Monitoring Access to National Curriculum Assessments

Research Background

Jamal Abedi and Tandi Clausen-May

October 2012

Ofqual/12/5220

Contents

The purpose of this document

This volume is a companion to *Guidance on Monitoring Access to National Curriculum Assessments* (Ofqual, 2012). This research background has four purposes:

- Explain statistical models used for computing Differential Item Functioning (DIF), with a focus on logistic regression.
- Refer to previous research on how pupils with different background variables may perform differently on individual items.
- Consider issues on the reliability and validity of the overall ability measure.
- Report the results of DIF analyses carried out in 2012 with item-level data from 2011 Key Stage 2 mathematics and science in England.

By reading this document you will learn more about DIF analysis methodology and how to prepare and present information on the status of test items across groups of pupils with different background characteristics.

As explained in the introduction to *Guidance on Monitoring Access to National Curriculum Assessments*, DIF analyses must be carefully considered because:

- There is always a disparity in numbers between groups of pupils with protected characteristics and the majority.
- Assessment experts may question the way in which pupils are categorised in the National Pupil Database and how they should be categorised in future.
- The statistical analyses must be a proportionate response to potential concern.
- DIF must always be a good complement to other work that is carried out to provide evidence on minimising bias.
- Background information for sub-groups of pupils need to be available.
- Ideally, whole-cohort item-level data needs to be available. This will be easier to achieve where marking takes place electronically.

Issues in DIF analysis: a detailed discussion

There are a number of issues that affect the outcome of DIF analyses that must be considered before conducting such an analysis. These include:

- the statistical models used for computing DIF.

- the impact of the pupils' background variables on DIF.
- the reliability and validity of the overall assessment outcome that is used to establish the ability levels of the pupils.

Statistical models used for computing DIF

There are several statistical procedures that can be used to identify differentially functioning test items. These approaches include:

- the Mantel-Haentzel statistic.
- logistic regression.
- simultaneous item bias test (SIBTEST).
- the Standardisation procedure.
- various item response theory-based approaches (Clauser and Mazor, 1998).

One of the most commonly used approaches in identifying DIF is the Mantel-Haenszel method (Holland and Thayer, 1988). This approach is quite simple to use, but as we explain later in this section, it has several major limitations. In the Mantel-Haenszel approach, the subjects in the focal group and the reference group are matched based on the total score of the test (the total number of correct responses on the test with all multiple-choice items).

To conduct Mantel-Haenszel analyses on items, there must be enough subjects in the focal and reference groups. The minimum number of subjects suggested is 100 subjects in the smaller group; the focal group and the reference group should have a total of at least 500 subjects (Petersen, 1988). This could create a limitation on using many of the pupil background variables, because in some levels of some of the pupil background variables there may not be enough subjects to perform the analysis.

The null hypothesis for the Mantel-Haenszel procedure is that the proportion of correct responses for an item (item i) divided by the proportion of the incorrect responses for the item is the same across the focal and reference groups. That is:

$$H_0: P_{Ri} / q_{Ri} = P_{Fi}/q_{Fi}$$

In other words, the odds of a correct answer to item i for the reference group are equal to the odds of a correct answer for the focal group.

The alternative hypothesis states that the odds of a correct answer to item i for the reference group are weighted by a coefficient called aj to the correct answer for the focal group. That is:

$$H_1: P_{Ri} / q_{Ri} = \alpha_j P_{Fi}/q_{Fi}$$
$$\text{Where } \alpha_j = P_{Ri}q_{Fi} / P_{Fi}q_{Ri}$$

(Roussos at al., 2000)

Diverse statistical software could be used to calculate Mantel-Haenszel statistics for identifying test items with DIF. However, statisticians identified several major limitations in the use of Mantel-Haenszel approach. As Roussos, Schnipke, and Pashley (2000) indicated, the Mantel-Haenszel is more relevant for the cases when items follow two-parameter-logistic (2PL) item response theory (difficulty and discrimination). This approach may not yield valid results when items fit the three-parameter-logistic (3PL) item response function. Also, the Mantel-Haenszel procedure is typically used with multiple-choice items. It may be difficult to use with tests where both multiple-choice and extended-constructed response items are used. Also, the Mantel-Haenszel approach may have less power in identifying Information 3 DIF or I3-DIF, which identifies items with R-square difference of at least 0.130, when they are presented (see Roussos and others, 2000).

