Classification Accuracy in Results from Key Stage 2 National Curriculum Tests

A report from the Ofqual Reliability Programme seminar on ‘Classification accuracy in Key Stage 2 National Curriculum tests’ held at the Institute of Education in London, 8th November 2010

Qingping He, Office of Qualifications and Examinations Regulation
Malcolm Hayes, Edexcel
Dylan Wiliam, Institute of Education

March 2011
Ofqual/11/4830

Contents

Summary

As part of its Reliability Programme, the Office of Qualifications and Examinations Regulation (Ofqual) held a seminar on the reliability of results from recent Key Stage 2 National Curriculum tests (NCTs) at the Institute of Education in London on 8th November, 2010. The seminar, which was organised by Professor Dylan Wiliam and attended by assessment researchers from academic and research institutions, awarding organisations, test development agencies and Ofqual, was intended to gain a further understanding of the reliability of Key Stage 2 National Curriculum tests through discussions on results from analyses of the 2009 and 2010 live test data using a number of widely used methods. These results showed that the classification accuracy indices for the Key Stage 2 National Curriculum tests range from about 90% for mathematics, 87% for science, to about 85% for English for the past two years. These figures are substantially higher than those suggested for the tests in the early years, indicating that the reliability of Key Stage 2 National Curriculum tests has been improved considerably since their introduction in the late 1990s.

Introduction

Assuming a range of reliability values and using simulations, Wiliam (2001) demonstrated that up to 30% of students taking the Key Stage 2 National Curriculum tests in English, mathematics and science at the age of 11 could be awarded a National Curriculum performance level that did not match their actual level of achievement as a result of unreliability in the test scores. This figure, which is referred to as level misclassification, has been frequently cited in the media since that time and has prompted subsequent debate about the reliability of test and public examination results in England among policy-makers, educational researchers, test developers and practitioners. More recent research however suggested that the current Key Stage 2 NCTs produce considerably lower rates of misclassification in results (Hutchison and Benton, 2009; Maughan et al., 2009; Newton, 2009). The seminar held on 8th November 2010 as part of the Ofqual Reliability Programme, at the Institute of Education in London and attended by a group of assessment researchers from academic and research institutions, awarding organisations, test development agencies and Ofqual (see Appendix A for names of the participants), was intended specifically to gain a further understanding of the reliability of current Key Stage 2 National Curriculum tests based on analyses of live test data collected from the 2009 and 2010 test series. The seminar discussed a range of topics, including:

- Different conceptions of assessment validity and threats to validity, including construct-irrelevant variance and construct under-representation.

- The relationship between reliability and validity: reliability should be viewed as one aspect of validity.

- The implications of inferences to be made from assessment outcomes for the operational definition of reliability and the choice of reliability indices to assist data interpretation. For example, where assessment results are reported using scores, the use of standard error of measurement would be appropriate; where results are reported using performance categories such as the National Curriculum levels used for National Curriculum tests, and grades used for GCSEs and GCEs, the use of classification accuracy would be appropriate. Classification accuracy refers to the degree that both true scores and observed scores on a test classify test-takers into the same performance categories.

- Conceptualising and interpreting reliability in the context of National Curriculum tests in England.

- Factors affecting classification accuracy in test results in general.

- Impact of different methods used for estimation on classification accuracy measures.

- Current classification accuracy or misclassification (which refers to the degree that true scores and observed scores classify test-takers into different performance categories) in results from the Key Stage 2 NCTs in science, mathematics and English.

- Ways to improve reliability and assessment quality in general.

This report provides a brief summary of the seminar and focuses on classification accuracy measures for Key Stage 2 National Curriculum tests in science, mathematics and English over the past two years.

Approaches to classification accuracy estimation

The Key Stage 2 National Curriculum tests are designed to assess students working at levels 3 to 5 of the National Curriculum in three subjects: science, mathematics and English. Those students who do not achieve the level 3 threshold in the tests but have scores within 3 marks of the threshold are assigned a compensatory level 2. Since students are classified into different National Curriculum performance levels, classification accuracy would be an appropriate indicator of the reliability for these tests.

Procedures for estimating classification accuracy

Methods under both the Classical Test Theory (CTT) or True Score Theory (TST) and Item Response Theory (IRT, which is sometimes referred to as the strong true score theory) frameworks have been developed to estimate classification accuracy and classification consistency indices for single-administered tests (see Hanson, 1991; Livingston and Lewis, 1995; Rudner, 2001, 2005; Lee, 2010). Classification consistency refers to the level of agreement between classifications based on observed scores from replications of the same measurement procedure. In general, estimation of classification accuracy involves:

- modelling (conditional) error score distribution if required
- estimating true score distribution based on actual observed score distribution
- estimating observed score distribution based on modelled true score distribution (and modelled error score distribution if required) if needed
- comparing modelled true score distribution with (actual) observed score distribution, taking into account cut scores set for the performance categories, to derive classification accuracy, which generally involves calculating the proportion of test-takers that are classified into the same performance categories by both the true scores and the (actual) observed scores.

The procedures outlined above can also be applied to the IRT framework on the IRT measurement scale.

Modelling error score distribution

Depending on the model employed, the estimation of classification accuracy may require explicitly information about error score distribution around the true scores at each true score level. In CTT, the observed score of a test-taker on a test is defined as being composed of a true score representing the average of observed scores over repeated testing and an error score representing the difference between the observed score on a particular testing occasion and the true score (see Lord, 1980; Traub and Rowley, 1991). Variation of observed scores between testing occasions for a test-taker is assumed to reflect the variation of error scores. The true score can be assumed to reflect the test-taker's true ability/trait in the construct being measured and the error score can be assumed to reflect the contribution from chance factors other than his/her true ability/trait (see Harvill, 1991). The reliability of test scores for a group of test-takers is defined as the proportion of observed score variance that is true score variance. Since the reliability estimates provide information on a specific set of test scores and cannot be used directly to interpret the effect of measurement on test scores for individual test-takers (Bachman and Palmer, 1996; Bachman, 2004), the standard error of measurement (SEM), which is defined as the average of the standard deviations of the error scores for individual test-takers, is introduced for this purpose (Harvill, 1991; Wiliam, 2001; Bachman, 2004). The SEM can be used to calculate confidence intervals for observed scores or true scores (Harvill, 1991). The SEM (SEM_{CTT}) in CTT can be estimated from the reliability estimate (r_{CTT}) and the standard deviation of observed scores (σ_{CTT}):

$$SEM_{CTT} = \sqrt{1 - r_{CTT}}\sigma_{CTT} \tag{1}$$

The use of a constant SEM to characterise the error score distributions for individual test-takers suggests that the error score distributions are the same for individual test-takers at different score points (equal error score variance). This is one of the limitations associated with CTT. There are however procedures which produce varying SEMs at different true score points (which in this case is termed conditional standard errors of measurement – CSEMs). Error scores may be assumed to be normally distributed at each true score level.

Other major limitations of CTT include that item and test statistics such as item difficulty and discrimination power and test reliability are dependent on the examinee sample from which they are derived (see Lord, 1980; Hambleton and Swaminathan, 1983; Hambleton et al, 1991; Bachman, 2004; Bond and Fox, 2007). As indicated previously, CTT also assumes equal variance of measurement errors for all test-takers. IRT, which models the performance of test-takers on individual test items and takes into account factors such as person ability and the characteristics of test items that affect the performance of the test-takers, such as item difficulty and discrimination power, overcomes some of these limitations in situations where test data fits the model. When test data meet IRT model assumptions (such as the unidimensionality and local independence assumptions required for unidimensional IRT models), some models such as the Rasch model may be used to construct objective measures (sample free item calibration and item-free person ability estimation; see Rasch, 1960; Wright and Stone, 1979). IRT models have been widely used to study error of measurement for both items and test-takers (Lord, 1980; Hambleton *et al*, 1991). In the case of ability measures for examinees, the standard error (standard deviation) (SEM_{IRT}) of a person ability measure (θ) is inversely proportional to the square root of the test information ($I(\theta)$):

$$SEM(\theta)_{IRT} = \sqrt{\frac{1}{I(\theta)}} \tag{2}$$

The measurement error in IRT is a function of the ability measure and can therefore be different at different locations on the measurement scale. Unlike CTT, in which an error score component is explicitly specified, IRT procedures make use of the probabilistic nature of the test data to estimate ability measures and associated error of estimation. It has been proposed that the standard error of measurement derived for ability measures in IRT could be used to estimate the CSEM for test scores in CTT (RMT, 2007):

$$CSEM(X)_{CTT} \approx \frac{1}{SEM(\theta)_{IRT}} \tag{3}$$

where $CSEM(X)_{CTT}$ is the conditional standard error of measurement at true (or observed) score X which corresponds to the ability measure θ. The conditional error measures may be assumed to be normally distributed.

Estimating true score distribution

In the case of CTT, true score distribution can be modelled based on actual observed score distribution and the assumed error score distribution or simply based on the observed score distribution (eg Lord, 1969; Hanson, 1991; Livingston and Lewis, 1995).

In the case of IRT, both the IRT measurement scale and the true score scale can be used. The true ability measure distribution is generally assumed to be the ability measures obtained from analysing the test data using a specified IRT model within the IRT software. In IRT, the true score is defined as the expected score from the IRT model which is the score that a test-taker is most likely to obtain given his/her ability measure and the parameter values of the items in the test. The true score distribution can therefore be estimated from the distribution of ability measures.

Estimating observed score distribution

Once the error score distribution at each true score level and the distribution of true scores are known, observed score distributions can be generated. In the case that the conditional observed score distribution at each true score level can be modelled directly from the true score, the observed score distribution can be modelled without explicit information on error score probability distribution. There are generally two ways to produce observed score distributions:

- Using simulations: Given the true score and the (conditional) error score probability distribution, the observed score of an examinee is simulated independent of other examinees, which is used to derive the simulated observed score distribution for the sample or population. Since the simulation is based on the probability distribution of error scores around true scores, the simulated observed score distribution will be slightly different for each simulation run.

- Using numerical integration: Given the total number of examinees at each true score level and the associated conditional error score probability distribution, these examinees are apportioned to an expected distribution of observed scores which can then be integrated over the true score range to derive the overall observed score distribution for the whole sample or population.

In the case of IRT, the observed ability measure distribution can similarly be estimated. If true score is used, the observed score distribution can also be derived based on the ability measure and the parameter values of the items in the test.

Estimating classification accuracy

Classification accuracy is estimated by comparing modelled true score distribution with observed score distribution:

- In the case of using simulations to estimate observed score distribution: Classification accuracy is calculated as the proportion of examinees with simulated observed scores in the same performance category as their true scores. The average of the classification accuracy over a large number of simulation runs can be used as the classification accuracy of the test.

- In the case of using numerical integration: The expected distribution of observed scores (conditional observed score distribution) at a given true score level can be used to derive the proportion of the examinees at that true score level with the observed scores in the same performance category as the true score (this proportion of examinees at a specific true score level is termed the conditional classification accuracy). The overall classification accuracy can be calculated by integrating the conditional classification accuracy over the entire true score range.

The actual observed score distribution may also be compared with the estimated true score distribution to derive classification indices.

Factors affecting classification accuracy

From the discussion outlined above, it can be seen that classification accuracy is affected by a range of factors, including:

- The measurement precision (SEM) or test reliability (which is generally used to estimate SEM for CTT). Other things being equal, higher measurement precision will result in higher classification accuracy or lower rate of misclassification.

- Score range and distribution.

- Number of performance categories that are used and the boundary locations of the performance categories. Other things being equal, higher number of performance categories will result in lower classification accuracy (note that in earlier years, the number of performance categories was larger, so that classification accuracy would be lower).

- Models that are used and the methods that are used to estimate model parameters. Different models and model parameters could produce different true score distributions or observed score distributions, which will affect the classification accuracy.

The different methods used

Six different methods have been used in the present study to estimate classification accuracy. These are briefly explained below.

Method 1 (M1):

- Assuming a beta distribution for true scores for each of the three subjects (model parameters are derived by fitting the modelled distribution to that of the observed scores).
- Assuming a constant SEM at different true score points and a normal error score distribution.
- Using simulations to derive observed score distributions and classification accuracy.

Method 2 (M2):

- Given the actual observed score distribution and assuming a normal probability distribution of true scores around an observed score with the standard deviation of the true scores assumed to be the same as the SEM. (The SEM should strictly be interpreted as the standard deviation of observed scores around a true score, and the standard deviation of the true scores around an observed score would be slightly smaller than the SEM; see Harvill, 1991).
- Generating the conditional true score distribution at each observed score level and estimating the corresponding conditional classification accuracy.
- Using numerical integration to derive the overall true score distribution and classification accuracy.

Method 3 (M3):

- Given the actual observed score distribution and assuming a normal probability distribution of true scores around an observed score, the standard deviation of true scores around an observed score is however a function of the observed score (the CSEM), which is assumed to be the inverse of the IRT derived SEM (see Equation 3).
- Generating the conditional true score distribution at each observed score level and estimating the corresponding conditional classification accuracy.
- Using numerical integration to estimate the overall true score distribution and classification accuracy.

Method 4 (M4) – The Livingston and Lewis method (Livingston and Lewis, 1995):

This method involves transforming the observed raw score distribution on a test to a distribution from a test (the transformed test) containing independent, identical and equally difficult dichotomous items:

- Estimating the effective test length for the transformed test. This is based on that the original test and the transformed test produce the same reliability measure. Scores on the transformed test form a new scale.
- Transforming the original raw score on to the new scale.
- Estimating the distribution of proportional true scores (calculated using the transformed observed scores) using a four-parameter beta model. The beta distribution parameters are estimated using the transformed observed score distribution.
- Estimating the observed score distribution at each true score level based on the the assumption that at each proportional true score level the observed score distribution is a binomial distribution. The conditional classification accuracy at each true score level is then estimated.
- Using numerical integration to estimate the overall observed score distribution and classification accuracy.

Method 5 (M5):

This is similar to Method 3, but the calculation is conducted on the IRT measurement scale:

- Analysing the data using a specified IRT model in the IRT software to estimate item and person parameters.
- The cut scores are converted on to the IRT ability scale.
- Given the actual observed person ability distribution and assuming a normal probability distribution of true ability measures around an observed ability measure. The standard deviation of the true ability measures around an observed ability measure is assumed to be the model-derived SEM exported from the IRT software.
- Generating the conditional true score distribution at each observed score level and estimating the corresponding conditional classification accuracy.
- Using numerical integration to estimate the overall true ability distribution and classification accuracy.

Method 6 – the Lee approach (see Lee, 2008, 2010):

This is also an IRT-based approach, which involves:

- Analysing the data using IRT software to estimate item and person parameters for the selected IRT model.
- Using the actual observed person ability distribution and the known item parameter values for the items in the test to estimate the true score distribution (IRT model expected score distribution).
- For a given true score, the probability that any observed scores (which can be estimated from the IRT model) will fall into the same performance category as the true score (the conditional classification accuracy) is calculated from the IRT model.
- Using numerical integration to estimate the overall observed score distribution and classification accuracy based on the distribution of conditional classification accuracy and the estimated true score distribution.

Classification accuracy in the 2009 and 2010 Key Stage 2 National Curriculum tests

The datasets

The data analysed are from the 2009 and 2010 live test series. These include mark distributions for the populations and item level data for each subject for a sample of over 3000 students for each series.

The Key Stage 2 science test consists of two test papers (Test A and Test B), each made up of 40 marks with a testing time of 45 minutes. The papers consist of a mixture of objective, short answer, and longer response questions. Scores from the two papers are combined to produce a composite score for the subject. The Key Stage 2 mathematics test has three subtests: Test A, Test B and a mental test. Test A and Test B are each worth 40 marks. Calculators are allowed for Test A but not for Test B. The mental test is worth 20 marks. Scores from the three subtests are combined to form the composite score for mathematics. The English test has two components, a reading component and a writing component. Both the reading and writing components are worth 50 marks. Again scores on the two subtests are aggregated to produce the overall score for the subject.

The composite scores for each subject are used to assign National Curriculum levels representing the achievement in the subject to the pupils following a rigorous standards setting process, which involves the use of both statistical information and professional judgement of the quality of sampled scripts. Test equating is also carried out to ensure the continuity of standards over time. Table 1 lists the final level boundary marks (cut scores) for the 2009 and 2010 test series, and Figures 1 to 3 show the mark distributions for both the populations and the samples that were used for item level data. As can be seen, the score distributions are negatively skewed.

Table 1: Level boundary marks (cut scores) for the 2009 and 2010 Key Stage 2 tests

Level	**Boundary marks**					
	Science		**Mathematics**		**English**	
	2009	**2010**	**2009**	**2010**	**2009**	**2010**
2	18	17	15	15	20	20
3	21	20	18	18	23	23
4	40	40	46	46	44	43
5	63	63	77	79	67	68

Figure 1: Score distributions of the Key Stage 2 2009 and 2010 science tests

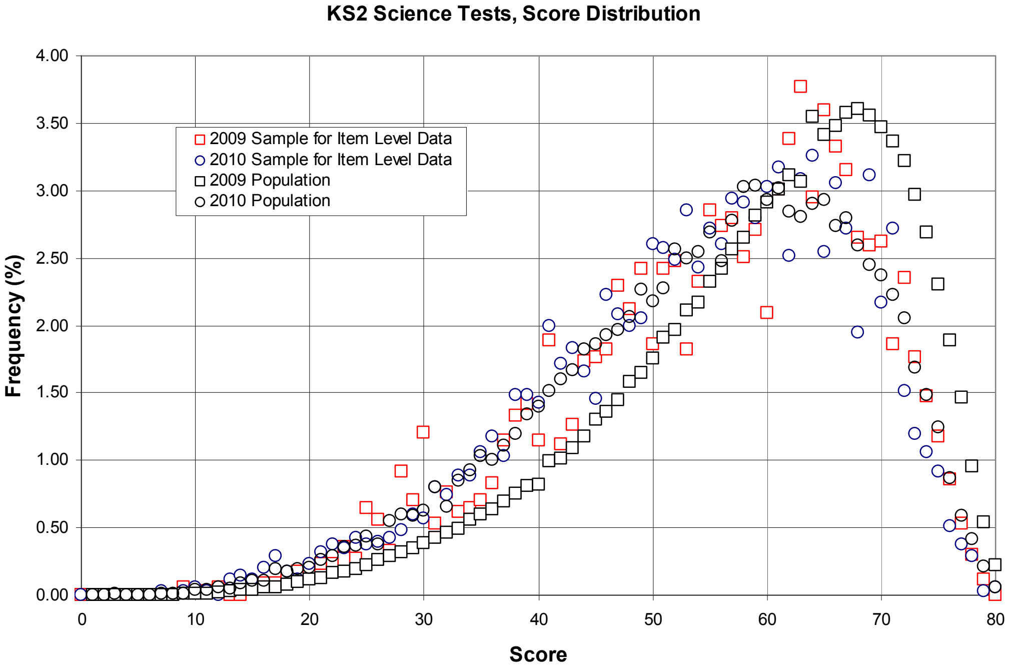


Figure 2: Score distributions of the Key Stage 2 2009 and 2010 mathematics tests

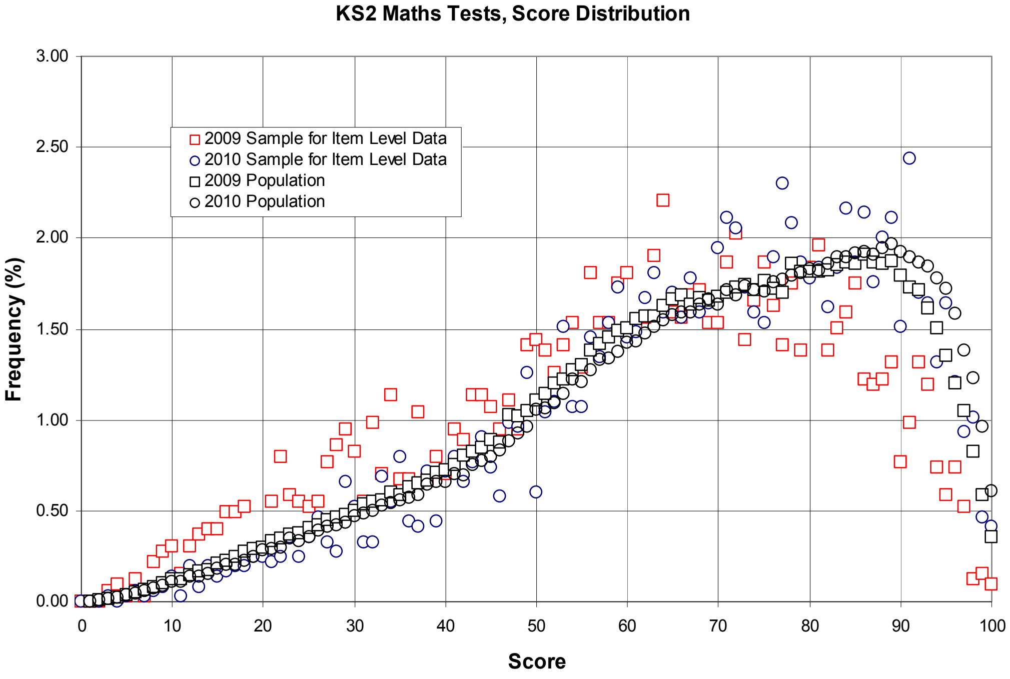
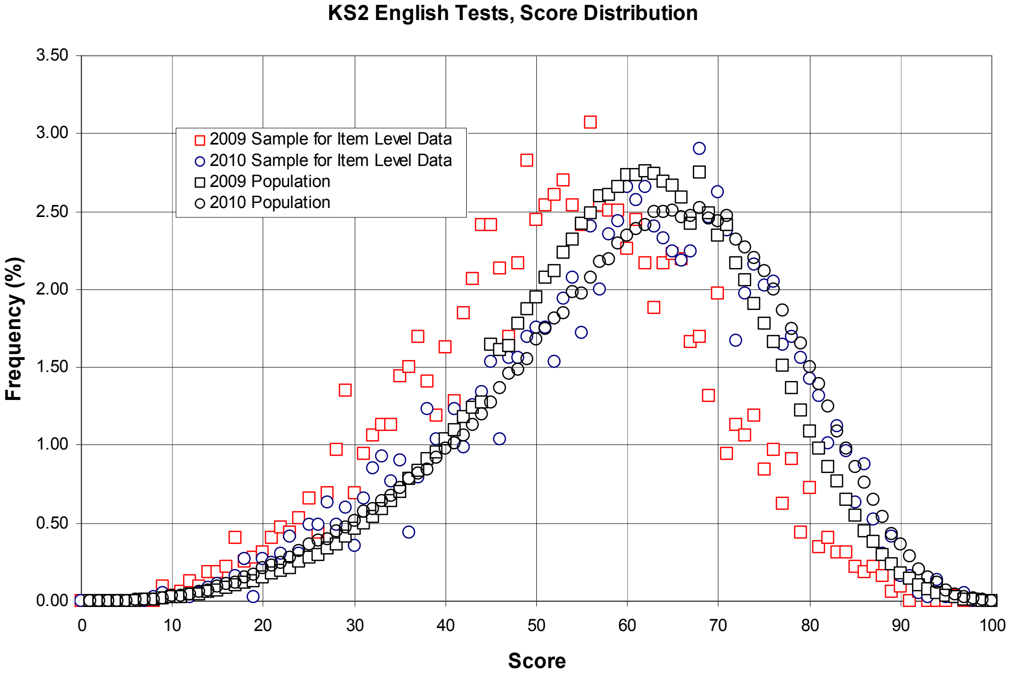


Figure 3: Score distributions of the Key Stage 2 2009 and 2010 English tests



Data analyses

The item level data from subtests in a subject were combined and analysed for Cronbach's alpha. Cronbach's alpha is a measure of consistency in test scores. Specifically, Cronbach's alpha refers to the degree to which groups of items in a test produce consistent or similar scores for individual test-takers (or consistency in test scores from different sets of items). As items in a test can be viewed as a sample from a domain of potential items, Cronbach's alpha may therefore be viewed as a measure of the extent to which the scores from test-takers on a test represent the expected scores from the entire domain. If items in a test also require human marking, Cronbach's alpha may also to some degree reflect the variability in test scores associated with the inconsistency in marking between markers. Table 2 shows the values of Cronbach's alpha for the three subjects, with highest values for the mathematics tests and the lowest for the English tests. For individual subjects, the values are similar for 2009 and 2010. For the science tests, values of Cronbach's alpha are also similar to those estimated for the pre-tests for 2005–2009 (see Maughan et al., 2009). These values are relatively high for tests of this kind, and significantly higher than those reported in earlier years.

Table 2: Sample sizes and Cronbach's alpha for the 2009 and 2010 Key Stage 2 tests

Subject	**Sample size**		**Number of items**		**Cronbach's Alpha**	
	2009	**2010**	**2009**	**2010**	**2009**	**2010**
Science	3395	26017	79	73	0.928	0.926
Mathematics	3265	3649	100	100	0.968	0.964
English	3189	3656	40	38	0.910	0.919

The IRT software WINSTEPS (http://www.winsteps.com/index.htm), which implements the Partial Credit Model (PCM) developed by Masters (Masters, 1982; Wright and Masters, 1982), was used to analyse the item level data for item and person measures. An inspection of model fit statistics indicated that the PCM model fits the data reasonably well.

Cronbach's alpha was assumed to represent a good approximation of the test reliability and was used to estimate the SEM. The model-derived SEMs exported from WINSTEPS were assumed to be close to the real SEMs. The methods described previously were then used to estimate classification accuracy for the tests. For the Livingston and Lewis method and the Lee method, the software systems BB-CLASS and IRT-CLASS developed by the Center for Advanced Studies in Measurement and Assessment (CASMA) at the University of Iowa were used (http://www.education.uiowa.edu/casma/computer_programs.htm, see Brennan, 2004; Lee and Kolen, 2008). IRT-CLASS implements a range of IRT models, including the PCM model.

Classification accuracy

Table 3 shows the range of classification accuracy values for the samples for 2009 and 2010 that were used to produce item level data, suggesting that the different methods produce slightly different estimates. This is expected, because different methods make different assumptions about the true scores and error scores and the extent to which these assumptions are met by the test data varies between the different methods. The accuracy values are generally about 90% for the mathematics tests, 87% for the science tests, and 85% for the English tests. These values are also comparable with those from recent studies by other researchers (Hutchison and Benton, 2009; Maughan et al., 2009; Newton, 2009). These values are substantially higher that those suggested for the tests in the early years of testing. This increase in classification accuracy is likely to be largely due to the increased reliability of Key Stage 2 National Curriculum tests and changes in the structure of the tests (see also Maughan et al., 2009).

Table 3: Classification accuracy for samples from the 2009 and 2010 Key Stage 2 tests estimated using different methods

Subject	**Accuracy (%)**					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
Science 2009	90	87	88	86	87	87
Science 2010	89	86	87	86	86	86
Mathematics 2009	92	89	90	89	89	89
Mathematics 2010	91	91	91	90	91	90
English 2009	87	84	87	83	86	85
English 2010	88	85	86	85	85	90

Assuming that the values of Cronbach's alpha estimated for the samples can be generalised to the populations, two of the methods (M2 and M4) were also used to estimate the classification accuracy for the populations (see Table 4). These values are closely similar to those estimated for the samples.

Table 4: Classification accuracy for the populations for the 2009 and 2010 Key Stage 2 tests estimated using two different methods.

Subject	**Accuracy (%)**	
	M2	**M4**
Science 2009	88	87
Science 2010	87	86
Mathematics 2009	90	90
Mathematics 2010	91	90
English 2009	87	85
English 2010	85	85

It is also interesting to look at classification accuracy at individual level (conditional level classification accuracy). Table 5 shows the accuracy at each level for the 2009 science test. It is clear that pupils at level 5 are most accurately classified, while pupils at level 2 are least accurately classified. It is however noticed that, since level 2 is only a compensatory level and covers a range of 3 marks (see Table 1), the low conditional classification accuracy is expected. Information expressed in Table 5 could be used to improve the overall classification accuracy.

Table 5: Conditional level classification accuracy for the 2009 Key Stage 2 science test estimated using different methods

Level	M2 (sample)	M5 (sample)	M2 for the population
N	72	75	72
L2	31	33	31
L3	83	82	83
L4	87	87	86
L5	89	90	91
Overall	87	87	88

Concluding remarks

Since pupils taking the National Curriculum tests are classified into different National Curriculum performance levels, classification accuracy would be an appropriate indicator of the reliability of the tests. Classification accuracy is affected by a range of factors, including measurement precision (test reliability), number of performance categories and mark distribution. The classification accuracy measures for the 2009 and 2010 Key Stage 2 National Curriculum tests estimated using a selection of methods are about 87% for the science tests, 90% for the mathematics tests, and 85% for the English tests. These values are considerably higher than the classification accuracy values suggested for the tests in their early years, reflecting improvement in quality of the tests (including reliability) as a result of the changes that have been made to improve the system. These values are also comparable with the accuracy values obtained by other researchers in recent years (see Hutchison and Benton, 2009; Maughan et al., 2009; Newton, 2009).

As implied earlier, all classification accuracy indices are estimates, based on certain mathematical models that inevitably make various assumptions about test scores. In many situations, the degree to which the model assumptions are met by the test data is difficult to evaluate. Although it is likely that the extent to which the real test data meet the assumptions of the models varies between the different methods, the classification accuracy values estimated from the different methods for the tests studied here are broadly similar, which suggests that the models represent the test data reasonably well.

The classification accuracy estimates are slightly different for the three subjects for the past two years, with mathematics having the highest accuracy. These differences to a certain degree reflect the difference in the nature of tasks assessed by the different subjects and the reliability in marking the test papers. While for the mathematics and science tests, answers can be reasonably objectively marked, marking of the English tests, particularly the writing component, is subject to potentially substantial human subjective judgement. Therefore, inconsistency in marking between markers would be expected to be higher for the English tests than for the mathematics and science tests, although procedures such as the development of a clear mark scheme and proper marker training have been adopted to improve marking reliability. It is also noticed that for all the three subjects, the standard error of measurement was estimated using Cronbach's alpha. For the mathematics and science tests, Cronbach's alpha may be assumed to be a good approximation of the test reliability, but the degree to which it also captures marking unreliability for the English tests is not entirely clear. It is also worth noting that Cronbach's alpha does not capture all aspects of the sources of variation in scores beyond the ability of the candidate. Further work on the effect of marking unreliability on Cronbach's alpha would be required.

It is important to realise that although efforts should be made to improve assessment reliability, some degree of unreliability in test scores or inaccuracy in classifications is inevitable in any educational assessments, including the Key Stage 2 National Curriculum tests. This is because variability that exists in the various factors in the assessment process affecting test scores cannot be completely eliminated. For example, the Key Stage 2 National Curriculum tests only sample contents and skills from across the whole of the Key Stage 2 National Curriculum programmes of study, different areas will be covered in different years, which would inevitably result in differences between the tests. Assessments use tasks of different formats to assess different types of knowledge and skills so that valid inference can be made from assessment results. Some tasks can be marked more consistently than others. It is certainly important to continue to explore ways of improving test reliability but this must be done with regard to other important factors such as validity and manageability.

It is also worth noting that both the reliability coefficient and the classification accuracy index are estimates of population parameters, and that they should be interpreted accordingly. The probability that a particular examinee is misclassified clearly depends on the position of his/her test score on the score scale. Examinees on or near the level boundary marks are more likely to be misclassified than those further away. It is also realised that the technical meaning of the term 'misclassification' in the context of educational measurement is different from that of its daily use. While the former is a measure of measurement error (not operational errors) and merely implies that variation in test scores can exist when the measurement procedure is repeated, the latter may suggest that something has gone wrong operationally.

References

Bachman, L. (2004) *Statistical Analyses for Language Assessment*. Cambridge, Cambridge University Press.

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice*. Oxford, Oxford University Press.

Bond, T. and Fox, C. (2007) *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd edn). Mahwah, NJ, USA, Lawrence Erlbaum.

Brennan, R. (2004) *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0) (CASMA Research Report No. 9) [Computer software and manual]. Iowa City, IA, Center for Advanced Studies in Measurement and Assessment, University of Iowa. Available online at: www.education.uiowa.edu/casma.

Hambleton, R. and Swaminathan, H. (1983) *Item Response Theory: Principles and applications.* The Netherlands, Kluwer-Nijoff.

Hambleton, R., Swaminathan, H., and Rogers, J. (1991) *Fundamentals of Item Response Theory.* Newbury Park, California, USA, Sage.

Hanson, B. (1991) *Method of Moments Estimates for the Four-Parameter Beta Compound Binomial Model and the Calculation of Classification Consistency Indexes*. ACT Research Report, 91–5. Iowa City, IA, ACT, Inc.

Harvill, L. (1991) 'Standard error of measurement'. National Council on Educational Measurement, ITEMS. Available online at: www.ncme.org/pubs/items/16.pdf.

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

Lee, W. (2008) *Classification Consistency and Accuracy for Complex Assessments using Item Response Theory* (CASMA Research Report No. 27). Iowa City, IA, Center for Advanced Studies in Measurement and Assessment, University of Iowa. Available online at: www.education.uiowa.edu/casma.

Lee, W. (2010) 'Classification consistency and accuracy for complex assessments using item response theory'. *Journal of Educational Measurement*, 47, 1–17.

Lee, W. and Kolen, M. (2008) *IRT-CLASS: A computer program for item response theory classification consistency and accuracy* (Version 2.0) [Computer software]. Iowa City, IA, Center for Advanced Studies in Measurement and Assessment, University of Iowa. Available online at: www.education.uiowa.edu/casma.

Livingston, S. and Lewis, C. (1995) ‘Estimating the consistency and accuracy of classifications based on test scores’. *Journal of Educational Measurement*, 32, 179–97.

Lord, F. (1969) ‘Estimating true-score distribution in psychological testing (and empirical Bayes estimation problem)’. *Psychometrica*’. 34, 259–99.

Lord, F. (1980) *Applications of Item Response Theory to Practical Testing Problems*. New Jersey, USA, Lawrence Erlbaum.

Masters, G (1982) ‘A Rasch model for partial credit scoring’. *Psychometrika*, 47, 149–74.

Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2009) *Partial Estimates of Reliability: Reliability in the Key Stage 2 Science Tests*. Coventry, UK, Ofqual. Available online at: www.ofqual.gov.uk/files/2009-11-partial-estimates-of-reliability-report.pdf.

Newton, P. (2009) ‘The reliability of results from National Curriculum testing in England’. *Educational Research*, 51, 181–212.

Rasch, G. (1960) *Probabilitistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark, Denmark Paedagogiske Institute.

RMT (2007) ‘Standard errors and reliabilities: Rasch and raw score’. *Rasch Measurement Transactions 2007*, 20, 4, 1086. Available online at: www.rasch.org/rmt/rmt204f.htm.

Rudner, L. (2001) ‘Informed test component weighting’. *Educational Measurement: Issues and Practice,* 20, 16–19.

Rudner, L. (2005) ‘Expected classification accuracy’. *Practical Assessment, Research and Evaluation,* 10, 13. Available online at: http://pareonline.net/pdf/v10n13.pdf.

Traub, R. and Rowley, G. (1991) ‘Understanding reliability’. National Council on Educational Measurement, ITEMS. Available online at: www.ncme.org/pubs/items/15.pdf.

Wiliam, D. (2001) ‘Reliability, validity, and all that jazz’. *Education*, 29, 17–21.

Wright, B. and Masters, G. (1982) ‘Rating scale analysis’. *Rasch Measurement*, Chicago, IL, USA, MESA Press.

Wright, B.D. and Stone, M.H. (1979) *Best Best Design: Rasch Measurement.* Chicago, IL, MESA Press, USA.

Appendix A: Seminar participants

Jo-Anne Baird	University of Bristol
Anton Bèguin	Cito
Paul Black	King’s College London
Tom Bramley	Cambridge Assessment
Barbara Donahue	Qualifications and Curriculum Development Agency (QCDA)
Malcolm Hayes	Edexcel
Qingping He	Office of Qualifications and Examinations Regulation (Ofqual)
Tina Isaacs	Institute of Education
Sarah Maughan	National Foundation for Educational Research (NFER)
Louise Maycock	Qualifications and Curriculum Development Agency (QCDA)
Paul Newton	Cambridge Assessment
Dennis Opposs	Office of Qualifications and Examinations Regulation (Ofqual)
Alastair Pollitt	CamExam
Dylan Wiliam	Institute of Education

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2011

Office of Qualifications and Examinations Regulation

Spring Place	2nd Floor
Coventry Business Park	Glendinning House
Herald Avenue	6 Murray Street
Coventry CV5 6UB	Belfast BT1 6DN

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

www.ofqual.gov.uk