

Measurement Invariance of the Key Stage 2 National Curriculum Science Sampling Test in England



February 2014

Ofqual/14/5376

Contents

Summary	2
1. Introduction.....	3
1.1 Measurement invariance and its importance	3
1.2 Approaches for investigating measurement invariance	3
1.3 The Key Stage 2 National Curriculum science sampling test in England	5
1.4 Aims of the present study	6
2. The partial credit Rasch model.....	7
3. Methods.....	9
3.1 The dataset	9
3.2 Data analysis.....	9
4. Results and discussion.....	10
4.1 Performance of different subgroups	11
4.2 Relationship between subtests and test internal consistency reliability	13
4.3 Test dimensionality and invariance of first factor loadings.....	15
4.4 Rasch modelling.....	18
5. Concluding remarks.....	35
6. References	37

Summary

Since 2010, the Key Stage 2 National Curriculum test in science that was taken by the whole national cohort at the age of 11 in England has been replaced with a sampling test. Pupils taking the sampling test come from a nationally representative sample of schools annually. The study reported in this paper compares the performance of different subgroups of the samples (classified by gender, ethnicity, English as an additional language, eligibility for free school meals, and special educational needs) in the test over the 2010 and 2011 test series. To study measurement invariance of the test across the different subgroups, factor analysis and Rasch modelling have been used to analyse item level data for differential item functioning (DIF) and differential test functioning (DTF) in order to investigate the functioning of the items in the tests and the overall tests across the subgroups. A range of measurement properties related to the science construct being measured by the tests, including dimensionality, item difficulty, expected scores on individual items and expected total test scores, have been examined. It is concluded that the Key Stage 2 science sampling tests function similarly across the different subgroups at both item level and the overall test level and provide a measure of pupils' attainment in science on the same measurement scale for all subgroups studied over the past two years.

1. Introduction

1.1 Measurement invariance and its importance

Educational and psychological tests are often used for making inferences about the performances of individuals in the underlying theoretical constructs or latent traits. They should therefore provide a valid and reliable measure of all test-takers' level on the latent construct being assessed. Since a test only represents one operationalization of the unobservable psychological construct, and its tasks or items only serve as indicators of the underlying construct, the accuracy of any inferences made based on individual test-takers' results about their level on the assessed construct will depend to a certain extent on the accuracy of the operationalization of the construct. Meaningful and fair comparison of test performances between test-takers from different subgroups within a population (or between test-takers from different populations) requires a high level of measurement invariance (MI) of the test across the different groups (or populations) (Dragow, 1984; Dragov and Kanfer, 1985; Reise, Widaman and Pugh, 1993; Carle, Millsap and Cole, 2008; Milfont and Fischer, 2010; Millsap, 2011; Zhang, Fokkema and Cuijpers et al., 2011; Dai, Han and Hu et al., 2011). Measurement invariance requires that a test measures the same construct in the same way for all subgroups of the population or that the measures for individuals from different subgroups are on the same measurement scale in terms of the underlying construct being assessed (Reise et al., 1993). In other words, measurement invariance requires that measures from a test be independent of other attributes of the test-takers (or invariant in relation to construct-irrelevant aspects of the test-takers such as gender or ethnicity), other than the attribute the test is intended to measure. Individual test-takers with similar standing on the latent trait should perform similarly on test items and the overall test, regardless of which group they come from.

Measurement invariance implies that the relationship between the properties of the measures from the test with respect to the underlying latent trait being measured by the test should be the same across the different subgroups of the population. However, a test is only an operationalization of the intended construct and is generally developed by test developers with particular knowledge of the domain of interest from which the test tasks are sampled and knowledge of the intended target population, and individuals from different subgroups of the population with the same level of the underlying trait being measured may perform differently on specific tasks in the test due to factors such as tasks requiring additional ability or trait or unfairly favouring pupils from specific subgroups. The extent to which measurement invariance holds can be investigated at both item level and the overall test level.

1.2 Approaches for investigating measurement invariance

The basic principle in conducting measurement invariance investigations is to examine the degree of similarity in the relationship between the properties of the

measures concerned and the underlying latent construct across different subgroups of the population or between different populations. Two frequently used approaches in measurement invariance studies are the confirmatory factor analysis (CFA) and item response theory (IRT) approaches (Reise et al., 1993; Meade and Lautenschlager, 2004; Dimitrov, 2006; Carle et al., 2008; Koh and Zumbo, 2008; Milfont and Fischer, 2010; Dai et al, 2011; Millsap, 2011; Zhang et al., 2011). In the CFA approach, the measured variables (observed variables or scores on test items) are expressed as linear functions of the latent variables (latent traits or variables used to represent the assessed construct) or factors with the loadings of the measured variables on the latent variables as the regression coefficients. The level of measurement invariance is investigated by comparing the similarities of the factor loadings of the measured variables (or covariance between test items) across the subgroups. Similar factor loadings across the groups would indicate a high level of measurement invariance.

In the IRT approach, the level of the latent construct being assessed for each test-taker and the characteristics of each of the items in the test are specified, and an item response model is used to describe the interactions between the test-takers and the test items for all subgroups. An IRT model generally uses a non-linear mathematical function (item response function (IRF), or item category probability curve (ICPC)) to describe the probability that an examinee will have a specific score on a particular item given his/her ability (trait level or person parameters) and the characteristics of the item (item parameters) (Lord, 1980; Hambleton, Swaminathan and Rogers, 1991; Yen and Fitzpatrick, 2006). The person and item parameters are estimated based on the responses from the test-takers on the items in the test using a variety of estimation techniques (see Embretson and Reise 2000; Linacre 2012; Simon 2008). For each item, an expected score for an examinee can be calculated using the corresponding IRFs. The relationship between the expected score on an item and the person ability is termed the item characteristic curve (ICC). The expected score on the test for a person can then be calculated as the sum of the expected scores on the items. The relationship between the expected score on the test and the person ability is termed the test characteristic curve (TCC). If the ability measures of persons from all groups are calibrated onto the same ability scale, the level of measurement invariance can be investigated by examining the degree of similarity (invariance) of the IRFs or ICCs and TCC (or the item parameter values) between the different subgroups. Similar item parameter values would indicate a high level of measurement invariance (that is, the IRFs or ICCs and TCCs are the same across different subgroups). Items with parameters significantly different between groups would indicate differential item functioning (DIF) between the groups (Raju, van der Linden and Fleer, 1995; Clauser and Mazor, 1998; Oshima and Morris, 2008), that is, test-takers from different subgroups with the same ability or trait have different probabilities of succeeding with the same item. If the aggregated effect of the functioning of the items in the test is different between the groups, the test

exhibits differential test functioning (DTF) between the groups (Raju, van der Linden and Fler, 1995; Clauser and Mazor, 1998; Flowers, Oshima, and Raju, 1998; Oshima and Morris, 2008). DIF items are a potential source of test bias.

1.3 The Key Stage 2 National Curriculum science sampling test in England

The Key Stage 2 National Curriculum tests in England include tests in three subjects: English (reading and writing), mathematics and science for pupils aged 11. The English and mathematics tests are taken by the whole national cohort and the science tests by pupils from a nationally representative 5 per cent sample of schools annually. These tests are designed to assess the performance of pupils at the end of Key Stage 2 (from ages 7 to 11) in the National Curriculum, which was introduced in the late 1980s in England (Qualifications and Curriculum Authority (QCA), 1999a; Whetton, 2009; Isaacs, 2010). The science sampling test began in 2010, involving a sample size over 25,000 pupils each year. Item level data for all pupils taking the science sampling test since 2010 are available as a result of the introduction of online marking.

The Key Stage 2 science tests are developed using standard test-development procedures. Items are created by experienced assessment experts and evaluated by review panels consisting of experts from a variety of relevant areas, including curriculum subjects, and inclusion and cultural awareness, and with different perspectives, such as teachers, local authorities and markers. Initially selected items are used to construct tests for pre-testing before they are used in live testing. The pre-testing process is well-defined and rigorous. The purposes of the pre-testing are to evaluate the quality of items further, to produce item statistics, to ensure the tests are at the appropriate difficulty level for the target population, and to produce initial performance level boundary scores by equating the test with an anchor test. Test equating is to ensure that the comparability of performance standards over time is maintained.

The live science sampling test consists of two components (Test A and Test B), each 45 minutes long and each with a maximum of 40 marks. The papers are composed of a mixture of objective and short-answer questions. Most of the items are dichotomous (i.e. items with the response scored as either correct or wrong). There are a few items that are marked out of two marks. Representative samples of schools are selected, and all pupils in Year 6 (the final year of the primary phase in England) from the selected schools sit the test on a specified day in May each year, unless they are assessed by their teachers to be operating below the level of the test. Pupils' scripts are scanned and marked by trained human markers online, with scores recorded on each item or sub-item for each pupil. Scores from the two components are aggregated to generate a composite score for the science subject. A standard-setting process (QCA, 2009b), which involves the use of both statistical

information and professional judgement of the quality of sampled pupils' work, is used to set thresholds for National Curriculum performance at levels 3, 4 and 5 for the mark distribution. Pupils are awarded a National Curriculum attainment level for the science subject based on their composite scores. Outcomes are aggregated and published nationally in order to monitor national attainment in science at the end of the primary phase.

1.4 Aims of the present study

In England we are currently conducting a research programme, the Assessment Validity Programme, investigating a number of issues concerned with the validity of national tests and public exams in England (Ofqual, 2012; Opposs and He, 2012). As part of this programme, the study reported in this paper aims to provide an evaluation of the level of measurement equivalence of the Key Stage 2 National Curriculum science sampling tests for 2010 and 2011 across different subgroups (classified by gender, ethnicity, English as an additional language, eligibility for free school meals, and special educational needs). Factor analysis and Rasch modelling have been used to analyse item level data for differential item functioning (DIF) and differential test functioning (DTF) in order to investigate the way items in the test and the overall test functioned across the subgroups.

2. The partial credit Rasch model

Item response theory and Rasch measurement models have been widely used in psychological and educational testing (Yen and Fitzpatrick, 2006; Wright and Stone, 1979; Wright and Maters, 1982). There are two types of item response theory models: unidimensional IRT (UIRT) models for items in a test measuring a single ability in common, and multidimensional IRT (MIRT) models for items measuring multiple abilities (see Embretson and Reise 2000; Reckase and Martineau 2004; Sijtsma and Junker 2006; Reckase 2009; He, 2012). In the present study, the measurement model used is the unidimensional Rasch model for dichotomous items along with its extension, the partial credit model (PCM), for polytomous items. The Rasch model for a dichotomous item can be expressed mathematically as (Rasch, 1960; Wright and Stone, 1979):

$$P(\theta, x) = \begin{cases} \frac{\exp(\theta - b)}{1 + \exp(\theta - b)} & \text{for } x = 1 \\ \frac{1}{1 + \exp(\theta - b)} & \text{for } x = 0 \end{cases} \quad (1)$$

where:

θ = person ability

b = the difficulty of the item

$P(\theta, x)$ = the probability of a person with ability θ scoring x on the item.

For the dichotomous Rasch model, when the item difficulty is close to the person ability, the test-taker will have a 50 per cent chance of answering the item correctly. The PCM for a polytomous item with a maximum available score of m can be expressed as (Masters 1982, 1999; Wright and Masters 1982):

$$P(\theta, x) = \begin{cases} \frac{\exp \sum_{k=1}^x (\theta - \delta_k)}{1 + \sum_{l=1}^m \exp [\sum_{k=1}^l (\theta - \delta_k)]} & \text{for } x = 1, 2, \dots, m \\ \frac{1}{1 + \sum_{l=1}^m \exp [\sum_{k=1}^l (\theta - \delta_k)]} & \text{for } x = 0 \end{cases} \quad (2)$$

In Equation (2), δ_k is the k^{th} threshold location of the item on the latent trait continuum, which is also referred to as the item step difficulty. The PCM reduces to the dichotomous Rasch model when $m=1$.

Two important assumptions are required for unidimensional item response theory and Rasch measurement models, including the partial credit Rasch model: unidimensionality and local independence. Unidimensionality requires that one ability or a single latent variable is being measured by the items in the test. Local independence requires that test-takers' responses to any questions in the test are statistically independent when the underlying ability influencing their performance on the whole test is held constant. That is, answers to any items in the test are independent of answers to any of the other items but only depend on the level of latent trait of the test-taker. In situations where items having the same stimulus are analysed as individual items, the assumption of local independence could be violated. Such items can be grouped to form a testlet and treated as one polytomous item and analysed using polytomous IRT and Rasch models (Thissen, Steinberg and Mooney, 1989; Yen, 1993).

An important feature of the Rasch model, and of IRT models in general, is that when test data meet the model assumptions and the model fits the data, values of model parameters used to characterise items that have been estimated using different samples drawn from the population will be similar (that is, invariance of model parameter estimates). This is the basis for conducting DIF and DTF analyses. The partial credit Rasch model has been widely used for investigating measurement invariance issues (Smith, 2004; Teresi, Ramirez, Lai et al., 2008; Schulz and Fraillon, 2009; Munkholm, Berg, Löfgren et al., 2010; Dai et al., 2011; Taylor and Lee, 2011; Pae, 2012). It will also be used in the present study to explore the measurement invariance of the Key Stage 2 National Curriculum science sampling tests discussed above.

3. Methods

3.1 The dataset

The data analysed in the present study are the pupils' item level data from the 2010 and 2011 live test series. For the 2010 test, there were 73 items (sub-items), with 66 sub-items marked out of 1 mark and 7 sub-items out of 2 marks (see Table 1). For the 2011 test, there were 10 sub-items worth a maximum of 2 marks, with the remaining 60 items being dichotomous. The sample size is similar for both years, with over 25,000 pupils taking the tests.

Table 1 Number of items and sample sizes for the 2010 and 2012 Key Stage 2 National Curriculum science sampling tests

Test	Maximum marks	No of items		No of pupils	
		2010	2011	2010	2011
Test A	40	34	35	25922	25561
Test B	40	39	35	25922	25561
Overall	80	73	70	25922	25561

3.2 Data analysis

Item level data from each year were analysed for some basic item and test statistics for the overall samples and different subgroups. The internal consistency reliabilities of the subtests and the overall test for the overall sample and individual groups were estimated using Cronbach's coefficient alpha (Cronbach, 1951). The relationship between the two subtests was also investigated through correlation analysis. SPSS was used to conduct factor analysis.

The Rasch analysis software WINSTEPS (www.winsteps.com/index.htm), which implements the partial credit Rasch model, was used to conduct Rasch analysis on the item level data (combined from the two subtests). Outputs from WINSTEPS include item and person measures, model-derived standard error of measurement for both persons and items, model fit statistics and other statistics. WINSTEPS also allows investigation of the degree to which the model assumptions are met by the test data and DIF of items between groups.

4. Results and discussion

Figure 1 shows the score distributions for the samples for the 2010 and 2011 tests. As can be seen, the score distributions are considerably negatively skewed. As with other Key Stage 2 National Curriculum tests, this might be partly due to policy decisions regarding test construction to ensure that the test targets pupils at level 4 and that the majority of pupils have a positive experience in taking the test. Further, the improvement in the performance of pupils since the introduction of the test in the late 1980s along with the effort to ensure similar performance level boundary marks in different test series and the level criteria may also have contributed to the negatively skewed score distributions. Although the shape of the score distributions for both Test A and Test B and the overall test for the two years are similar, the average scores for the 2011 tests are slightly higher than those for the 2010 tests, indicating that the 2011 tests, particularly Test A, are easier relative to the respective sample in comparison with the 2010 tests (also see Table 2). The average percentage scores from the samples on the tests are 67 per cent for the 2010 test and 69 per cent for the 2011 test respectively.

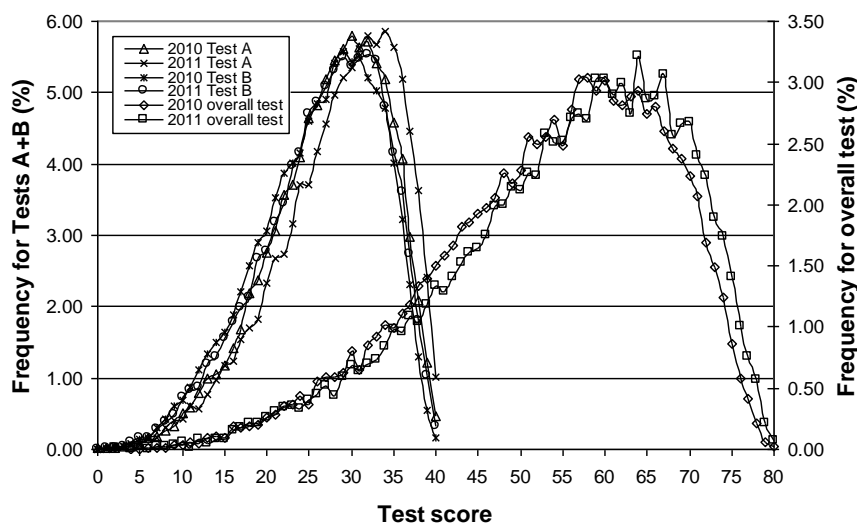


Figure 1 Score distributions of the 2010 and 2011 science subtests and the overall tests

Table 2 Mean scores and standard deviations of the 2010 and 2011 Key Stage 2 science tests

	2010		2011	
	Mean	St Dev	Mean	St Dev
Test A	27.39	6.92	28.48	7.14
Test B	26.22	7.17	26.76	7.18
Overall	53.61	13.52	55.24	13.75

4.1 Performance of different subgroups

Pupils in the samples were classified into a number of subgroups according to ethnicity, English as the first language, English as an additional language (EAL), eligibility for free school meals (FSM), special educational needs (SEN), and gender (see Table 3). Table 4 shows the number of pupils in each subgroup, their overall mean score and the standard deviation of scores for the 2010 and 2011 tests. As can be seen from Table 4, boys and girls on average performed similarly for both years. Pupils with SEN performed substantially less well than pupils without SEN. Similarly, pupils with EAL performed less well than pupils with English as their first language. Black pupils performed below the average of the whole samples (see Table 2).

Table 3 Codes used for classifying subgroups

Group	Code	Explanation
Ethnic group	A	Asian
	B	Black
	C	Chinese
	G	Any other ethnic group
	M	Mixed
	U	Unclassified
	W	White
Language group	E	English as the first language
	O	Other language as the first language (EAL)
	U	Unclassified
Free school meal (FSM)	N	Not eligible for FSM
	Y	Eligible for FSM
Special educational needs (SEN)	N	Non-SEN
	S	Having a formal statement of SEN
	Y	Assessed by the school as having SEN
Gender	F	Female
	M	Male

Table 4 Sample size, mean score and standard deviation of scores for different subgroups

Group		2010			2011		
		No of p	Mean	St Dev	No of p	Mean	St Dev
Ethnic group	A	2213	51.88	14.25	2569	52.55	15.17
	B	1129	50.00	14.19	1242	51.08	14.62
	C	91	59.57	11.92	79	61.99	12.52
	G	295	50.75	15.39	262	50.68	15.22
	M	1029	54.96	13.44	1029	55.99	13.57
	U	157	52.26	14.49	158	54.03	13.94
	W	21008	53.95	13.32	20222	55.84	13.39
Language group	E	22356	54.03	13.28	21548	55.91	13.31
	O	3530	50.97	14.67	3990	51.67	15.38
	U	36	53.67	14.05	23	48.13	17.46
Free school meals	N	21568	54.80	13.07	21022	56.59	13.29
	Y	4354	47.75	14.15	4539	48.99	14.12
Special educational needs	N	19710	56.65	11.82	19690	58.21	12.01
	S	386	39.72	15.71	416	43.95	15.63
	Y	5826	44.27	13.88	5455	45.38	14.40
Gender	F	12675	53.51	13.39	12655	55.51	13.73
	M	13247	53.71	13.64	12906	54.97	13.76

4.2 Relationship between subtests and test internal consistency reliability

Values of correlation coefficient between the two subtests based on the item level data from the overall sample are provided in Table 5. These are 0.84 for both the 2010 and 2011 tests. These values are similar to those reported for the 2005 to 2009 Key Stage 2 science pre-tests (Maughan, Styles and Lin et al., 2012). The correlation between two subtests reflects the strength of a linear relationship between the scores from the subtests and to a degree the magnitude of errors associated with the observed scores. If the two subtests are designed to be equivalent forms, then the correlation between the two subtests can also be interpreted as a measure of the equivalent forms reliability of the test. Equally, if the subtests are assumed to assess similar constructs, the correlation could be viewed as a measure of concurrent validity of the subtests.

Table 5 Correlations between subtests and Cronbach's alpha for the subtests and the overall tests

Test	Cronbach's alpha		Correlation (A and B)	
	2010	2011	2010	2011
Test A	0.85	0.87	0.84	0.84
Test B	0.87	0.86		
Overall	0.93	0.93		

The internal consistency reliabilities of the two subtests represented by Cronbach's alpha vary from 0.85 to 0.87 for the two years (see Table 5). These values are also similar to those for the 2005 to 2009 Key Stage 2 science pre-tests reported by Maughan et al. (2012). 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 in a test). As items in a test can be viewed as a sample from a domain of potential items, Cronbach's alpha may also be viewed as a measure of the extent to which the scores from test-takers on a test represent the expected scores of equivalent tests from the entire domain. At the overall test level, values of Cronbach's alpha are 0.93 for both years. Values of Cronbach's alpha for the overall test scores for the subgroups vary from 0.91 to 0.94, and are close to 0.93 for the overall samples (see Table 6). This suggests that the items in the tests functioned similarly across the subgroups. For the group with unclassified language, Cronbach's alpha was not calculated due to small sample sizes.

Although Cronbach's coefficient alpha is probably the most widely used measure of the internal consistency reliability of a test, its use as a measure of both test reliability and internal structure or unidimensionality has been widely criticised by psychometricians (for example Green, Lissitz and Mulaik, 1973; Cortina, 1993; Sijtsma, 2009a, 2009b, 2012; Revelle and Zinbarg, 2009; Yang and Green, 2011). It has been shown that coefficient alpha is not a measure of test unidimensionality but only a lower bound to test reliability (Green et al., 1973; Cortina, 1993; Sijtsma, 2009a). However, a high value of Cronbach's alpha is a necessary condition of unidimensionality (Cortina, 1993; Green et al., 1977).

Table 6 Cronbach's alpha for subgroups on the overall tests

Group		2010	2011
		Cronbach's Alpha	Cronbach's Alpha
Ethnic group	A	0.93	0.94
	B	0.93	0.93
	C	0.91	0.93
	G	0.94	0.94
	M	0.93	0.93
	U	0.93	0.93
	W	0.92	0.93
Language group	E	0.92	0.92
	O	0.94	0.94
	U		
Free school meals	N	0.92	0.92
	Y	0.93	0.93
Special education need	N	0.91	0.91
	S	0.94	0.94
	Y	0.92	0.93
Gender	F	0.93	0.93
	M	0.93	0.93
Overall		0.93	0.93

4.3 Test dimensionality and invariance of first factor loadings

The issue of dimensionality has been an important consideration for test development, evaluation and use (Cortina, 1993; Furr and Bacharach, 2008; Furr, 2011). A clear understanding of the dimension structure of a test would facilitate test score construction and interpretation. To assess the dimensionality of the 2010 and 2011 science sampling tests investigated here, exploratory factor analysis was carried out. Figure 2 shows the distributions of the eigenvalues of the factors (latent variables) for the overall samples. As can be seen, the eigenvalues decrease considerably from the second factor, suggesting that the first factor is the dominant common factor shared by all items in the tests. For individual subgroups, a similar pattern exists.

Table 7 shows the ratios of the first to the second eigenvalues and the second to the third eigenvalues for the different subgroups and the overall samples for the two tests. Apart from the Chinese group, which has a small sample size, these ratios are similar for the subgroups and the overall sample for both years, suggesting again that

the items functioned similarly across the groups. Further, the ratio of the first to the second eigenvalues is substantially higher than the ratio of the second to the third eigenvalues, indicating that there is a strong common dimension that runs through all the items in the tests. Figure 3 compares the loadings of items on the first factor estimated for individual subgroups with the loadings on the first factor for the overall sample for the 2010 and 2011 tests. Apart from the Chinese group, a strong linear relationship exists between these factor loadings, which again suggests that the test and its items functioned similarly across the different groups.

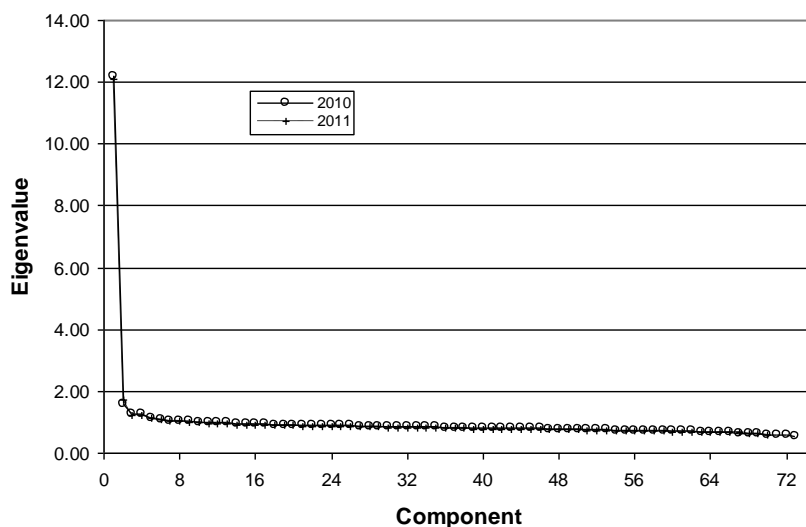


Figure 2 Scree plot of factor eigenvalues for the 2010 and 2011 Key Stage 2 science tests

Table 7 The ratios of the first to the second eigenvalues and the second to the third eigenvalues for different subgroups and the overall samples

Group		2010		2011	
		1st/2nd	2nd/3rd	1st/2nd	2nd/3rd
Ethnic group	A	8.33	1.14	8.01	1.26
	B	7.41	1.12	6.89	1.22
	C	3.22	1.32	4.25	1.04
	G	7.30	1.06	6.38	1.20
	M	6.94	1.22	7.10	1.22
	U	5.44	1.16	4.97	1.16
	W	7.71	1.22	6.84	1.23
Language group	E	7.69	1.22	6.88	1.21
	O	8.52	1.14	7.80	1.32
	U				
Free school meals	N	7.47	1.22	6.89	1.23
	Y	8.04	1.14	6.55	1.24
Special educational needs	N	6.75	1.14	6.46	1.08
	S	7.39	1.13	6.31	1.19
	Y	7.18	1.17	6.38	1.33
Gender	F	7.58	1.25	7.02	1.23
	M	7.94	1.20	6.99	1.29
Overall		7.79	1.22	7.01	1.39

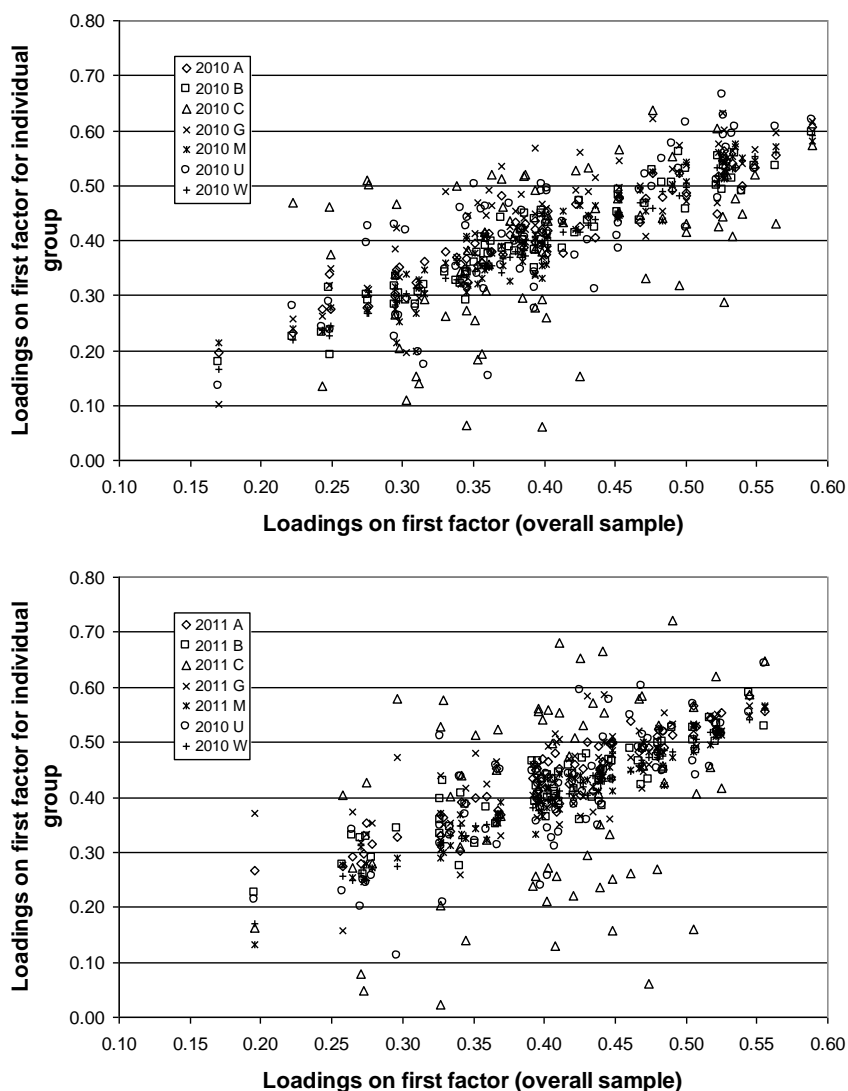


Figure 3 Relationship between loadings on the first factor for different subgroups and those for the overall sample for the 2010 and 2011 tests

4.4 Rasch modelling

In the Rasch model, because the probability of succeeding at a score category on an item is specified as a function of the difference between person ability and item difficulty, the person parameters and item parameters are therefore placed on the same measurement scale. For example, persons and items can be placed on the same unidimensional line used to represent the intended construct and then compared directly. Figure 4, the item distribution map, compares the distribution of difficulty of the items and the distribution of person ability for the 2010 and 2011 tests. In this figure, the vertical lines are the Rasch continuum representing the construct being measured by the tests. The item measures are placed to the right

side of the lines and person measures to the left of the lines (the scale is displayed on the left side of the figure). The unit used for the Rasch measurement scale is logits (log-odds unit). It is to be noted that the Rasch or IRT item response function (IRF) is invariant when certain linear transformations are applied to both person ability trait and item parameters, and this creates indeterminacy when establishing the IRT or Rasch measurement scales (that is, the origins of person trait and item parameters cannot be determined independently). One possible way to deal with this for the Rasch model is to set the average difficulties of the items in the test to zero to determine the values for item parameters, which is the case for Figure 4. In Figure 4, M stands for mean of item measures or person measures, S for one standard deviation and T for two standard deviations of the respective measures. Items near the top are most difficult items, while those near the bottom are the easiest items. Similarly, pupils near the top are high-performing pupils, while those near the bottom are the lowest-performing pupils.

Compared with Figure 1, Figure 4 shows that the distribution of pupils on the Rasch ability scale is more symmetrical than on the raw score scales. This is because although raw scores are sufficient statistics for estimating the Rasch model parameters, Rasch ability is not a linear function of raw scores – the ability of pupils with extreme raw scores will be stretched outwards. The use of the Rasch scale therefore removes the ceiling effect (and the floor effect) associated with raw scores. The difficulties of the items are also relatively symmetrically distributed.

In the Rasch model, when a person's ability matches the difficulty of the items, maximum information or minimum measurement error will be obtained. Therefore, Figure 4 suggests that the items in both the 2010 and 2011 tests, particularly the 2011 test, are relatively easy for the corresponding pupil samples, because the means of the person measures are considerably higher than the means of the item measures for both years (see also Table 2). If it can be assumed that the 2011 sample is similar to the 2010 sample in terms of the underlying pupil ability, Figure 4 and Table 2 would suggest that the 2011 test is slightly easier than the 2010 test. More difficult items would be needed to provide more accurate measurement for those high-performing pupils. However, the tests are designed primarily to provide a reliable measure of pupils' attainment in science at level 4. In the case of large gaps between item difficulties in the item distribution, pupils with ability close to the gaps will also be less precisely measured. Gaps along the construct continuum could be an indication of construct under-representation (Baghaei, 2008), which could impact on the validity of test score interpretation.

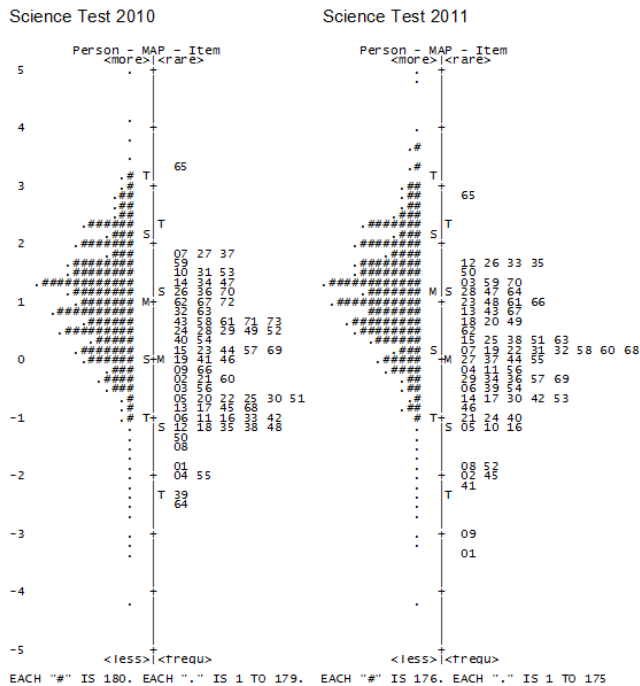


Figure 4 The Rasch item distribution map for the 2010 and 2011 Key Stage 2 science sampling tests

Figure 5 illustrates the test characteristic curves for the two tests. Because the mean of the item measures were set to 0 logits, pupils with ability close to 0 will have a mean score of about half of the maximum available marks on the tests. As can be seen, an approximate linear relationship between the latent ability and the raw score exists in the region near the middle of the ability distribution. However, in the extremes of the ability continuum, the relationship is not linear. As with Figure 4, Figure 5 also suggests that the 2011 test is relatively easier for higher-performing pupils in relation to the whole sample compared with the 2010 test. For average and low-performing pupils, the two tests performed similarly.

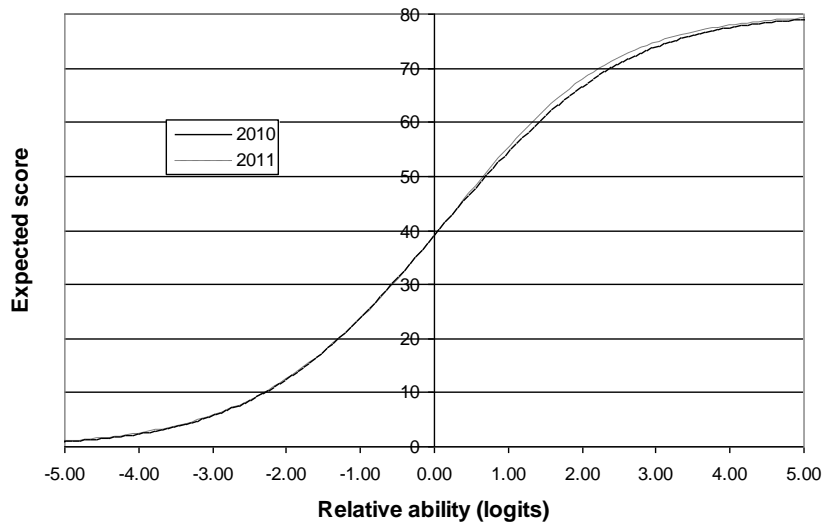


Figure 5 Test characteristic curves for the 2010 and 2011 tests

As an example of using the Rasch item distribution map to compare the performance of items between groups in a test, Figure 6 shows the item maps for the 2010 test for pupils not eligible for free school meals versus those eligible for free school meals, resulting from separate item calibration (that is, responses for the respective group were used to estimate item parameters for the group concerned) with the mean of the item difficulties set to 0 for each group. Although the distributions of item difficulty measures and person ability measures show a similar pattern, the mean of the person measures for pupils not eligible for FSM is substantially higher than that for pupils eligible for FSM. This is consistent with the average raw score of the former being higher than that of the latter listed in Table 4.

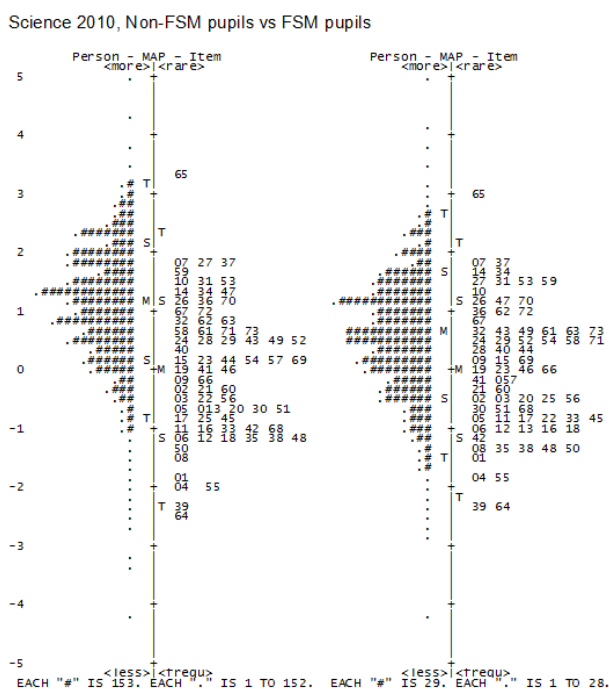


Figure 6 The Rasch item distribution map for pupils not eligible for FSM and those eligible for FSM for the 2010 test

Model fit and test dimensionality

The extent to which items in a test meet the model assumptions (local independence and unidimensionality assumptions) needs to be investigated. Violation of model assumptions can invalidate the interpretation of results. Unidimensionality and local independence can be evaluated using factor analysis and analysis of residual correlation matrices (Yen, 1993; Smith 2002; Reeve and Fayers, 2005; Reckase, 2009; also see previous discussion). However, as indicated by Hambleton et al. (1991), as long as a coherent scale can be constructed by the items, strict unidimensionality will not be needed because IRT and Rasch analysis is relatively robust to violations of the unidimensionality assumption (also see Bèguin 2000; Hanson and Bèguin 2002).

The application of IRT and Rasch models to analyse test data also assumes that the model reflects the functioning of the test items correctly. As indicated by Barnard (1999), an evaluation of how well the model chosen fits the data is essential in IRT and Rasch modelling to ensure the usability of test results. Embretson and Reise (2000) and Reckase (2009) outline procedures for assessing the fit of IRT models. The degree to which the model fits the test data can be evaluated using model fit

statistics which are normally calculated and exported by the software used for IRT and Rasch item calibration.

As illustrated by the high eigenvalues for the first factor from factor analysis and the high values of the internal consistency reliabilities of the tests discussed above, there is a dominant dimension that runs through the items in the test for both the 2010 and 2011 series. Principal components analysis (PCA) using WINSTEPS suggests that the Rasch model accounts for 31.4 per cent and 30.8 per cent of the total variance for the two tests respectively (see Table 8). Analysis of the residuals (the residual is defined as the difference between a person's observed score on an item and his/her Rasch model predicted score on the item) indicated that the first contrast in the residuals, in eigenvalue terms, accounts for 1.7 per cent of the unexplained variance, the second contrast 1.5 per cent of the unexplained variance, and the third contrast 1.4 per cent of the unexplained variance, for both the 2010 and 2011 tests. These residual contrasts are of relatively equal importance in explaining the variance unexplained by the Rasch model, and therefore no meaningful second dimension could be constructed for the original responses. These tests can be essentially treated as unidimensional (see Pae, 2012; Linacre, 2012). Further analysis of paired item residuals suggested that, for the 2010 test, there were only two pairs of items with a residual correlation greater than 0.20. The first pair had a residual correlation of 0.24 and the second pair had a residual correlation of 0.22. The maximum correlation between paired item residuals was 0.126 for the 2011 test. The mean correlations between paired item residuals were -0.012 for the 2010 test and -0.013 for the 2011 test respectively (see Table 8). It was noted that an initial analysis of the 2011 test data suggested that two sub-items had a correlation of over 0.75 between their residuals, indicating local dependence between the two items. These two items were then combined into a two-mark polytomous item and the data were re-analysed. As can be seen from Table 8, the standard deviations of the correlations are 0.021 and 0.019 for the two tests respectively, which are relatively small, suggesting that they are close to their mean values. The correlations between pairs of item residuals are small, suggesting that the final structure of the items in the tests exhibit local item independence for the two tests (Yen, 1993; Yen, Bené and Huynh, 2000).

Table 8 Variance explained by the Rasch model for the 2010 and 2011 tests, contrasts of residuals and correlations between paired item residuals

	2010	2011
Variance explained by model (%)	31.40	30.80
Unexplained variance (%)	68.60	69.20
Explained by first contrast (%)	1.70	1.70
Explained by second contrast (%)	1.50	1.50
Explained by third contrast (%)	1.40	1.40
Explained by fourth contrast (%)	1.30	1.30
Correlation between pairs of item residuals		
Average	-0.012	-0.013
Standard deviation	0.021	0.019

Frequently used Rasch model fit statistics include some of the residual based fit statistics such as unweighted fit mean squares (Outfit) and weighted fit mean squares (Infit) for items and persons (see Wright and Master, 1982; Wu and Adams, 2007; Linacre, 2012). Both Infit and Outfit have an expected value of 1. The extent to which the values of Infit or Outfit depart from 1 reflects the level of dissimilarity between the shapes of the observed item characteristic curve and the theoretical ICC. For items with Infit less than 1, their observed ICC will be steeper than the model ICC (over-fit, or more discriminative), while items with Infit greater than 1 will have observed ICC flatter than the model ICC (under-fit or less discriminative). Items and persons with Infit statistics in the range from 0.70 to 1.30 are normally regarded as fitting the Rasch model reasonably well (Keeves and Alagumalai, 1999). However, some researchers set the range of acceptable values for Infit even wider, from 0.60 to 1.40 (Tan and Yates, 2007; Wong, McGrath and King, 2011).

Figure 7 shows the item Infit and Outfit statistics for the 2010 and 2011 tests for the overall samples. Almost all items have Infit values within the range 0.80–1.20, suggesting that they fit the Rasch model well. Figure 8 depicts the distribution of person Infit statistics. Over 99 per cent of the pupils have Infit values within the range 0.70–1.4, indicating that pupils also fit the Rasch model reasonably well.

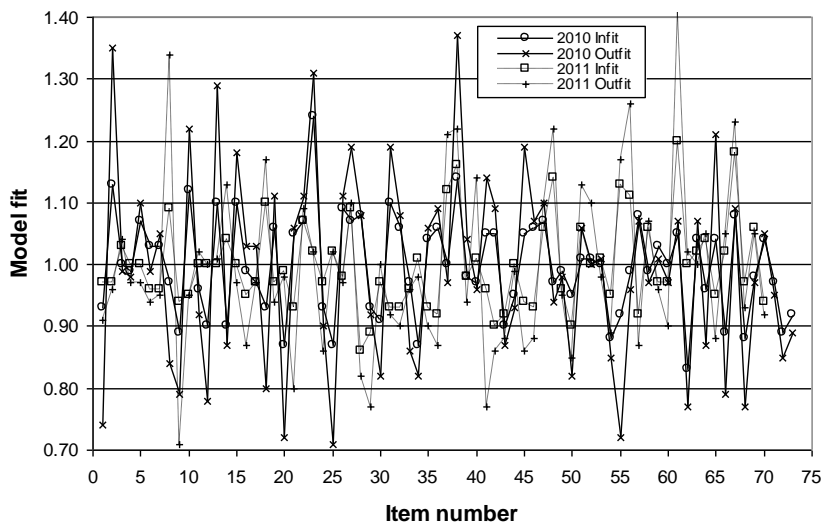


Figure 7 Item Infit and Outfit statistics for the 2010 and 2011 tests

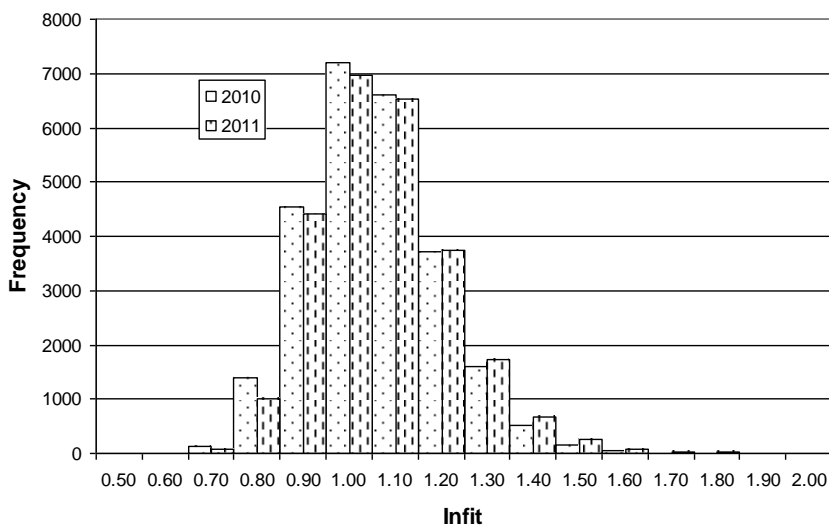


Figure 8 Distribution of person Infit statistics

Functioning of items across subgroups

As an example of using item characteristic curves to examine the performance of persons from different groups on the same items, Figure 9 compares the observed score ICCs and model ICCs for Item 32 and Item 44 from the 2010 test for all pupils, girls and boys. As can be seen, for these two items, girls and boys performed similarly across the ability range, and the observed ICCs are close to the model ICCs, with Item 32 slightly under-fit and Item 44 slightly over-fit. When the relative

ability is 0 (that is pupils' ability is the same as the difficulty of the item), the expected score is half of the maximum mark available on the item. Item 44 has a difficulty of 0.20 logits, and Item 32 has a difficulty of 0.87 logits, which is 0.67 logits harder than Item 44 as reflected in Figure 9.

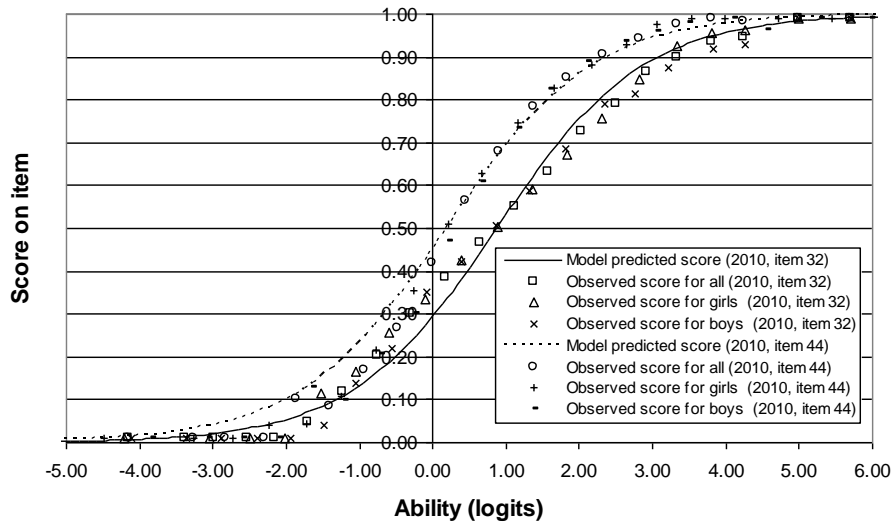


Figure 9 Observed score ICCs and model ICCs for two items from the 2010 test for all pupils, girls and boys

Figure 10 compares the difficulty of items estimated using pupil responses from different ethnic groups with those estimated using the overall sample for the 2010 and 2011 tests. As is expected, strong linear relationships exist in item difficulty between the different groups. A large gap can be seen in the difficulty range from 2 logits to 3 logits on the ability axis for the 2010 test. Inclusion of items in that range would produce more precise measures for pupils with ability in that range.

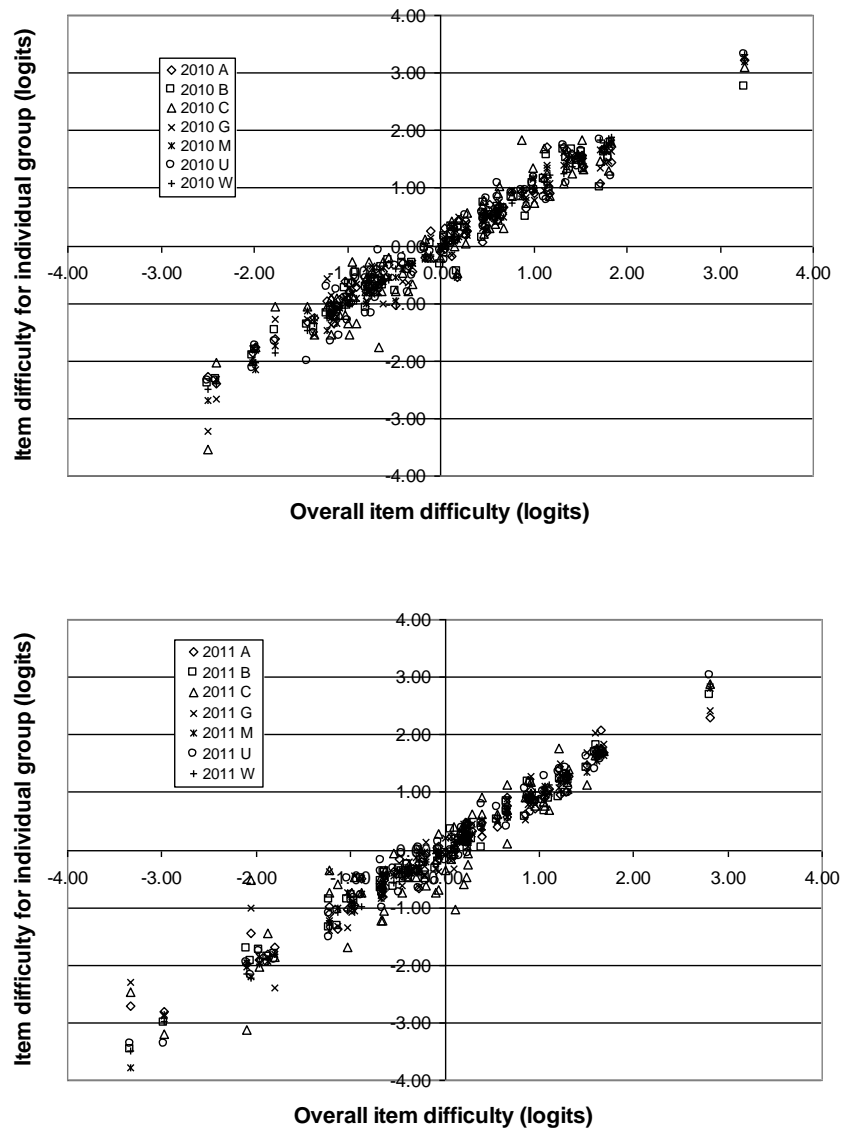


Figure 10 The relationship between item difficulties for individual ethnic groups and the item difficulties for the overall sample for the 2010 and 2011 tests

Figure 11 shows the differences in item difficulties between those estimated based on responses from different ethnic groups and those estimated using the responses from white pupils for the two tests. Here white pupils are taken as the reference group and the other ethnic groups the focal groups. Items with values of differences in difficulty between groups less than 0.43 logits are regarded as non-DIF items, items with values from 0.43 logits to 0.64 logits as slight to moderate DIF items, and values greater than 0.64 logits as moderate to large DIF items (see Tristan, 2006; Linacre, 2012). Apart from a few items for the Chinese group and the group for pupils of any other ethnicity, the differences for most of the items between the focal groups

and the reference group are within the range from -0.40 logits to 0.40 logits, suggesting that these items showed negligible DIF. For the groups with small sample size, the Chinese group, the group for pupils of other ethnicities and those unclassified, there were a number of items that showed substantial DIF. The significance level of DIF for an item between two groups can be tested using a statistic test such as a t-test, which will involve the use of the difference in item difficulty and the standard errors of the difficulty measures for the two groups. However, when the sample size is large, the standard error of the item difficulty measure can be small, and this can result in significant DIF for some items even though the magnitude of DIF may be small. Quite a few items showed significant DIF between groups based on the t-test. To account for this effect of sample size on the significance level of DIF between groups, Tristan (2006) suggested the use of a normalised standard error of estimation for the item difficulty measures. When the normalised standard errors were used, most of the items showed insignificant DIF between the focal groups and the reference group.

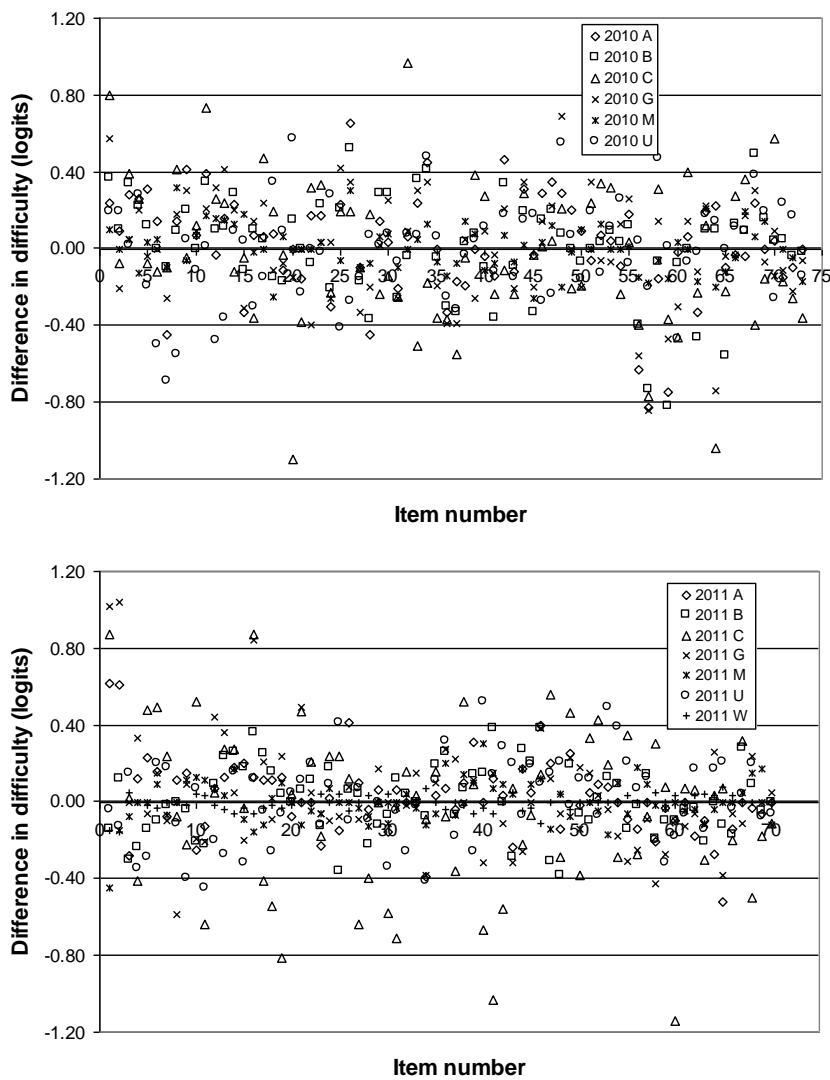


Figure 11 Differences in item difficulty between the focal groups and the reference group for the 2010 and 2011 tests

Figure 12 shows the differences in item difficulties between those estimated based on responses from girls and those estimated based on responses from boys (reference group); and the differences in item difficulty between the difficulties estimated using responses from pupils with EAL and the difficulties estimated using responses from pupils whose first language is English (reference group), for the 2010 and 2011 tests. Girls and pupils with English as an additional language are treated as focal groups. For the 2010 test, most items showed negligible DIF between girls and boys. Item 59 showed large DIF between pupils with EAL and those whose first language is English, with EAL pupils performing considerably better than other pupils with similar ability. There were also a few items showing slight to moderate DIF.

However, for most of the items, DIF is negligible and insignificant between the focal groups and the reference groups.

For the 2011 test, again, most items show negligible DIF between girls and boys. There were a couple of items that showed moderate DIF between pupils with EAL and pupils with English as their first language. These items appeared to be more difficult for pupils with EAL than for pupils with English as their first language who were at a similar ability level.

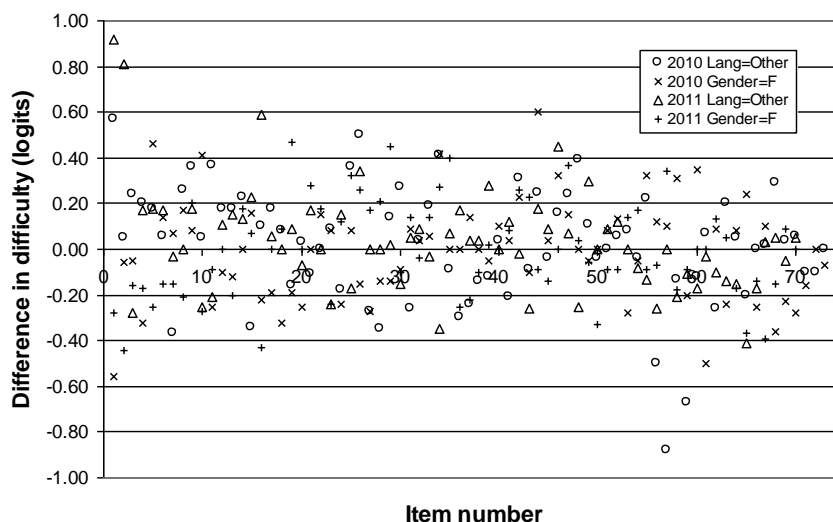


Figure 12 Differences in item difficulty between boys and girls and between pupils with EAL and pupils with English as their first language

Functioning of tests across subgroups

As comparisons between test-takers are normally made using the overall test scores, equivalence at the overall test level, which represents the aggregated effect of equivalence at item level, would be important (that is, for there to be no DTF between different subgroups). For the Rasch model, this can be investigated by examining the similarity of the test characteristic curves between different subgroups. Figure 13 compares the TCCs for the different ethnic groups with the TCC for the whole sample for the 2010 and 2011 tests. The curves almost completely overlap each other, suggesting that the overall test functioned similarly across the different groups. To look at the curves in more detail, the difference score between the expected score for an individual group and that for white pupils (the reference group) was calculated, and the distributions of the difference scores are shown in Figure 14. As can be seen from Figure 14, the 2010 test favoured some groups at the lower middle ability range but other groups at the higher ability range very slightly, compared with the reference group. For example, pupils from the Chinese group with the ability near -2.0 logits

would on average have achieved 0.20 mark higher than the average of white pupils, but over 0.30 marks lower at the ability near 1.0 logits. In contrast, black pupils would on average have achieved 0.30 marks lower than the average of white pupils at the ability near -2.0 but 0.30 higher at the ability near 1.0 logits. These difference scores are, however, minimal compared with the error scores associated with the estimated ability measures (equivalent to about 3 marks) and the average score on the test from the overall sample (53.61). For the 2011 test, the patterns are similar, with the magnitude of difference scores slightly larger.

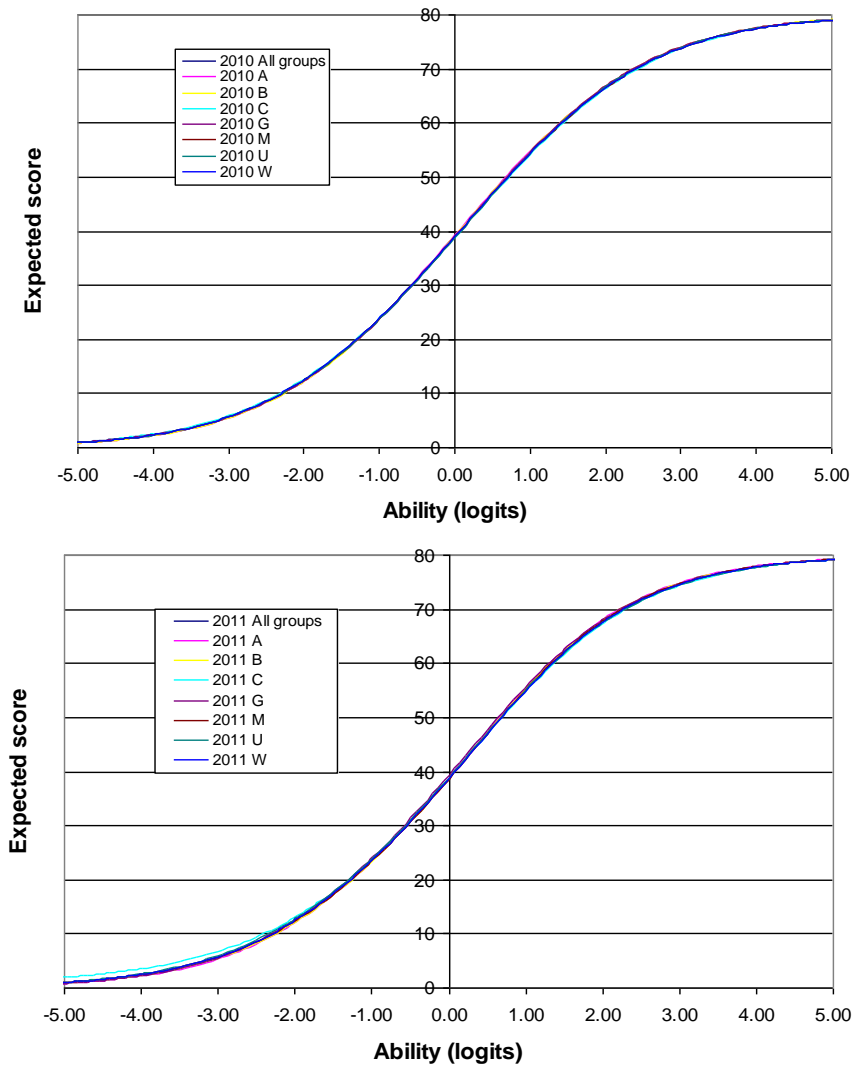


Figure 13 Test characteristic curves for different groups and the overall sample for the 2010 and 2011 tests

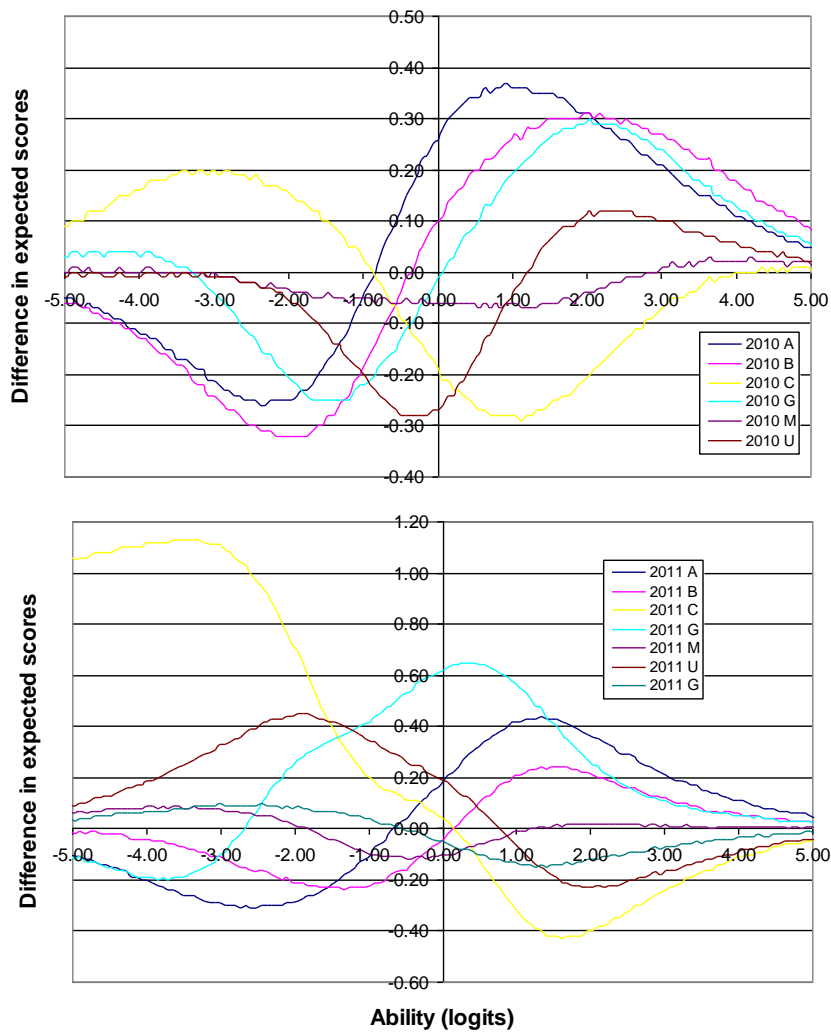


Figure 14 Difference in expected test scores between different ethnic groups and the reference group for the 2010 and 2011 tests

Figure 15 shows the expected difference scores between girls and boys and between EAL pupils and pupils with English as their first language for the 2010 and 2011 tests. Girls performed very slightly better than boys at the lower ability range but less well than boys at the higher ability range. EAL pupils performed not as well as other pupils at the lower ability range but better than others at the higher ability range on the test. These patterns are also similar for both the 2010 and 2011 tests. Again, the magnitude of the difference scores is not substantial.

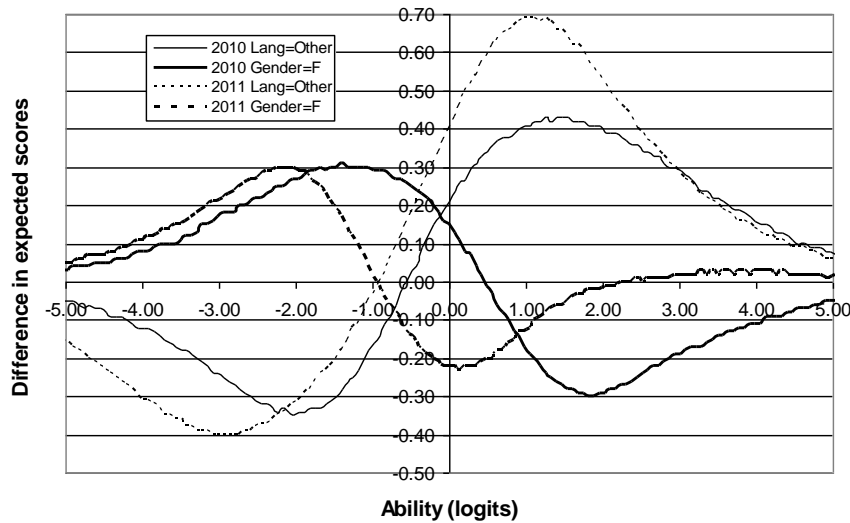


Figure 15 Difference in expected test scores between girls and boys and between pupils with EAL and pupils with English as their first language for the 2010 and 2011 tests

To investigate further the functioning of the overall test for EAL pupils and other pupils for the 2011 test, another approach was employed. This involved comparing the ability measures of EAL pupils, calculated using the item difficulties estimated, with the EAL pupils' responses, with the ability measures calculated using the item difficulties estimated with responses from other pupils. The results are shown in Figure 16. Both measures are almost on the identity line, indicating that the test functioned similarly for both groups. Figure 17 illustrates the relationship between the ability measures for girls estimated based on calibrating the items using responses from girls and their ability measures estimated based on calibrating the items using the responses from boys for the 2011 test. Again, the two measures are almost on the identity line.

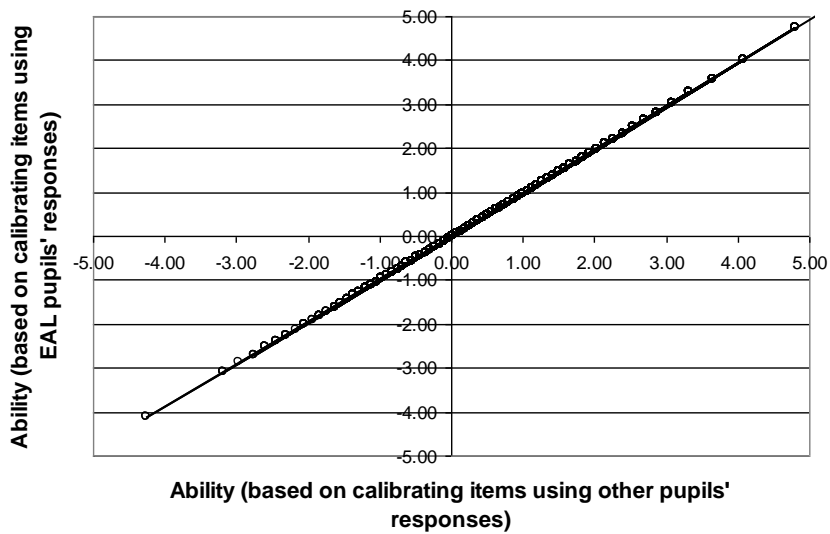


Figure 16 Relationship between person measures for EAL pupils estimated using their own responses to calibrate the items and those estimated using other pupils' responses to calibrate the items for the 2011 test

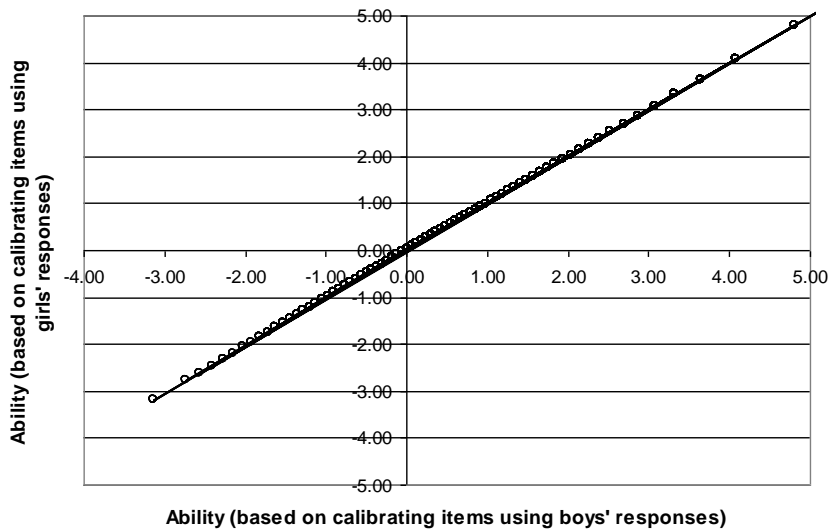


Figure 17 Relationship between person measures for girls estimated using their own responses to calibrate the items and the measures estimated using responses from boys to calibrate the items for the 2011 test

5. Concluding remarks

In this study, the performance of different groups on the Key Stage 2 National Curriculum science sampling test for the 2010 and 2011 series was examined. It was found that boys and girls performed similarly on the test for the two years studied. Black pupils and pupils with EAL performed less well than the average of the samples. Pupils with special educational needs performed considerably below the average of all pupils.

The two subtests correlate highly with each other, and the internal reliabilities for the subtests and the overall tests are high for the overall samples and the subgroups investigated. Results from factor analysis of the raw responses and the Rasch model residuals suggest that the tests can be essentially treated as a unidimensional measure of the intended construct.

A high level of measurement invariance is important for ensuring that a test assesses the same construct in the same way for different groups so that individuals from different subgroups are treated equally and that group comparison can be fair and meaningful. The degree of measurement equivalence of the science sampling tests across different groups in the samples at both item level and the overall test level was investigated using factor analysis and Rasch modelling. The strong correlations between the loadings on the first factor of the raw responses indicated that there was structural equivalence across the different groups. The Rasch model fits the test data for both years reasonably well. Results from the Rasch analysis indicated that most of the items in the tests functioned similarly across the different groups, with only a very few items showing large and significant DIF between some subgroups, suggesting a high level of measurement invariance in item parameters across the groups. At the overall test level, the tests also functioned in the same way across the groups, with negligible differential test functioning. It can therefore be concluded that the tests provide a measure of pupils' attainment in science on the same measurement scale for all subgroups over the two-year period studied.

Further investigation of the items showing DIF would be needed to identify the causes of DIF in order to improve future item writing. The Rasch item map could be a useful tool for constructing more effective test that targets pupils with different levels of ability to reduce measurement error further. However, the design of a test may also need to meet certain policy and other requirements.

It is noted that the present study only looked at the internal functioning of the Key Stage 2 science sampling tests across different subgroups. A more comprehensive validation of the test for its stated purposes would require the evaluation of evidence from other sources, including content representativeness in relation to the curriculum, demonstration of cognitive processes involved in pupils' responses to questions, and the appropriateness of the established performance standards in relation to the

attainment levels defined for the curriculum. Work in these areas is currently in progress.

Acknowledgements

The authors would like to thank the UK Department for Education and QCDA for providing item level data and pupil background information, and Sarah Maughan, Gordon Stobart and Colin Watson for their comments on an early draft of the paper.

This is a preprint of an article whose final and definitive form has been published in *Research Papers in Education* ©, 2012. *Research Papers in Education* is available online at: www.informaworld.com/. The article is available at: www.tandfonline.com/doi/full/10.1080/02671522.2012.742133.

6. References

- Baghaei, P. (2008). 'The Rasch Model as a Construct Validation Tool.' *Rasch Measurement Transactions* 22. 1145–46.
- Béguin, A. (2000). *Robustness of Equating High-Stakes Tests*. PhD thesis. The Netherlands: University of Twente.
- Barnard, J. (1999). 'Item Analysis in Test Construction.' In G. Masters and J. Keeves (eds.) *Advances in Measurement in Educational Research and Assessment*. 195–206. New York, USA: Pergamon.
- Carle, A., Millsap, R. and Cole, D. (2008). 'Measurement Bias Across Gender on the Children's Depression Inventory: Evidence for Invariance From Two Latent Variable Models.' *Educational and Psychological Measurement* 68. 281–302.
- Clauser, B. and Mazor, K. (1998). 'An NCME Instructional Module on Using Statistical Procedures to Identify Differentially Functioning Test Items.' *Educational Measurement: Issues and Practice* 17. 31–44.
- Cortina, J. (1993). 'What is Coefficient Alpha? An Examination of Theory and Applications.' *Journal of Applied Psychology* 78. 98–104.
- Cronbach, L. (1951). 'Coefficient Alpha and the Internal Structure of Tests.' *Psychometrika* 1. 297–334.
- Dai, G., Han, K., Hu, H., and Colarelli, S. (2011). 'Cultural Differences and Measurement Invariance of Selection Tools: A Case of Examining Chinese NEO PI-R Conscientiousness Scale.' *Journal of Chinese Human Resource Management* 1. 95–114.
- Dimitrov, D. (2006). 'Comparing Groups on Latent Variables: A Structural Equation Modeling Approach.' *Work* 26. 429–36.
- Drasgow, F. (1984). 'Scrutinizing Psychological Tests: Measurement Equivalence and Equivalent Relations with External Variables are Central Issues.' *Psychological Bulletin* 95, 134–35.
- Drasgow, F., and Kanfer, R. (1985). 'Equivalence of Psychological Measurement in Heterogeneous Populations.' *Journal of Applied Psychology* 70. 662–80.
- Embretson, S., and Reise, S. (2000). *Item Response Theory for Psychologists*. New Jersey, USA: Lawrence Erlbaum Associates.
- Flowers, C., Oshima, T. and Raju, N. (1999). 'A Description and Demonstration of the Polytomous DFIT Framework.' *Applied Psychological Measurement* 23. 309–26.

Furr, R. and Bacharach, V. (2008). *Psychometrics: An Introduction*. Thousand Oaks, CA: Sage Publications.

Furr, R. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. London, UK: Sage Publications.

Green, S., Lissitz, R. and Mulaik, A. (1977). 'Limitations of Coefficient Alpha as an Index of Test Unidimensionality.' *Educational and Psychological Measurement* 37. 827–838.

Hambleton, R., Swaminathan, H. and Rogers, H. (1991). *Fundamentals of Item Response Theory*. London, England: Sage Publications.

Hanson, B., and Béguin, A. (2002). 'Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in Common-Item Equating Design.' *Applied Psychological Measurement* 26 3–24.

He, Q. (2012). 'On-Demand Testing and Maintaining Standards for General Qualifications in the UK Using Item Response Theory: Possibilities and Challenges.' *Educational Research* 54. 89–112.

Isaacs, T. (2010). 'Educational Assessment in England.' *Assessment in Education* 17. 315–34.

Keeves, J. and Alagumalai, S. (1999). 'Item Banking.' In G. Masters and J. Keeves (eds.) *Advances in Measurement in Educational Research and Assessment*. The Netherlands: Elsevier Science. 23–42.

Koh, K. and Zumbo, B. (2008). 'Multi-Group Confirmatory Factor Analysis for Testing Measurement Invariance in Mixed Item Format Data.' *Journal of Modern Applied Statistical Methods*. 471–77.

Linacre J. (2012). *Winsteps® Rasch Measurement Computer Program User's Guide*. Beaverton, Oregon: Winsteps.com.

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.

Masters, G. (1999). 'Partial Credit Model.' In G. Masters and J. Keeves (eds.) *Advances in Measurement in Educational Research and Assessment*. The Netherlands: Elsevier Science. 98–109.

- Masters, G., and Keeves, J. (1999). *Advances in Measurement in Educational Research and Assessment*. The Netherlands: Elsevier Science.
- Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2012). 'Partial Estimates of Reliability: Parallel Form Reliability in the Key Stage 2 Science tests. In D. Opposs and Q. He (eds.) *Ofqual's Reliability Compendium*. Coventry, UK: Ofqual. 67–90.
- Meade, A. and Lautenschlager, G. (2004). 'A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance.' *Organizational Research Method* 7. 361–88.
- Milfont, T. and Fischer, R. (2010). 'Testing Measurement Invariance Across Groups: Applications in Cross-Cultural Research.' *International Journal of Psychological Research* 3. 111–21.
- Millsap, R. (2011). *Statistical Approaches to Measurement Invariance*. New York, USA: Routledge.
- Munkholm, M., Berg, B., Löfgren, B. and Fisher A. (2010). 'Cross-Regional Validation of the School Version of the Assessment of Motor and Process Skills.' *American Journal of Occupational Therapy* 64, 768–75.
- Ofqual (2012). *The Assessment Validity Programme*. Coventry, Ofqual.
www.ofqual.gov.uk/standards/validity/.
- Opposs, D. and Q. He (2012). *The Assessment Validity Programme*. Coventry: Ofqual.
- Oshima, T. and Morris, S. (2008). 'Raju's Differential Functioning of Items and Tests (DFIT).' *Items – Instructional Topics in Educational Measurement* 27.3. National Council on Measurement in Education.
- Pae, H. (2012). 'A Psychometric Measurement Model for Adult English Language Learners: Pearson Test of English Academic.' *Educational Research and Evaluation* 18. 211–29.
- Qualifications and Curriculum Authority (QCA) (1999a). *Science: The National Curriculum for England, Key Stages 1–4*. London: Department for Education and Employment and Qualifications and Curriculum Authority.
- Qualifications and Curriculum Authority (QCA) (1999b). *Test Development, Level Setting and Maintaining Standards*. London, QCA.
<http://webarchive.nationalarchives.gov.uk/20090608182316/testsandexams.qca.org.uk/18939.aspx>.

Raju, N., van der Linden, W. and Fleer, P. (1995). An IRT-Based Internal Measure of Test Bias with Applications for Differential Item Functioning. *Applied Psychological Measurement* 19. 353–68.

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

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, USA: Springer-Verlag.

Reckase, M. and Martineau, J. (2004). *The Vertical Scaling of Science Achievement Tests*. Paper commissioned by the Committee on Test Design for K–12 Science Achievement Center for Education National Research Council.
www7.nationalacademies.org/bota/Vertical%20Scaling.pdf.

Reise, S., Widaman, K. and Pugh, R. (1993). 'Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance.' *Psychological Bulletin* 114. 552–66.

Revelle, W. and Zinbarg, R. (2009). 'Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma.' *Psychometrika* 74. 145–54.

Schulz, W. and Fraillon, J. (2009). *The Analysis of Measurement Equivalence in International Studies Using the Rasch Model*. Paper presented to the symposium 'Rasch Measurement: Present, Past and Future' at the European Conference on Educational Research (ECER), Vienna, Austria.

Sijtsma, K. (2009a). 'On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha.' *Psychometrika* 74. 107–20.

Sijtsma, K. (2009b). 'Reliability Beyond Theory and into Practice.' *Psychometrika* 74. 169–73.

Sijtsma, K. (2012). 'Future of Psychometrics: Ask What Psychometrics Can Do for Psychology.' *Psychometrika* 77. 4–20.

Sijtsma, K. and Junker, B. (2006). 'Item Response Theory: Past Performance, Present Developments, and Future Expectations.' *Behaviormetrika* 33. 75–102.

Simon, M. (2008). *Comparison of Concurrent and Separate Multidimensional IRT Linking of Item Parameters*. PhD Thesis. Minneapolis, USA: University of Minnesota.

Smith, E. 2002. 'Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals.' *Journal of Applied Measurement* 3. 205–31.

- Smith, R. (2004). 'Detecting Item Bias with the Rasch Model.' *Journal of Applied Measurement* 5. 247–65.
- Taylor, C. and Lee, Y. (2011). 'Ethnic DIF in Reading Tests with Mixed Item Formats.' *Educational Assessment* 16. 35–68.
- Tan, J. and Yates, S. (2007). 'A Rasch Analysis of the Academic Self-Concept Questionnaire.' *International Education Journal* 2007. 470–84.
- Tennant, A., and Pallant, J. (2007). 'DIF Matters: A Practical Approach to Test if Differential Item Functioning Makes a Difference.' *Rasch Measurement Transactions* 20. 1082–84.
- Teresi, J., Ramirez, M., Lai, J. and Silver, S. (2008). 'Occurrences and Sources Of Differential Item Functioning (DIF In Patient-Reported Outcome Measures: Description of DIF Methods, and Review of Measures of Depression, Quality of Life and General Health.' *Psychology Science Quarterly* 50. 538–612.
- Thissen, D., Steinberg, L. and Mooney, J. (1989). 'Trace Lines for Testlets: A Use of Multiple-Category Response Models.' *Journal of Educational Measurement* 26. 247–60.
- Tristan, A. (2006). 'An Adjustment for Sample Size in DIF Analysis.' *Rasch Measurement Transactions* 20. 1070–71.
- Whetton, C. (2009). 'A Brief History of a Testing Time: National Curriculum Assessment in England 1989–2008.' *Educational Research* 51. 137–59.
- Wong, H., McGrath, C. and King, N. (2011). 'Rasch Validation of the Early Childhood Oral Health Impact Scale.' *Community Dent Oral Epidemiology* 39. 449–57.
- Wright, B. and Masters, G. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago, USA: MESA Press.
- Wright, B. and Stone, M. (1979). *Best Test Design: Rasch Measurement*. Chicago, USA: MESA Press.
- Wu, M. and Adams, R. (2007). *Applying the Rasch Model to Psycho-Social Measurement: A Practical Approach*. Educational Measurement Solutions, Melbourne.
- Yen, S., Bené, N. and Huynh, H. (2000). *The Effect of Content Integration on the Construct Validity of Reading Performance Assessment*. Paper presented at the 2000 Annual Meeting of the National Council on Measurement in Education. New Orleans, USA.

Yen, W. and Fitzpatrick, A. (2006). 'Item Response Theory.' In R.L. Brennan (ed.) *Educational Measurement* (4th Edition. Washington, DC: American Council on Education/Praeger. 111–54.

Yen, W. (1993). 'Scaling Performance Assessment: Strategies for Managing Local Item Dependence.' *Journal of Educational Measurement* 20. 187–213.

Yang, Y. and Green, S. (2011). 'Coefficient Alpha: A Reliability Coefficient for the 21st Century?' *Journal of Psychoeducational Assessment* 29. 347–63.

Zhang, B., Fokkema, M., Cuijpers, P., Li, J., Smits, N. and Beekman, A. (2011). 'Measurement Invariance of the Center for Epidemiological Studies Depression Scale (CES-D) Among Chinese and Dutch Elderly.' *Medical Research Methodology* 20. 11–74.

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 2014

© Crown copyright 2014

You may re-use this publication (not including logos) free of charge in any format or medium, under the terms of the [Open Government Licence](#). To view this licence, visit [The National Archives](#); or write to the Information Policy Team, The National Archives, Kew, Richmond, Surrey, TW9 4DU; or email: psi@nationalarchives.gsi.gov.uk

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

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