Logistic regression

The logistic regression approach as outlined by Zumbo (1999) is more suitable for researching questions related to DIF by subgroups of pupils, such as English as an additional language and pupils with disabilities. It is important to realise that different statistical models for computing DIF may provide different results. Some of these approaches produce fewer items showing DIF and some produce greater numbers of such items. This being the case, it is helpful to conduct DIF analyses using multiple approaches (at least two different DIF models) and to cross-check the results. If major discrepancies between the outcomes of different approaches are found then a close examination of the DIF items should be carried out.

The impact of the pupils' background variable on DIF

Students with different background variables, such as different special educational or assessment needs, may perform differently on individual items. This was demonstrated in a study conducted on data from three different states in the US where Abedi, Leon and Kao (2008) found many test items that were identified as having DIF for pupils with disabilities as a whole group. However, when the pupils were divided by type of disability, different patterns of DIF were observed. For example, the set of items that were identified as having DIF for pupils with severe learning disabilities was quite different from the items that were identified as having DIF for pupils with moderate learning difficulties and from those found for hearing impaired pupils. The study did not identify many items which had DIF across all categories of disability.

Reliability and validity of the overall ability measure

In order to identify items with DIF, the performance of two groups, the focal group and the reference group, must be matched based on their overall ability level, which is often referred to as the ‘conditioning’ variable. In an item response theory approach this is based on the latent ability estimate, or theta. In a classical approach to DIF such as the Mantel-Haentzel statistic, it is based on the pupils’ total scores on the test. It is of paramount importance that the total score used for pupil ability estimation be reliable and valid. In terms of reliability, the test must have high internal consistency; that is, measure a single construct or a single dimension (Cortina, 1993). The overall test score must be valid in terms of the content and the construct.

It is important to exclude test items that are identified as having DIF from the total test scores so that the total test score is not affected by items that could possibly be biased toward either the focal group or the reference group. For doing this, DIF can be conducted in two phases:

1. All test items are included in the computation of the total score.

2. The total score is recomputed by excluding items that were identified as DIF, and DIF analysis is conducted again using the new total score.

Some examples of the application of DIF

DIF analysis is often used to examine group differences between specific racial or ethnic groups or between males and females. For example, Hauser and Kingsbury (2004) explored differential functioning across pupil groups based on ethnicity and based on gender on items from the Idaho Standards Achievement Test. Zenisky, Hambleton and Robin (2004) explored gender DIF in a large-scale science assessment. Other studies have also examined incidences of DIF for pupils with limited English proficiency (Snetzler and Qualls, 2000).

DIF analyses have also been conducted for pupils with disabilities. Abedi, Leon and Kao (2008) conducted DIF analyses comparing the performance of pupils with disabilities (as the focal group) in reading with mainstream pupils as the reference group. Results indicated that pupils with disabilities performed differently on items with unnecessarily complex charts and graphs and complex linguistic structure. DIF analyses have also been used to examine the effects of accommodations that are provided to pupils with disabilities during testing (Bolt, 2004; Cohen and others, 2005; Koretz and Hamilton, 1999).

Application of DIF to National Curriculum assessments

The selection of variables to be used as a basis for the formation of focal and reference groups depends largely on assessment purpose, type and policies. For example, linguistic and cultural factors may be considered to be important for some assessments but less so for others. To identify major variables that could be used for the design of DIF analyses that are relevant and useful to the developers of national curriculum assessments such as key stage tests, several sources of information could be used:

- Review of literature on the sources of possible bias in National Curriculum assessments for subgroups of pupils. For instance, research on assessment for subgroups of pupils suggests that unnecessary linguistic complexity and convoluted charts and graphs make assessments less accessible to pupils in general and to those at risk of academic failure in particular. To examine the impact of these factors, the performance of focal groups on National Curriculum Assessment items with complex linguistic structures and difficult charts and graphs could be compared with the performance of pupils in the reference group. DIF analyses would then determine whether or not these features may lead to bias in the assessments.

- Information from focus groups that discuss accessibility issues in National Curriculum assessments. People with different backgrounds who are familiar with National Curriculum assessments and its target population can provide feedback and suggestions on how to control extraneous variables that may be the sources of unnecessary difficult language and specialised knowledge that is not related to the aim of the National Curriculum assessments.. The information collected on likely sources of construct irrelevant variances can then be used to establish focal and reference groups.

