able to cover part of the material, and ‘external’ methods, which compared the result with performance on other assessments, such as the ‘live’ test result and teacher assessment. They found that the less close the replication was in terms of time or material, the lower the estimate of the reliability. They emphasised, however, that a low estimate of reliability might not necessarily mean a poor test. A teacher’s estimate could differ from a test result because the teacher was assessing the individual’s general competence, while the test result reflected the performance at a moment in time. Performance is also likely to differ between the pre-test and live test because the children would be more motivated for the latter.

Grouping scores into a smaller number of grades – in this case levels two, three or four – also decreased reliability. If an individual was very close to a boundary, then it could be a 50:50 chance of them being awarded the higher or lower level. For the pre-test under consideration, the study’s best estimate of the proportion of pupils correctly classified by the pre-test was 82.6 percent.

Publishing data

The idea that Ofqual might produce a format in which examining bodies could make many different reliability statistics available to the public gained some support from the afternoon’s working groups. The working groups also explored the idea that a database of reliability be collected, using anonymous statistics from the awarding bodies. This could be used to inform the debate about the levels of reliability for different qualifications.

‘The press might have a field day to begin with, but workshops carried out by Ofqual have suggested that the public has a fairly sophisticated view,’ said Andrew Boyle, Ofqual’s Head of Assessment Research.

‘They will have no truck with markers who do not mark accurately, but if on the day their son was asked something he had not revised for and got a B rather than an A they would probably say that was fair enough. People are not expecting examinations to be probes in the brain, even if the results are sometimes used that way.’

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by The Office of Qualifications and Examinations Regulation in 2009.

The Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk