The Reliability of Results from National Curriculum Assessments, Public Examinations and Qualifications

An interim report of the Ofqual Reliability of Results Programme

Dennis Opposs and Qingping He

July 2010

Ofqual/10/4756

Contents

Abstract

The Office of Qualifications and Examinations Regulation (Ofqual) in England is carrying out a two-year research programme to investigate the reliability of results from National Curriculum assessments and public examinations in order to develop regulatory policy on reliability. Since it started in 2008, the programme has made substantial progress in the areas of:

- generating evidence of reliability of results from National Curriculum assessments
- reviewing theories and models used to produce and interpret reliability measures
- exploring public perceptions of unreliability in examination results.

This paper provides a brief summary of the results obtained from the programme to date and the research projects that are currently in progress. It also outlines potential areas that, it is expected, Ofqual will explore during the development of reliability policy.

Background

England is a country in which much educational assessment takes place (Black and Wiliam, 2005). There are the following major assessment occasions in the English system:

- whole cohort National Curriculum assessments for 11-year-olds in English and mathematics
- public examinations, including standardised qualifications typically taken at the ages of 16 and 18
- large and diverse suites of vocational qualifications, which may be taken by candidates in schools, further education institutions or in workplaces as part of on-the-job training.

Some assessment systems (such as the National Curriculum assessments and the 16-plus examinations – mainly GCSEs) produce data that are also used for public evaluations of institutions and individual professionals, in addition to providing information about individual student attainments in specific subject areas.

Reliability, in educational measurement terms, refers to the consistency of results on a given measure from repeated measurements under equivalent conditions and is an important indicator of the quality of an assessment. Although results have a huge impact on learners' lives, as with any measurements, assessment results contain inaccuracies. Although it is generally realised that assessment results contain inaccuracies and substantial work has been carried out to study the reliability of assessments, there is considerable variability in how measurement uncertainty is represented and reported in different parts of the world (see Bradshaw and Wheater, 2010). While in the United States and some other countries assessment results are sometimes reported as raw scores or scaled scores together with the associated standard error of measurement (Bradshaw and Wheater, 2010; Phelps et al., 2010), in England assessment organisations tend to report learners’ performance levels or grades for National Curriculum assessments and public examinations without any indication at all of the likely error-rates involved. However, it has been suggested that there is a duty to communicate about the reliability of assessment results to the public (see, for example, American Educational Research Association (AERA) et al., 1999, Standard 2.1; Newton, 2005a, 2005b; Phelps et al., 2010). It is important that the degree of inconsistency in test and examination results is investigated, interpreted and understood appropriately.

There has been little sustained and systematic attempt to evaluate the reliability of results from England's assessment systems, and little understanding of the public's knowledge of and attitudes towards unreliability in assessment results. To address this, Ofqual is conducting a two-year research programme, involving:

- generating evidence of reliability of results from National Curriculum assessments, public examinations, and other qualifications
- interpreting and communicating reliability evidence
- exploring public perceptions of unreliability in exam results
- developing policy to regulate the reliability of assessments

with a view to improving the national assessment systems further.

The Ofqual Reliability of Results Programme

Aims and objectives

As indicated earlier, the primary aim of the Ofqual Reliability of Results Programme is to gather evidence to inform Ofqual on developing regulatory policy on reliability. The main objectives of the programme include the following:

- to generate evidence of reliability of results from a number of major National Curriculum assessments, public examinations and qualifications offered by assessment agencies and awarding organisations in England
- to stimulate, capture and synthesise technical debate on the interpretation of reliability evidence generated from this programme and other reliability studies
- to investigate how results and the associated errors are reported internationally, and what procedures are adopted by assessment providers to communicate results and measurement errors to the users
- to explore public understanding of and attitudes towards assessment inconsistency
- to stimulate national debate on the significance of the reliability evidence generated by this programme and by other reliability studies
- to help improve public understanding of the concept of reliability
- to develop Ofqual policy on reliability.

Programme structure

To achieve the aims and objectives set out for the programme, the programme is structured into three strands:

- Strand 1: Generating evidence on the reliability of results from a selection of national qualifications, examinations and other assessments in England through empirical studies.
- Strand 2: Interpreting and communicating evidence on reliability.
- Strand 3: Investigating public perceptions of reliability and developing regulatory policy on reliability.

Advisory groups

A Technical Advisory Group (TAG), made up of educational assessment experts, was appointed at the start of the programme. The group has been advising work on Strands 1 and 2, including advice on the methodologies to be used and the selection of qualifications, examinations and other assessments to be investigated. They are also responsible for reviewing reports from research projects funded under this programme.

A Policy Advisory Group was also appointed to provide advice on work for Strand 3 of the Reliability Programme. The Policy Advisory Group is made up of representatives from a wide range of stakeholders, including assessment experts, assessment providers, employers, communications experts, teachers, students and parents. The Policy Advisory Group has been advising Ofqual on engagement with key stakeholders and communication of reliability evidence to the public.

Activities

A variety of activities have been undertaken to meet the programme objectives, including:

- commissioning research projects to awarding organisations and research institutions to generate evidence on reliability of results from National Curriculum assessments, public examinations and vocational qualifications; reviewing measurement theories and models used to study reliability; reviewing techniques used for producing and interpreting reliability measures; gauging public perceptions on reliability; and investigating international approaches to the representation and reporting of assessment results and measurement errors

- participating in national and international conferences to exchange ideas and experiences with other assessment researchers, policy makers and practitioners on issues related to reliability

- organising technical seminars involving assessment experts and communications experts to discuss issues related to reliability to reach consensus on the interpretation, evaluation and communication of reliability evidence to the wide public

- participating in and organising public events to raise public awareness of assessment reliability and to help the public to understand the concept of reliability.

Summary of results from the programme to date

Reliability evidence

Key Stage 2 science assessment

As part of Strand 1 of the Reliability Programme, the National Foundation for Educational Research (NFER) was commissioned to conduct a research project studying the reliability of results from National Curriculum assessments in science, which are administered to all pupils in England at the age of 11 over a period of five years. This study provided robust evidence of reliability of results from Key Stage 2 science assessments (see Maughan et al., 2009). The researchers studied the internal reliability of individual tests used and compared the consistency of results from different versions (or parallel forms) of the same test. A variety of reliability indices, including internal consistency coefficient for individual tests, correlation coefficients between parallel forms, Kappa statistics for individual tests, and classification accuracy and consistency indices for individual tests and between parallel forms, were produced using widely used procedures. Measures of reliability have been appropriately interpreted in the context of National Curriculum assessment in England.

The Key Stage 2 science assessment consists of two papers (Paper A and Paper B). These papers are each made up of 40 marks and consist of a mixture of objective, short answer and longer response questions. The papers each have a time allowance of 45 minutes. Scores from the two papers are combined to produce a composite score, which is then used to assign a level representing the achievement in science by the pupil. Each year a subset of pupils takes an equivalent test, which is used as the following year's live test shortly before the current year's live test. By using the levels from the pre-test and live test to produce a cross-tabulation of results for the pupils for each year studied, the researchers were able to investigate the consistency of the levels awarded to the pupils from the two tests (see Maughan et al., 2009, for detailed description of the level-setting procedures used for the live test and the pre-test). As an example, Table 1 compares the percentages of pupils that were assigned to different levels by the 2004 live test (A+B) and the 2005 pre-test (A+B). The percentages of pupils who were awarded the same level on each version of the test (the bold numbers in the table) can be added up to provide an indication of the overall consistency between the live test and the pre-test, which is 73 per cent in this case.

Table 1 Percentages of pupils who were classified into the different performance categories by the 2005 pre-test (A+B) and 2004 live test (A+B) (from Maughan et al., 2009)

	2004 Live test (A+B)				
2005 Pre-test (A+B)		Below L3	L3	L4	L5
	Below L3	**<1**	1	0	0
	L3	<1	**8**	4	<1
	L4	<1	4	**29**	9
	L5	0	0	9	**35**

Table 2 shows the percentage agreement in classification by the tests for each of the years investigated. There would appear to have been an improvement in the classification consistency of the tests over the five-year period, with the last three years being better than the first two. It was shown that almost all of the remainder of the pupils were classified into the adjacent levels, with less than 1 per cent of pupils being awarded more than one level different in four of the five years (see Maughan et al., 2009).

Table 2 Percentages of pupils who were classified into the same performance categories by the pre-tests (A+B) and 2004 live tests (A+B) (based on Maughan et al., 2009)

Tests	Consistency (%)
2005 pre-test vs 2004 live test	72
2005 pre-test vs 2004 live test	74
2007 pre-test vs 2006 live test	79
2008 pre-test vs 2007 live test	79
2009 pre-test vs 2008 live test	79

Maughan et al. (2009) also computed the correlation coefficient for each pair of the pre-tests and live tests. Table 3 lists the raw score correlation coefficients for each pair and Cronbach's alpha for individual tests. Values of the Cronbach's alpha for live test papers have been published by the Qualifications and Curriculum Development Agency (QCDA) on its website.

Table 3 Raw score correlation coefficients between the pre-tests and live tests, and Cronbach's coefficient alpha for individual tests (based on Maughan et al., 2009)

Year of comparison	Tests	Correlation	Cronbach's alpha
04/05	Pre-test (A+B) vs live test (A+B)	0.85	0.92 vs *
05/06	Pre-test (A+B) vs live test (A+B)	0.81	0.93 vs 0.92
06/07	Pre-test (A+B) vs live test (A+B)	0.85	0.92 vs 0.93
07/08	Pre-test (A+B) vs live test (A+B)	0.86	0.94 vs *
08/09	Pre-test (A+B) vs live test (A+B)	0.88	0.94 vs*

* Cronbach's alpha for live test was not available.

In classical test theory (CTT), the correlation between two parallel forms is the reliability of the test, and the correlations between the pairs of the tests are generally lower than the Cronbach's alpha values for individual tests. This is expected because Cronbach's alpha is only an internal reliability measure of the test, which only reflects the combined effect of errors from sources associated with items in the specific test and markers. The correlation between two tests, on the other hand, reflects the contributions to the overall inconsistency in results from both test items in the individual test forms and markers, and the occasions under which the tests were administered (that is, including both test item and marker-related and occasion-related errors).

Maughan et al. (2009) also investigated the decision accuracy and consistency of results based on a single administration of the test using item response theory (IRT). Decision accuracy is defined as the proportion of pupils that would be awarded the same performance levels by both the true scores and the observed scores of the pupils on the test. Decision consistency refers to the proportion of pupils that would be awarded the same performance levels by two sets of observed scores on two parallel forms of the same test. For the 2009 pre-test, the decision accuracy and consistency were estimated to be 0.89 (or 89 per cent) and 0.84 (or 84 per cent) respectively. Misclassification, which is defined as 1-decision accuracy, is frequently used to indicate the level of inconsistency in awarding the performance levels by the true scores and observed scores. For the 2009 pre-test, this is 0.11 (or 11 per cent).

Maughan et al. (2009) also used Newton's (2009) concept of classification 'correctness' and the method that Newton proposed to investigate level misclassification further. Classification 'correctness' is defined as the probability that a pupil is awarded the 'correct' level on the basis of just one of the two test administrations and where 'correctness' is taken to be correspondence with a pupil's 'true' level. Newton (2009) proposed a simple relationship between the inconsistency (defined as the proportion of pupils who were awarded a different level by the two versions of the test – the live test and the pre-test), and the classification correctness *p*: inconsistency=2*p*(1-*p*). Newton's formula can be applied to tests that can be assumed to be strictly parallel, because *p* is assumed to be the same for the two versions of the test. Table 4 lists classification correctness and misclassification (defined as 1-classfication correctness) for the individual tests.

Table 4 Classification correctness (%) and misclassification (%) for the pre-tests and the live tests (based on Maughan et al., 2009)

Year of comparison	Tests	Correctness (%)	Misclassification (%)
04/05	Pre-test (A+B) vs live test (A+B)	83	17
05/06	Pre-test (A+B) vs live test (A+B)	85	15
06/07	Pre-test (A+B) vs live test (A+B)	88	12
07/08	Pre-test (A+B) vs live test (A+B)	88	12
08/09	Pre-test (A+B) vs live test (A+B)	88	12

Table 4 shows that between 83 per cent and 88 per cent of pupils would be correctly classified by the individual tests. It is noted that for the 2009 pre-test, the classification correctness estimated using Newton's method is closely similar to the classification accuracy estimated using the IRT method (0.89 or 89 per cent). The level misclassification figures are substantially lower than the 30 per cent figure suggested by Wiliam in 2001 (see Wiliam, 2001). As suggested by Maughan et al. (2009), the reason for the discrepancy between their level misclassification figures and that of Wiliam’s is not clear, although they recognised that the different results were produced using different methodologies and that Wiliam’s work was based on mathematical simulations rather than actual data.

The 2008 Key Stage 2 English reading pre-test

Used as a case study to illustrate how various reliability measures can be estimated and interpreted, Hutchison and Benton (2009) investigated the reliability of the 2008 Key Stage 2 English reading pre-test, which was conducted in 2007 (see also later discussions). The test was made up of 34 items allowing a total of 50 marks to be achieved. For the sample of pupils from 60 schools involved in their analysis, the test had a mean of 28.5 and a standard deviation of 9.1. Table 5 shows Cronbach’s alpha and IRT-based classification accuracy and consistency for the test. The reliability measures for this test are generally lower than those for the science tests discussed previously. This is expected as this test was shorter and contained more open-ended questions requiring human marking than the science tests. An IRT-based misclassification was estimated to be 17 per cent, or about 83 per cent of the pupils were classified correctly.

Table 5 Internal consistency reliability and IRT-based classification accuracy and consistency of the 2008 Key Stage 2 English reading pre-test (based on Hutchison and Benton, 2009)

Number of pupils	Cronbach’s alpha	IRT-accuracy (%)	IRT-consistency (%)
1387	0.88	83	76

Hutchison and Benton (2009) also compared the results for the pupils from the pre-test with the results from an anchor test, the live test and teacher assessment (TA). Teacher assessment levels were collected as part of their assessment development trials. Table 6 shows some additional reliability indices for the pre-test. The correlations between the pre-test scores and the anchor test scores and between the pre-test and the live test scores were higher than the correlation between the pre-test and the teacher assigned levels. In terms of classification consistency, the values are again lower than those for the science tests.

Table 6 Correlations and consistencies between the 2008 Key Stage 2 reading pre-test and the other assessments (based on Hutchison and Benton, 2009)

External measures of reliability	Comparison with scores on an anchor test	Comparison with scores on the 2007 live Key Stage 2 reading test	Comparison with teacher assessment levels
Number of pupils	637	1387	1387
Score correlation	0.846	0.812	0.766
% of pupils with improved level on alternative form	11.6	22.6	12.5
% of pupils with reduced level on alternative form	17.7	7.4	21.3
Consistency (%)	70.6	70.0	66.1

Reliability theories and models

A number of research projects were also commissioned to review measurement theories and models that are used to study reliability and the techniques that are used to produce and interpret reliability indices.

The report produced by Hutchison and Benton (2009) gives an insightful explanation of the measurement process, and a clear description of the different forms of reliability and the commonly used reliability indices under both CTT and IRT. The report provides a relatively comprehensive list of procedures that are commonly used to estimate these indices. This report also presents a clear description of how measurement error is related to reliability and how it should be interpreted. Clear descriptions of the assumptions involved in the use of the different forms of reliability measures and the sources of unreliability they account for are also provided in the report. A case study using a Key Stage 2 English reading pre-test was conducted to demonstrate how the various reliability indices can be estimated and interpreted (see previous discussions). The researchers also explored the use of alternative terms of reliability that could be understood by non-technical audiences.

The report produced by Johnson and Johnson (2009) provides an insightful explanation of the essential distinction between classical test theory and generalisability theory (G-theory): a single undifferentiated error component versus the possibility of identifying multiple error sources in assessment results. The authors looked at the procedures involved in using CTT and G-theory to investigate score reliability. Their work clearly illustrated the usefulness of using G-theory in the early developmental stages of tests and examinations. They explained how measurement models can be used in a decision study (D-study) to design a test with pre-specified measurement precision. G-theory can be used to explore the effect of various factors such as the number of tasks and the number of markers on the reliability of the test being designed, and to ensure that the acceptable degree of score reliability is reached before the test is used in live testing situations. G-theory studies can also be used to monitor the results from live testing, to ensure that the required level of score reliability is maintained during testing.

The report produced by He (2009) investigates how the reliability of composite scores is affected by the reliabilities of component scores, weights assigned to individual components and the interrelationships between component scores. He conducted a relatively comprehensive review of the literature on methodologies for researching the reliabilities of tests and examinations, particularly in terms of multivariate techniques applicable for multi-component examinations, which is of great relevance to the examinations featured in the UK. The author looked at ways of forming composite scores from component scores and summarised the procedures developed for CTT, G-theory and IRT that are widely used for studying the reliabilities of composite scores composed of weighted component scores.

Representing and reporting of assessment results and measurement errors

International approaches to representing and reporting of assessment results and errors

An important area that the Ofqual Reliability Programme is trying to explore is how assessment results and associated errors are reported internationally, and what procedures are employed by assessment providers to communicate results and errors to the users.

The report produced by Bradshaw and Wheater (2010) provides evidence in these areas. The authors searched relevant literature and examples of assessments to identify how results are represented, what level of detail is reported and what steps are taken to quantify and report on error internationally. They also looked at the rationales that were behind the use of different reporting systems. These researchers developed a taxonomy for classifying approaches to the reporting of assessment results, and used this taxonomy to classify a range of international assessments. Key findings from this study include (see Bradshaw and Wheater, 2010):

- the way results are reported depends on the intended use of the results and to whom the results are to be reported
- two opposing issues must be weighed up when deciding on the level of detail of results reporting. These are the:
    - increased reliability when few grades are reported
    - greater information when many are reported.
- a selection of international assessments have been classified using the developed taxonomy. The classification is by three main areas:
    - a description of the assessment, which includes at what stage of secondary education the assessment is used, the purpose, who makes the award, the mode and method of the assessment, and whether the assessors are external or internal
    - how the results are represented, for instance by grades, scores or a profile and the numbers of these
    - whether error or uncertainty is reported.
- few examples were located of reporting uncertainty or error in their results to learners. An introduction of the reporting of error in high-stakes qualifications

would need careful handling to ensure this did not result in misinterpretation and a loss of confidence in the system.

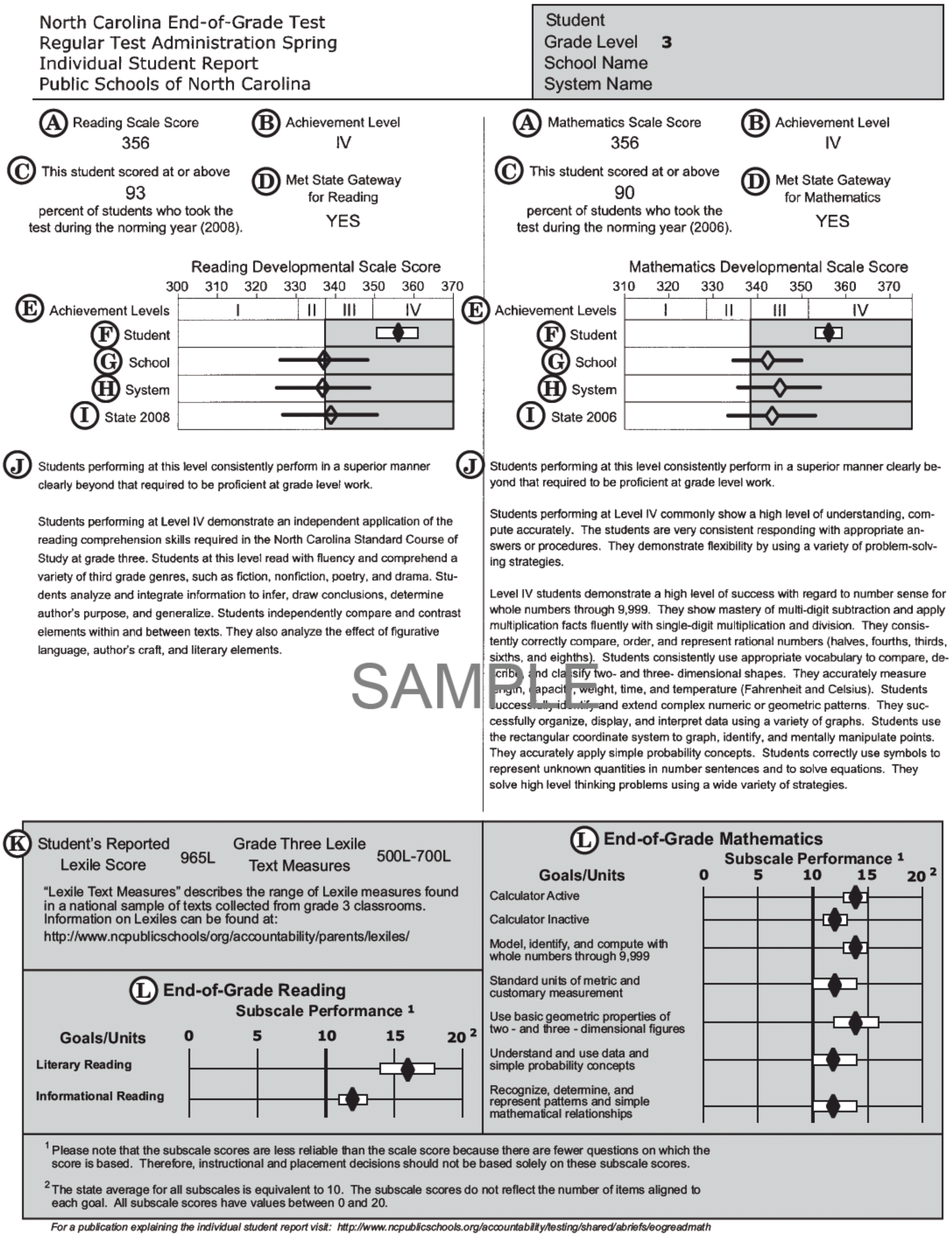
Reporting of measurement uncertainty and reliability in USA

Following the report by Bradshaw and Wheater (2010), a group of researchers led by Richard Phelps in the United States were commissioned to investigate how measurement uncertainties were represented and reported in US high-stakes tests (see Phelps et al., 2010). These researchers conducted web-based searches, which were followed up where needed with telephone calls, and contacted key researchers at relevant entities involved in reporting test results in the US. Based on the evidence they collected, these researchers discussed the prevalence of the reporting of measurement uncertainty in high-stakes tests and the degree of ease or difficulty with which ordinary citizens may access such information. They found that the degree of transparency with measurement uncertainty issues varies. Transparency seems to be greater for educational than for licensure tests, for mostly objective than for mostly essay tests, for larger programmes than for smaller programmes. These researchers also found that transparency seemed to improve if the role of test contractors was greater and the role of state government was smaller.

With educational tests, they found that many of the states in the US highlight imprecision along with the student scores on the parent/student reports (more states now are reporting score bands. See figure 1 for an example of the kind of reports commonly used). But all states prepare technical manuals, which are usually readily available to those who want them. With licensure examinations, the situation is mixed. Some provide information about uncertainty on the candidate report itself and more reliability information in a yearly technical document. Others make available various technical reports and papers summarising reliability information. Others produce reports with substantial detail that are not released to the public.

The researchers found that totality of uncertainty is not reported to all stakeholders in US educational and licensure testing programmes. It would be difficult for the average parent to find a full range of measurement uncertainty statistics for their children’s tests, for example. The researchers conclude that the average parent would not be looking for this degree of technical information, which explains why technical manuals are not found on the home page of testing programme websites. Documents that better respond to the typical consumer’s needs are placed at the forefront and the technical manuals are placed behind. Despite this, they are not hidden and there seems not to be any effort to hide information.

Figure 1 North Carolina End-of-Grade Test student score report (adapted from Phelps et al., 2010)



North Carolina End-of-Grade Test
Regular Test Administration Spring
Individual Student Report
Public Schools of North Carolina

Student
Grade Level 3
School Name
System Name

A Reading Scale Score
356

B Achievement Level
IV

C This student scored at or above
93
percent of students who took the test during the norming year (2008).

D Met State Gateway for Reading
YES

A Mathematics Scale Score
356

B Achievement Level
IV

C This student scored at or above
90
percent of students who took the test during the norming year (2006).

D Met State Gateway for Mathematics
YES

J Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work.

Students performing at Level IV demonstrate an independent application of the reading comprehension skills required in the North Carolina Standard Course of Study at grade three. Students at this level read with fluency and comprehend a variety of third grade genres, such as fiction, nonfiction, poetry, and drama. Students analyze and integrate information to infer, draw conclusions, determine author's purpose, and generalize. Students independently compare and contrast elements within and between texts. They also analyze the effect of figurative language, author's craft, and literary elements.

J Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work.

Students performing at Level IV commonly show a high level of understanding, compute accurately. The students are very consistent responding with appropriate answers or procedures. They demonstrate flexibility by using a variety of problem-solving strategies.

Level IV students demonstrate a high level of success with regard to number sense for whole numbers through 9,999. They show mastery of multi-digit subtraction and apply multiplication facts fluently with single-digit multiplication and division. They consistently correctly compare, order, and represent rational numbers (halves, fourths, thirds, sixths, and eighths). Students consistently use appropriate vocabulary to compare, describe, and classify two- and three- dimensional shapes. They accurately measure length, capacity, weight, time, and temperature (Fahrenheit and Celsius). Students successfully identify, and extend complex numeric or geometric patterns. They successfully organize, display, and interpret data using a variety of graphs. Students use the rectangular coordinate system to graph, identify, and mentally manipulate points. They accurately apply simple probability concepts. Students correctly use symbols to represent unknown quantities in number sentences and to solve equations. They solve high level thinking problems using a wide variety of strategies.

SAMPLE

K Student's Reported Lexile Score 965L
Grade Three Lexile Text Measures 500L-700L

"Lexile Text Measures" describes the range of Lexile measures found in a national sample of texts collected from grade 3 classrooms. Information on Lexiles can be found at:
http://www.ncpublicschools/org/accountability/parents/lexiles/

L End-of-Grade Reading
Subscale Performance [1]

L End-of-Grade Mathematics
Subscale Performance [1]

Goals/Units
0
5
10
15
20 [2]
Calculator Active
Calculator Inactive
Model, identify, and compute with whole numbers through 9,999
Standard units of metric and customary measurement
Use basic geometric properties of two - and three - dimensional figures
Understand and use data and simple probability concepts
Recognize, determine, and represent patterns and simple mathematical relationships

[1] Please note that the subscale scores are less reliable than the scale score because there are fewer questions on which the score is based. Therefore, instructional and placement decisions should not be based solely on these subscale scores.

[2] The state average for all subscales is equivalent to 10. The subscale scores do not reflect the number of items aligned to each goal. All subscale scores have values between 0 and 20.

For a publication explaining the individual student report visit: http://www.ncpublicschools.org/accountability/testing/shared/abriefs/eogreadmath

Public perceptions of reliability

Surveys of perceptions of A levels and GCSEs

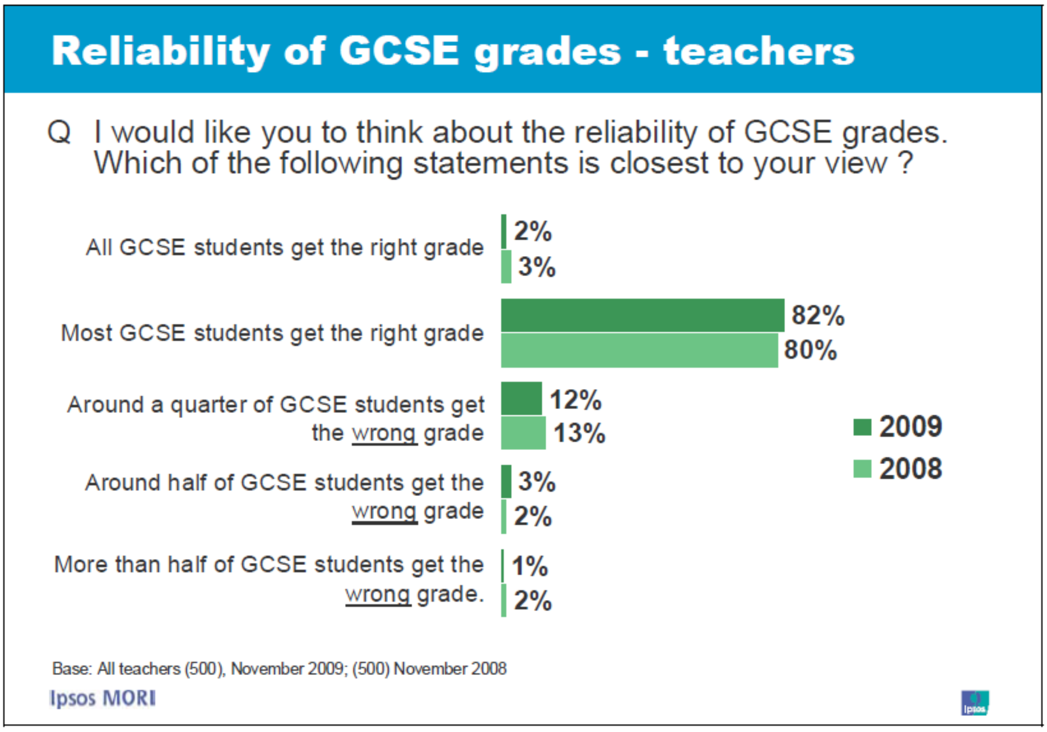
Ipsos MORI conducts a survey of perceptions of A level and GCSE that is now in its eighth wave (Ipsos MORI, 2010). The most recently reported wave of the survey was conducted in the winter of 2009, and reported findings based on samples of:

- A level and GCSE teachers
- A level and GCSE students, and their parents
- the general public.

Both the 2009 and the 2010 surveys included questions about the reliability and accuracy of examination results (Ipsos MORI, 2010). Large majorities of teachers, parents and students thought that most or all students got the correct grade at GCSEs and A levels (for example, see figure 2). Respondents also gave reasons that they perceived as being likely to cause candidates to get the wrong grade in examinations, which included:

- students performing better or worse than expected in examinations or coursework
- inaccurate marking and poorly designed exam papers.

Figure 2 Teachers' views on the accuracy of GCSE grades (adapted from Ipsos MORI, 2010)



Qualitative investigations of perceptions of reliability

As part of Strand 3 of the Reliability Programme, research projects were commissioned from Ipsos MORI and from the Assessment and Qualifications Alliance (AQA) to investigate public perceptions of reliability using workshops and focus groups. These studies focused on the following aspects of reliability:

- the assessment process
- factors affecting the performances of students on examinations
- the reliability concepts and measurement error
- the different types of error in exam results: preventable human mistakes versus inevitable random measurement error
- factors contributing to measurement error in exam results
- the level of acceptance towards human error and measurement error in exam results.

The research conducted by Ipsos MORI in January 2009 used two workshops in London and Birmingham to investigate the opinions of different groups about reliability and unreliability (Ipsos MORI, 2009). Research participants were drawn from the following groups:

- teachers
- students
- parents
- members of the general public
- employers
- examiners.

The sessions started with an analogy to an error occurring in medical treatment; this was used as a substantial input to help workshop participants understand the concepts under discussion. Researchers understood that giving such substantial input to participants whose opinions and attitudes one was trying to discover ran the risk of biasing them. However, the belief was that participants would probably not have developed views on reliability in test scores and so it was felt important to give them contextualisation of this sort.

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 examinations system, but they could be intolerant of ‘preventable errors’ (Ipsos MORI, 2009). Sometimes participants appeared to be making a distinction between inherent and preventable error, and other times not. Some research participants stated that their attitude to error depended upon whether the error changed a student’s grade or mark (Ipsos MORI, 2009). They considered grade-related error to be more consequential than mark-related. 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 typographical errors’ in papers (Ipsos MORI, 2009), while examiners could be more sanguine – taking the view that what was important was that any mistakes that did occur were rectified. There was evidence that students were aware that some inconsistency between human markers was inherent in subjects such as English. However, there were also statements that such inherent error should be minimised or even eliminated. There was considerable discussion on ‘test-related error’ (Ipsos MORI, 2009). Students and the general public were able to debate whether and how examinations can and should sample from curricula. This was a sophisticated debate about the validity of qualifications systems.

Chamberlain from AQA (2010) conducted qualitative research to follow up Ipsos MORI's (2009) work. Chamberlain collected data via 10 focus groups, with samples of:

- job-seekers
- employees
- employers
- Postgraduate Certificate in Education (PGCE) students
- teachers.

Like Ipsos MORI, Chamberlain designed her research with the assumption that she would have to take steps to mitigate participants' lack of knowledge of key elements of the reliability concept. Chamberlain used vignettes as a technique to introduce reliability to her research participants. The vignettes were:

> very short stories or scenarios involving fictional characters in specific dilemmas which were related to the research context and relevant to the lives and educational experiences of the participants. (Chamberlain, 2010)

Chamberlain's prior assumption that respondents would have limited awareness of reliability was confirmed by the data. All the participants except secondary school teachers lacked awareness of reliability concepts. The secondary teachers had more developed views, often based on experience of dealing with re-marks or appeals. The participants tended to be fairly trusting of the examinations process; trusting in the professionalism and training of subject experts. Once again, secondary school teachers' views differed from other groups; some respondents had acted as moderators in order to mediate the influence of external examiners. Participants felt it would be useful for reliability information to be communicated to the public in general terms, but were opposed to specification quantification of unreliability (for example, via an indication of the amount of uncertainty associated with a grade) on a candidate's exam certificate.

Communicating about reliability with the public

Boyle et al. (2009) conducted research looking at issues with communicating unreliability in test scores to the public. These researchers suggested two reasons for the difficulty in communicating the reliability concepts with the public.

1. The concept of reliability is complex and hard to explain succinctly.

2. Unreliability seems like an intrinsically bad news story.

They cited two sources of evidence for these reasons. Firstly, literature describing the media environment that surrounds examination results in England is summarised, which gives a history of assessment organisations' attempts at communicating with the public and is used to make suggestions for how such bodies might communicate better. The second source of evidence is the findings from the 2009 Ipsos MORI work (2009) discussed above, which provides the researchers an initial feel for the tolerance that different sectors of the public have for different sources of measurement inaccuracy in examination results. The researchers then conclude by suggesting ways to improve each of the issues with unreliability as a media story. The problem of complexity is addressed by allowing people to interact with the message via multiple media, using varied analogies and so on. In terms of the negativity of the story, the suggested response is not to try to make this into a good news story. Rather, the aspiration is to communicate the message that many assessment results contain an element of unreliability to the public in a manner that allows people to become more sophisticated users of those results.

Ofqual reliability seminars

Interpretation and communication of reliability evidence

One of the objectives of the Reliability Programme is to investigate how reliability evidence should be best produced, interpreted, evaluated and communicated to the users of assessment results. In addition to the various commissioned research projects that looked at various aspects of assessment reliability, Ofqual also held a seminar involving leading assessment experts and communications experts to discuss these issues (see Ofqual, 2009).

The discussions at the seminar focused on the following major topics:

- factors that affect the reliability of results from assessments
- definition and meaning of different forms of reliability
- theories and models that are used to study reliability
- statistical methods that are used to produce reliability estimates
- discussion on the empirical evidence of reliability from case studies
- representing and reporting assessment results and reliability estimates/ measurement errors
- improving reliability and implications
- disseminating reliability statistics
- educating the public to understand reliability concepts.

There was suggestion at the seminar that factors that could affect the reliability of assessment results and the way they interact with each other should be investigated. There was debate about the meaning of the term ‘reliability’ as to whether factors that the awarding organisations have little control should be included, and the views on this were divided. There was also debate about the different statistical methods that are used to produce reliability estimates and the impact such estimates would have on level or grade misclassification for National Curriculum assessments and general qualifications. Results from both simulation investigations and empirical studies were presented at the seminar and the differences in results between the various methods were discussed. There was discussion on how the different reliability indices could be affected by the use of different score scales used for reporting assessment results.

There was strong agreement on the importance of being more open with the public about the factors that can affect the accuracy of assessment results. 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? Delegates agreed that these were all aspects that should be discussed, whether or not they are included in any stricter definition of 'reliability'.

The participants realised the importance of a high level of reliability in assessment results. However, there was a balance that must be reached between improving reliability and the impacts on students in terms of what is to be measured, and assessment providers in terms of financial costs. Also, it was agreed that increasing reliability should not comprise validity.

It was realised that there was a need to educate the users of assessment results to understand the concept of reliability and the existence of inevitable measurement uncertainty in results from assessment systems.

Reliability policy and its implications for awarding organisations

Ofqual held a further seminar to discuss findings from some of the commissioned research projects, the implications of the findings for the development of regulatory policy on reliability, and the impact of such policy on assessment providers. Participants of the seminar included assessment researchers from academic and research institutions, awarding organisations and test agencies, the Qualifications and Curriculum Development Agency and Ofqual. The seminar involved presentations from researchers followed by group and plenary discussions.

The presentations covered a range of areas related to assessment reliability, including:

- the identification of factors that influence the reliability of results
- the review of measurement theories and models that are used to study the reliability of assessment results
- the review of techniques that are used to produce and interpret reliability measures and their limitations
- the investigation of methods that are used to study the reliability of results for different forms of assessment
- international approaches to representing and communicating assessment results, and associated errors to users of results.

Ofqual presented potential reliability policy alternatives and discussed the advantages and disadvantages of the different options (see later discussions).

The group and plenary discussions focused on the following topics:

- tension in managing public confidence while exploring and improving reliability
- operational issues for awarding organisations in producing reliability information
- particular challenges posed by the Reliability Programme in vocational qualifications.

Areas discussed at the seminar included:

- which reliability measures should be reported and how they should be published:
    - ways to represent results
    - ways to represent measurement errors
    - ways to communicate reliability measures to the public.
- constraints on reporting reliability measures:
    - human resources: the requirement of the necessary technical expertise
    - financial resources: the requirement of necessary financial costs. This is especially important for small assessment providers
    - operational difficulties: these would include the collection of the necessary data for producing reliability measures. Qualifications sharing components or units face particular challenges for producing qualification level reliabilities, as data for shared components/units are difficult to collect (for example, qualifications supported by the Qualifications and Credit Framework (QCF) may contain shared units). It is also difficult to conduct reliability studies for some components or units (for example, teacher assessments and competence-based assessments in vocational qualifications). Some qualifications have small candidate entries and could be difficult and expensive to produce reliability measures.
- issues with improving reliability:
    - reliability only represents one aspect of the quality of an assessment

  - financial implications
  - implications for technical expertise
  - validity issues: improving reliability should not comprise the validity of the assessment.

- reliability and qualification structure: component reliabilities and the overall qualification level reliability to a certain extent are affected by the structure of the qualification (for example, item types and testing time or length, and number of components/units in a qualification). Awarding organisations, QCDA, Ofqual and other regulators need to work together when designing new assessment specifications
- education of the public to understand the concept of reliability:
  - the reason why understanding measurement precision is important
  - how reliability measures should be interpreted.

One representative from an awarding body gave a presentation at the seminar on Ofqual's policy on reliability – a view from an awarding body's perspective. The presentation and the discussions that followed covered a range of aspects related to the reliability of assessment results, including:

- what examinations leading to qualifications are trying to measure
- sources of error under the framework of classical test theory
- what counts as, and should be reported as, reliability
- what practical and affordable research can be done to better understand the relative importance of the sources of error in a general sense
- routine reporting and related issues:
  - purposes of reporting reliability information
  - what is practical and affordable routinely
  - unintended consequences
  - reporting strategy: start general at system level and move towards specific qualifications as understanding grows.

Views on reporting reliability information from an international perspective

Ofqual held a joint discussion group with NFER at the 2009 AEA-Europe annual conference to gather views on representing and reporting reliability information from an international perspective. The discussions focused on the following topics.

- what do users of outcomes want?
- what are the main issues in reporting and using results and associated errors?
- is it important to report measurement error in results?
- what is the best practice in representing and reporting results?
- what is the best practice in representing and reporting measurement error?

Views expressed by participants included:

- reliability studies should be built into the assessment quality assurance process
- information on reliability (or misclassification or measurement error) should be in the public domain
- the introduction of information about reliability (particularly misclassification or measurement error) should be managed carefully to ensure that the public have confidence in the assessment system
- education of the public to understand the concept of reliability or measurement error is seen to play an important part to alleviate the problem of misinterpretations of measurement error by the media
- the reporting of results and measurement errors can be complex since results are normally used by multiple users, each of whom may have different requirements
- reliability indices should be reported at population level
- standard error of measurement should be reported at individual test-taker level.

Projects in progress

Several projects are currently being undertaken by awarding organisations, individual researchers, Technical Advisory Group members and Ofqual researchers to:

- produce further evidence of reliability of GCSE and GCE components and qualifications, and vocational qualifications
- conduct a review of reliability studies on teacher assessments
- conduct quantitative investigation of public perceptions on reliability
- analyse findings from the Reliability Programme and provide advice on Ofqual reliability policy.

New evidence of reliability

Classification accuracy and consistency in GCSEs and A level examinations offered by AQA

This project is being undertaken by researchers from AQA to look at the reliability of GCSE and A level units in the form of grade misclassification. The scope of the research is limited to a selection of units composed of objective, short answer or structured response test items that were considered to allow the assumption of reliable marking. The researchers will use two models to derive the reliability estimates: an IRT model and the CTT model, employing the procedure developed by Livingston and Lewis (1995). A comparison of the reliability estimates from the various units will be conducted to investigate the various factors that affect reliability estimates.

Estimates of reliability of qualifications

This project is being conducted by researchers from Cambridge Assessment. The aim of the project is to generate reliability estimates for a selection of GCSE and A level qualifications. This research intends to address the following questions.

- What are the most effective measures of assessment score/grade reliability that can be readily calculated?
- What ways are there of combining and presenting reliability information about assessment scores and grades?

The research will investigate the influence of the following sources of error on reliability of results:

- test-related sources of error

- marking-related sources of error
- grade-related sources of error.

The research will produce reliability estimates at both unit and qualification level.

Estimates of reliability of vocational qualifications

This project is being conducted by researchers from City & Guilds and Cito in the Netherlands. The aim of the project is to investigate the reliability of competence-based qualifications. The researchers are collecting and analysing naturally occurring and experimental data to study whether the procedures used to make binary decisions maximise the consistency of decisions. The reliabilities of the individual assessments of the qualification as well as the overall qualification level reliability will be investigated for a selection of subjects. Appropriate methods used for analysing the impacts of sources of error and assessment methods on reliability will be identified.

Quantifying and interpreting component reliability

This project is being conducted by researchers from Assessment Europe. The aims of the project are to:

- carry out generalisability analyses of selected 2009 GCSE and A level datasets
- offer capacity-building support for generalisability analysis to examining board personnel, partly through interpretive feedback and partly through collaboration in analysis
- produce a G-theory exemplification report featuring selected analyses and commentated results, for submission to Ofqual for general awarding body circulation and/or publication.

The researchers will collaborate with awarding organisations in the UK to analyse their data and provide support for capacity building for examining board personnel. The researchers intend to look at three types of dataset for potential analyses:

- objective test results
- marker standardisation data
- live marking data.

It is expected that the relative contributions from the different error sources to the overall measurement precision for the selected components will be assessed.

Review of reliability studies of teacher assessments

This project is being conducted by Assessment Europe. The principal aim of this project is the production of a comprehensive and critical review of the literature on the reliability (and more general validity) of teacher assessments, with a particular focus on relevant activity on the part of awarding organisations in England, Wales and Northern Ireland. The research will address the following main research questions:

- What is the nature of the tasks assigned to teachers as the basis for forming judgements about learners' knowledge, skills or abilities?
- What rules and procedures are in operation that guide or standardise the conditions under which learners produce the evidence that their teachers use to assess them?
- What is the nature of learners' work – reports or artefacts – that teachers are required to assess, and what rules or requirements govern these?
- What is the nature of any formal marking schemes that teachers use to arrive at their assessments, and what procedures are in place for checking reliability?
- What methods are employed to check on the reliability of sets of submissions, and what are the criteria that would trigger action to address discrepancies?
- What scaling or other adjustment methods are employed to the assessments before aggregation with assessment results to arrive at final awards, and what are the potential effects of these on the overall reliability of those final awards?

The review will embrace the role and practice of teacher assessment within secondary-sector academic and vocational examinations in the UK.

Quantitative investigation of public perceptions of reliability

This project is being conducted by researchers from Ofqual. The qualitative investigations of stakeholders' perspectives into reliability discussed previously had elements that sought to 'teach' participants about reliability – the Ipsos Mori (2009) research used a workshop format with a substantial initial input and the Chamberlain (2010) research used vignettes as part of a focus group approach. This might have helped the participants to understand the concept of reliability and the factors that could introduce uncertainty in exam scores, and develop views on measurement error. The group discussions could also have influenced the opinions of the participants about error in exam results. Furthermore, the small sample size of these studies makes it inappropriate to make any generalisation of the findings. The Ipsos MORI (2010) survey only addressed some narrow aspects of reliability of examination results. This research seeks to contribute further to a developing understanding of attitudes to reliability and unreliability using an objective online questionnaire survey, and explores the awareness of the public in the following areas:

- Knowledge of and experience in the examination process and confidence in the national examinations system.
- Understanding of the factors that affect the performances of students on examinations and factors that introduce uncertainty into exam scores.
- Attitudes towards different types of assessment error (including human mistakes and measurement inaccuracy).
- Approaches to trust in general.

Data collected will also be used to investigate:

- how attitudes to unreliability is related to knowledge and understanding of the reliability concept
- how attitudes to unreliability is related to confidence and belief in the exam system and approaches to trust
- how confidence and belief in the exam system is related to trust.

Technical Advisory Group report

The Technical Advisory Group of the Ofqual Reliability Programme will produce a report on the programme that will cover the following areas:

- the remit of the Ofqual Reliability Programme
- a summary of the results from the Reliability Programme and implications for the development of Ofqual policy on reliability
- areas for further study
- advice on possible Ofqual reliability regulatory policy.

Potential Ofqual reliability policy

Findings from the Reliability Programme will be analysed and their implications for Ofqual will be evaluated. Ofqual will develop policy on reliability based on these findings. Although detailed reliability policy will be developed at a later stage once all the evidence gathered has been evaluated, it is expected that during that development, Ofqual will explore a wide range of areas, including:

- promoting the use of reliability studies as part of the assessment quality assurance process
- promoting the use of standardised procedures for marking assessments
- promoting the use of standardised procedures for producing reliability measures (including underlying assumptions and limitations, interpretations)
- promoting the use of standardised procedures for reporting exam results and associated errors (including interpretations)
- setting reliability standards and monitoring the reliability of assessments and qualifications
- promoting public understanding of reliability concepts
- promoting the use of procedures to improve assessment reliability.

References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing.* AERA, Washington, DC.

Black, P. and Wiliam, D. (2005) ‘Lessons from Around the World: How Policies, Politics and Cultures Constrain and Afford Assessment Practices’. *The Curriculum Journal*, 16(2), 249–261.

Boyle, A., Opposs, D. and Kinsella, A. (2009) ‘No News is Good News? Talking to the Public about the Reliability of Assessment’. Paper presented at the 35th International Association for Educational Assessment (IAEA) Annual Conference in Brisbane, Australia, 13–18 September, 2009. Available online at: http://www.ofqual.gov.uk/files/2009-09-iaea-no-news-is-good-news.pdf

Bradshaw, J. and Wheater, R. (2010) *International Survey of Results Reporting*. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/Ofqual_10_4705_International_Survey_of_Results_Reporting_08_03_10_(2).pdf

Chamberlain, S. (2010) *Public Perceptions of Reliability*. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/Ofqual_10_4708_public_perceptions_reliability_report_08_03_10.pdf

He, Q. (2009) Estimating the Reliability of Composite Scores. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/2010-02-01-composite-reliability.pdf

Hutchison, D. and Benton, T. (2009) *Parallel Universes and Parallel Measures: Estimating the Reliability of Test Results.* Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/2010-02-01-parallel-universes-and-parallel-measures.pdf

Johnson, S. and Johnson, R. (2009) *Conceptualising and Interpreting Reliability.* Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/2010-02-05-conceptualising-and-interpreting-reliability.pdf

Ipsos MORI (2009) *Public Perceptions of Reliability in Examinations*. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/2009-05-14_public_perceptions_of_reliability.pdf

Ipsos MORI (2010) *Perceptions of A levels and GCSEs – Wave 8*. Available online at: www.ofqual.gov.uk/files/2009-02-26-ofqual-perceptions-of-alevels-gcses.pdf

Livingston, S. A., and Lewis, C. (1995) 'Estimating the Consistency and Accuracy of Classifications Based on Test Scores'. *Journal of Educational Measurement*, 32, 179–197.

Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2009) *Partial Estimates of Reliability*. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/files/2009-11-partial-estimates-of-reliability-report.pdf

Newton, P.E. (2005a) 'The Public Understanding of Measurement Error'. *British Education Research Journal*, 31, 419–442.

Newton, P.E. (2005b) 'Threats to Professional Understanding of Assessment Error'. *Journal of Education Policy*, 20, 457–483.

Newton, P.E. (2009) 'The Reliability of Results from National Curriculum Testing in England'. *Educational Research,* 51, 181–212.

Ofqual (2009) *The Reliability Programme – Technical Seminar Report*. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/research-and-statistics/research-reports/92-articles/20-reliability

Phelps, R., Zenisky, A., Hambleton, R. and Sireci, S. (2010) *On the Reporting of Measurement Uncertainty and Reliability for U.S. Educational and Licensure Tests*. Ofqual, Coventry, UK. Available online at: http://www.ofqual.gov.uk/research-and-statistics/research-reports

Wiliam, D. (2001) 'Reliability, Validity, and all that Jazz'. *Education,* 3–13, 29 (3), 17–21.

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 2010

© Crown copyright 2010

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

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

www.ofqual.gov.uk