- Interviews with experts in the field of assessment who are knowledgeable about the country’s educational issues.

DIF analyses using National Curriculum assessments data could be conducted with the focal groups identified by these strategies, with the mainstream pupil population providing the reference group. However one must note that up to 2012, the item-level data collected for Key Stage 2 English and Mathematics was neither large nor representative, if compared to the item-level data available for Key Stage 2 Science. Therefore the Key Stage 2 Mathematics sample we have used as part of our analysis, presented in the next section, was a convenient sample with very limited generalisability. Such limitations can be addressed in future, provided that the responsible body decides to invest in the collection of rigorous samples of item-level data.

To have a more robust view of DIF in Key Stage 2 assessments, we also conducted DIF analysis on the Key Stage 2 Science for 2011 with large item-level data. In this document we first present the results of DIF conducted on the 2011 Key Stage 2 mathematics tests and then present and interpret the DIF findings on the 2011 Key Stage 2 Science test.

Post-hoc DIF analysis to improve the item development process: Key Stage 2 mathematics tests, 2011

In order to support *Guidance on Monitoring Access to National Curriculum Assessments*, data from the Key Stage 2 tests for 2011 in mathematics and science were used to conduct DIF analyses. Different DIF methodologies were carefully reviewed to select an approach or series of approaches that would provide more reliable outcomes given the structure of the data used in this study. As mentioned in the recommendations (pp.4–6), logistic regression methodology was deemed to be the most relevant approach for the analyses of the data.

We have used Table X in *Guidance on Monitoring Access to National Curriculum Assessments* (page 38) to organise the outputs obtained from the logistic regression approach to DIF.

Preliminary analyses

DIF analyses were performed on the 2011 Key Stage 2 mathematics and science test outcomes to obtain some information about the status of test items across groups of pupils with different background characteristics (gender, English as an additional language, free school meals and so on). Results of these preliminary analyses showed that:

- The DIF analytical methodology used in the analyses was quite powerful and provided clear outputs for interpretations.
- There are few items in the 2011 Mathematics and Science tests with DIF.

Reporting DIF analysis

DIF analyses were performed on the 2011 Key Stage 2 mathematics test items and on the science Key Stage 2 test items. Data from all test items from the three mathematics tests that were administered in 2011 were analysed. These tests are:

- Mental mathematics (MMM) test with 20 items, MMM_Q1 to MMM_Q20.
- Mathematics Test A (calculator not allowed) (MA) test with 40 items. MA_Q1 to MA_Q25ii (there were multiple items for some of the questions).
- Mathematics Test B (calculator allowed) (MB) test with 40 items. MB_Q1 to MB_Q26ii (there were multiple items for some of the questions).
- Science Test A (SA) test with 36 items.
- Science Test B (SB) test with 35 items.

Selection of variables to be used as the basis for forming focal and reference groups largely depends on assessment purpose, type and policies. For the mathematics content area (the content of analyses presented here), linguistic factors may play an important role since unnecessary linguistic complexity may be a major source of construct irrelevant factors. Some test items may have more complex linguistic structures than other items. Therefore, in order to select grouping variables that are relevant and useful to National Curriculum assessments, we decided to use pupils' language background status, specifically pupils with English as additional language. We compared item-level performance (DIF) of English as an additional language pupils with the performance of pupils whose first language is English. Students in the non-English as an additional language group were identified as the reference group and those in English as an additional language were identified as the focal group.

Students' gender was used as another grouping variable. This selection was justified based on the review of literature in test biases and DIF analyses and also based on Ofqual's policy and practice in previous years. We believe DIF analyses outcome by gender and English as an additional language will be informative for test item writers and for reporting the results of large-scale assessment.

Students' status on free school meal eligibility was also used as a grouping variable. Students were grouped into 'non-eligible' for free school meals as the reference group and 'eligible' for free school meals as the focal group.

Another major decision was on the statistical procedure for carrying out DIF analyses. There are several statistical procedures that can be used to identify differentially functioning test items. These approaches include the Mantel-Haentzel statistic, logistic regression, simultaneous item bias test (SIBTEST), the standardisation procedure and various item response theory-based approaches (Clauser and Mazor, 1998). We decided to use the logistic regression approach as outlined by Zumbo (1999) since it provides more powerful outcomes and, more importantly, the logistic regression approach identifies both uniform and non-uniform DIF (see, Zumbo, 1999). For the polytomous items, such as ordinal item responses (for example, graded response, Likert-type responses, scores from the extended item response) the ordinal regression approach is commonly used. The linear regression, with the continuous criterion variable, is an alternative to ordinal regression, assuming equal interval between score points.

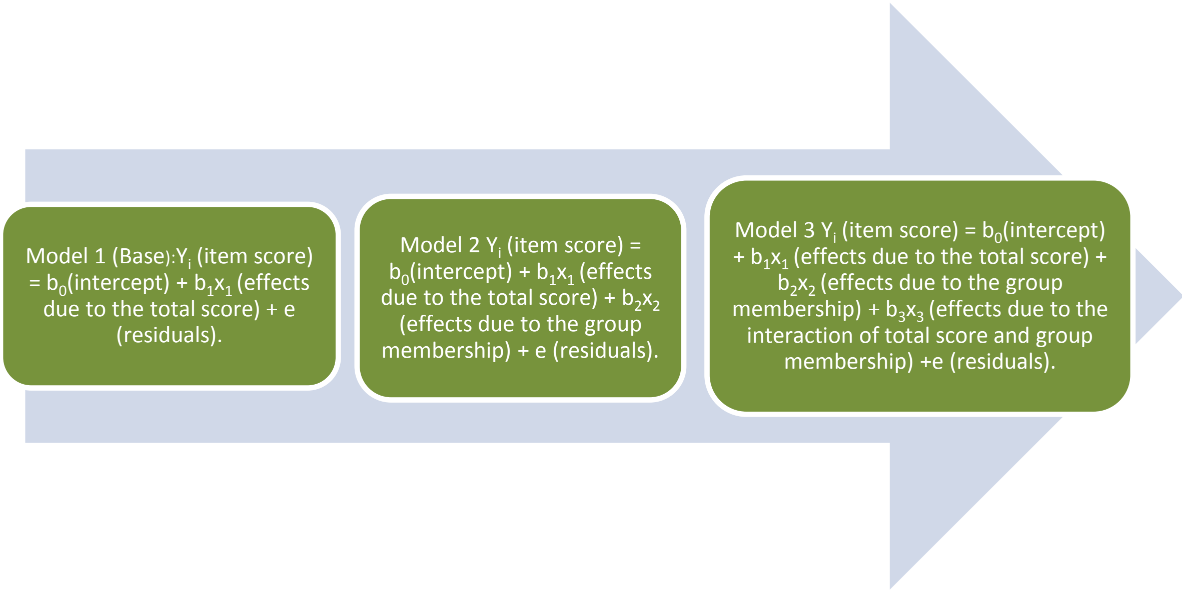
For each dichotomous test item, three logistic regression models were created and computed.

- In the first model, the item score (0 or 1 in dichotomous items) was used as the criterion variable. The total score was used as the predictor.

- In the second model, the item score was used as the criterion variable. Total test score and the group membership (focal versus reference group) were used as predictors. This model demonstrated uniform DIF.

- In the third model, item score was used as the criterion variable. Total test score, the group membership, plus the interaction of the total test score and the group membership were used as the predictors. This model provided information on a combination of uniform and non-uniform DIF.

Below is a representation of the three models:



- **Model 1 (base)**: Y_i (item score) = β_0(intercept) + β_1x_1 (effects due to the total score) + ε (residuals).

- **Model 2 (uniform DIF)**: Y_i (item score) = β_0(intercept) + β_1x_1 (effects due to the total score) + β_2x_2 (effects due to the group membership) + ε (residuals).

- **Model 3 (uniform and non-uniform DIF)**: Y_i (item score) = β_0(intercept) + β_1x_1 (effects due to the total score) + β_2x_2 (effects due to the group membership) + β_3x_3 (effects due to the interaction of total score and group membership) +ε (residuals).

An item was identified as potential DIF if the Chi-square for the third model (with total, group and total by group) has a significant p-value (Type I error rate) at or above the 0.01 level. Zumbo (1999) recommended an effect size of at least 0.130 R^2 of the difference between the R-square of the base model (the model with the total test score only as the predictor) and the R-square of the full model (the model with total score plus group membership plus interaction of total score and group).

Table 1 presents a summary of the logistic regression analyses for the three Key Stage 2 2011 mathematics tests (MMM – mental mathematics; MA – Test A with calculator not allowed; and MB –Test B with calculator allowed) and the grouping variable to identify items as DIF by gender. Table 2 presents similar results for the three mathematics tests by English as an additional language. The structure of the data in both tables is the same and is described below:

- Column 1 is the test item number.
- Column 2 (MR) is the mean of the test items for the reference group.
- Column 3 (MF) is the mean for the focal group.
- Column 4 is the Chi-square for the uniform and non-uniform DIF.
- Column 5 is the Chi-square significance for the uniform/non-uniform DIF.
- Column 6 is the Chi-square for the uniform DIF.
- Column 7 is the Chi-square significance for the uniform DIF.
- Column 8 is the R-square for the uniform/non-uniform DIF.
- Column 9 if the R-square for the uniform DIF
- Column 10 is DIF designation based on all the data presented in previous columns

Based on the data presented in Tables 1 and 2, we label each item as follows ('I' stands for information to item writers):

- 'No DIF (N)'.
- 'I1-DIF', which means that the item exhibits some differential functioning (the Chi-square for the uniform and non-uniform DIF is significant at the .05 level) but does not have a large enough R-square to be categorised as 'C' DIF. So item reviewers should take a look to see if there is any potential for differential functioning
- 'I2-DIF', which identifies items that have more evidence of DIF than the I1-DIF, (the Chi-square for the uniform and non-uniform DIF is significant beyond .01 level) but not having large enough R-square to be labelled as 'I3-DIF' or 'C' DIF. Again, this is a process of identification of items that may differentially perform across the focal and the reference group to provide information for item writers.
- 'I3-DIF' or 'C-DIF', which identifies items with R-square difference of at least .130 as per Zumbo's recommendation.

Table 3 presents mathematics items with No-DIF, I1-DIF, I2-DIF and I3-DIF or C-DIF for each of the three tests. Items with C-DIF would need the utmost attention since they could be biased against either the focal group or the reference group. Items with I2-DIF would need some attention, but the level of DIF would not be critical. Items with I1-DIF would need a quick review, but the differential performance may not be serious.

As the data in Table 3 shows, none of the items in any of the three mathematics tests were identified as I3-DIF. The majority of items had No DIF. There were some items with I1 and I2 DIF which may need to be reviewed. Certain experts would recommend a review of items with ‘I3’ or ‘C’ DIF first and then items with I2-DIF. If time permits, items with I1-DIF could also be reviewed.

In the review of items, the means of items in the reference (MR) and focal groups (MF) (Columns 2 and 3) should be considered. If the mean for the focal group is higher then the DIF favours the focal group and vice versa (please see Tables 1 to 3).

As indicated above, DIF analyses were performed on the 2011 Key Stage 2 Science tests. The grouping variable was pupils’ status on eligibility for free school meals. The reference group for the analyses included pupils who are not eligible for free school meals and the focal group consisted of pupils who are eligible for free school meals. The analyses performed for identification of items as DIF were the same as those conducted for Key Stage 2 mathematics tests; therefore, we do not present details of the analyses.

Table 4 presents DIF results for the Key Stage 2 2011 Science Test A (KSA) and Table 5 presents DIF results for Key Stage 2 2011 Science Test B (KSB).

Data in Tables 4 and 5 present evidence that the results of DIF for Science test items were similar to those presented for mathematics test items, indicating that no Science test items were identified as ‘I3’ DIF. However, there were many items with the potential of DIF that could be reviewed to make sure no DIF was present in the items.

DIF: summary results

The tables below report DIF by gender and English as an additional language for Key Stage 2 (2011) Mathematics tests A and B and by free school meals for Key Stage 2 (2011) Science tests A and B.

Table 1: DIF analyses for Key Stage 2 (2011) MMM (mental mathematics), MA (Test A) and MB (Test B) by gender*

Test item number	**Mean of test items reference group**	**Mean of the focal group**	**Chi-square for the uniform and non-uniform DIF**	**Chi-square significance for the uniform/non-uniform DIF**	**Chi-square for the uniform DIF**	**Chi-square significance for the uniform DIF**	**R-square for the uniform/non-uniform DIF**	**R-square for the uniform DIF**	**DIF designation based on all the data presented**
Item	MR	MF	χ^2-U/N	Sig U/N	χ^2-U	Sig U	R^2 U/NU	R^2 U	DIF
MMM_Q1	.91	.92	4.95	.084	4.11	.043	.003	.002	N
MMM_Q2	.89	.92	13.84	.001	13.65	.000	.006	.006	I2
MMM_Q3	.86	.80	40.37	.000	38.32	.000	.011	.010	I2
MMM_Q4	.80	.71	49.26	.000	48.44	.000	.011	.011	I2

MMM_Q5	.64	.52	46.91	.000	43.27	.000	.010	.009	I2
MMM_Q6	.91	.91	.06	.971	0	.981	.000	.000	N
MMM_Q7	.90	.91	2.94	.230	2.19	.139	.001	.001	N
MMM_Q8	.76	.74	.18	.915	.06	.815	.000	.000	N
MMM_Q9	.78	.78	4.74	.093	3.33	.068	.002	.001	N
MMM_Q10	.82	.79	3.63	.163	1.66	.197	.001	.000	N
MMM_Q11	.55	.62	84.81	.000	72.07	.000	.019	.016	I2
MMM_Q12	.69	.56	62.96	.000	62.15	.000	.014	.014	I2
MMM_Q13	.59	.59	8.59	.014	6.07	.014	.002	.001	I1
MMM_Q14	.61	.55	4.54	.103	4.54	.033	.001	.001	N
MMM_Q15	.60	.60	4.66	.097	4.27	.039	.001	.001	N
MMM_Q16	.76	.76	2.98	.225	2.91	.088	.001	.001	N
MMM_Q17	.69	.67	6.5	.039	.05	.830	.002	.000	I1

MMM_Q18	.50	.51	19.21	.000	16.67	.000	.004	.004	I2
MMM_Q19	.38	.41	31.36	.000	29.39	.000	.007	.006	I2
MMM_Q20	.37	.28	24.96	.000	24.331	.000	.006	.005	I2
MA_Q1	.83	.85	4.61	.100	3.53	.060	.002	.002	N
MA_Q2	.92	.93	1.09	.581	.86	.354	.001	.000	N
MA_Q3A	.93	.92	.67	.716	.15	.698	.001	.000	N
MA_Q3B	.84	.81	4.88	.087	3.60	.058	.002	.001	N
MA_Q4	.89	.89	.32	.851	.01	.915	.001	.000	N
MA_Q5	.83	.81	3.02	.221	2.28	.131	.001	.001	N
MA_Q6A	.88	.91	10.25	.006	8.65	.003	.005	.004	I2
MA_Q6B	.86	.85	.58	.747	.21	.65	.000	.000	N
MA_Q7A	.86	.79	35.76	.000	34.77	.000	.012	.012	I2

MA_Q7B	.88	.85	9.42	.009	8.70	.003	.004	.004	I2
MA_Q8i	.84	.85	2.65	.266	.80	.373	.001	.000	N
MA_Q8ii	.82	.83	2.46	.293	1.49	.223	.001	.001	N
MA_Q9	.88	.81	37.52	.000	37.36	.000	.012	.012	I2
MA_Q10	.88	.88	.43	.806	.10	.755	.000	.000	N
MA_Q11i	.86	.86	2.11	.349	1.34	.247	.001	.000	N
MA_Q11ii	.73	.74	7.75	.021	4.80	.028	.002	.001	I1
MA_Q12i	.77	.81	28.51	.000	28.24	.000	.008	.008	I2
MA_Q12ii	.67	.72	28.24	.000	28.21	.000	.007	.007	I2
MA_Q13a	.75	.78	7.95	.019	7.94	.005	.002	.002	I1
MA_Q13b	.59	.61	4.18	.124	4.17	.041	.001	.001	N
MA_Q14	.69	.62	16.42	.000	16.04	.000	.005	.005	I2
MA_Q15	.74	.78	21.13	.000	21.06	.000	.006	.006	I2

MA_Q16a	.73	.72	.15	.929	.01	.931	.000	.000	N
MA_Q16b	.57	.53	5.26	.072	3.44	.064	.001	.000	N
MA_Q16c	.63	.62	2.78	.249	.11	.739	.001	.000	N
MA_Q17	.70	.73	30.39	.000	15.91	.000	.009	.005	I2
MA_Q18	.53	.56	8.82	.012	8.34	.004	.002	.002	I1
MA_Q19a	.69	.67	.94	.624	.93	.336	.000	.000	N
MA_Q19b	.53	.46	18.15	.000	13.70	.000	.005	.004	I2
MA_Q20	.61	.58	1.69	.429	1.10	.295	.000	.000	N
MA_Q21i	.55	.54	8.10	.017	1.94	.163	.002	.001	I1
MA_Q21ii	.46	.42	1.48	.477	.40	.528	.001	.000	N
MA_Q22a	.45	.42	3.91	.142	.76	.385	.001	.000	N
MA_Q22b	.61	.54	21.13	.000	20.89	.000	.006	.006	I2
MA_Q23a	.45	.36	27.19	.000	26.82	.000	.007	.007	I2

MA_Q23b	.53	.48	6.20	.045	5.73	.017	.002	.002	I1
MA_Q24i	.59	.59	1.11	.575	.12	.731	.001	.000	N
MA_Q24ii	.42	.41	.55	.759	.01	.911	.000	.000	N
MA_Q25i	.33	.30	.93	.630	.41	.524	.000	.000	N
MA_Q25ii	.25	.23	.35	.838	.11	.743	.000	.000	N
MB_Q1	.92	.89	17.05	.000	12.58	.000	.008	.006	I2
MB_Q2	.91	.92	9.16	.010	8.90	.003	.005	.005	I1
MB_Q3i	.91	.92	4.69	.096	4.40	.036	.002	.002	N
MB_Q3ii	.84	.84	1.62	.446	1.62	.204	.000	.000	N
MB_Q4i	.97	.97	1.48	.477	.453	.501	.001	.000	N
MB_Q4ii	.89	.90	1.92	.384	1.78	.182	.001	.001	N
MB_Q5a	.81	.79	.33	.849	.03	.861	.000	.000	N

MB_Q5b	.87	.84	5.71	.058	4.17	.041	.002	.001	N
MB_Q6a	.94	.92	5.87	.053	3.67	.055	.004	.002	N
MB_Q6b	.87	.87	.48	.786	.37	.545	.000	.000	N
MB_Q7a	.83	.77	17.88	.000	17.86	.000	.006	.006	I2
MB_Q7b	.75	.72	.71	.700	.39	.531	.001	.001	N
MB_Q8	.79	.74	10.09	.006	10.01	.002	.002	.002	I2
MB_Q9	.77	.78	3.88	.144	3.67	.055	.001	.001	N
MB_Q10a	.69	.64	8.78	.012	7.25	.007	.003	.003	I1
MB_Q10b	.70	.69	1.74	.420	.51	.477	.000	.000	N
MB_Q11	.81	.84	26.31	.000	16.46	.000	.009	.006	I2
MB_Q12	.51	.49	1.45	.485	.01	.905	.000	.000	N
MB_Q13i	.76	.73	3.05	.217	.17	.678	.001	.000	N
MB_Q13ii	.74	.71	1.47	.479	.05	.821	.001	.000	N

MB_Q14a	.65	.66	6.19	.045	3.23	.072	.002	.001	I1
MB_Q14b	.68	.64	.86	.65	.86	.355	.000	.000	N
MB_Q15i	.82	.82	4.12	.127	3.14	.076	.002	.001	N
MB_Q15ii	.80	.80	2.49	.288	2.02	.155	.001	.001	N
MB_Q16	.62	.50	47.28	.000	47.26	.000	.012	.012	I2
MB_Q17	.76	.71	4.33	.115	4.33	.037	.001	.001	N
MB_Q18	.50	.50	3.72	.155	3.72	.054	.001	.001	N
MB_Q19i	.61	.57	.28	.871	.23	.633	.000	.000	N
MB_Q19ii	.60	.56	.42	.811	.32	.574	.000	.000	N
MB_Q20a	.64	.66	7.38	.025	7.20	.007	.002	.002	I1
MB_Q20b	.55	.54	.49	.781	.47	.493	.000	.000	N
MB_Q21	.63	.66	20.49	.000	20.38	.000	.006	.006	I2
MB_Q22i	.62	.53	22.56	.000	22.08	.000	.006	.006	I2

MB_Q22ii	.42	.36	6.05	.049	5.33	.021	.002	.002	I1
MB_Q23	.51	.51	4.74	.093	4.65	.031	.001	.001	N
MB_Q24	.55	.43	43.13	.000	41.87	.000	.011	.011	I2
MB_Q25i	.57	.58	18.17	.000	17.03	.000	.005	.005	I2
MB_Q25ii	.34	.37	30.03	.000	29.61	.000	.007	.007	I2
MB_Q26i	.54	.51	.37	.833	.01	.916	.000	.000	N
MB_Q26ii	.32	.26	5.64	.06	5.51	.019	.002	.002	N

*Spring 2011. Reference group, male =13,680 (51 per cent). Focal group, female = 13,163 (49 per cent).

Table 2: DIF analyses for Key Stage 2 (2011) MMM, MA and MB Mathematics tests by English as an additional language**

Item	**MR**	**MF**	**χ^2-U/N**	**Sig U/N**	**χ^2-U**	**Sig U**	**R^2 U/NU**	**R^2 U**	**DIF**
MMM_Q1	.92	.89	5.72	.221	5.64	.060	.003	.003	N
MMM_Q2	.90	.94	26.60	.000	17.07	.000	.011	.007	I2
MMM_Q3	.84	.77	15.49	.004	15.31	.000	.004	.004	I2
MMM_Q4	.77	.68	24.32	.000	22.16	.000	.005	.005	I2
MMM_Q5	.59	.54	4.63	.327	2.50	.287	.001	.001	N
MMM_Q6	.91	.88	6.87	.143	4.86	.088	.003	.003	N
MMM_Q7	.90	.94	18.02	.001	17.60	.000	.008	.008	I1
MMM_Q8	.76	.64	35.49	.000	35.49	.000	.010	.010	I2
MMM_Q9	.78	.76	5.14	.273	2.66	.265	.002	.001	N
MMM_Q10	.81	.83	12.86	.012	8.73	.013	.004	.002	I1

MMM_Q11	.57	.69	56.52	.000	54.51	.000	.013	.012	I2
MMM_Q12	.64	.58	7.87	.097	2.88	.237	.001	.000	N
MMM_Q13	.59	.63	14.70	.005	13.32	.001	.003	.003	I2
MMM_Q14	.58	.58	6.10	.192	3.24	.198	.001	.001	N
MMM_Q15	.60	.62	6.28	.179	5.32	.070	.002	.001	N
MMM_Q16	.77	.72	5.17	.270	2.82	.245	.002	.001	N
MMM_Q17	.69	.63	5.00	.288	4.44	.109	.001	.001	N
MMM_Q18	.52	.42	17.26	.002	15.49	.000	.004	.003	I2
MMM_Q19	.39	.48	45.48	.000	37.65	.000	.010	.008	I2
MMM_Q20	.33	.29	4.82	.307	2.43	.297	.001	.000	N
MA_Q1	.84	.81	3.50	.479	.87	.647	.002	.001	N
MA_Q2	.93	.92	5.17	.270	.59	.746	.003	.000	N

MA_Q3A	.93	.91	3.26	.515	1.09	.579	.002	.001	N
MA_Q3B	.83	.78	6.79	.147	3.56	.169	.002	.001	N
MA_Q4	.89	.87	2.75	.601	2.74	.254	.002	.002	N
MA_Q5	.82	.83	5.18	.269	4.98	.083	.002	.002	N
MA_Q6A	.89	.92	8.67	.070	6.80	.033	.004	.003	N
MA_Q6B	.85	.83	4.45	.349	.54	.765	.002	.000	N
MA_Q7A	.82	.83	3.47	.483	1.60	.450	.001	.000	N
MA_Q7B	.88	.81	14.82	.005	14.80	.001	.006	.006	I2
MA_Q8i	.85	.84	4.65	.326	.54	.764	.001	.000	N
MA_Q8ii	.82	.82	4.93	.294	.33	.849	.002	.000	N
MA_Q9	.85	.82	2.25	.690	.27	.874	.000	.000	N
MA_Q10	.88	.83	9.27	.055	5.34	.069	.004	.003	N
MA_Q11i	.86	.86	15.67	.003	1.28	.528	.006	.000	I2

MA_Q11ii	.74	.73	13.95	.007	.93	.628	.004	.000	I2
MA_Q12i	.79	.76	4.81	.307	1.64	.440	.002	.001	N
MA_Q12ii	.70	.67	1.28	.865	1.18	.56	.000	.000	N
MA_Q13a	.77	.72	4.88	.300	4.81	.090	.001	.001	N
MA_Q13b	.59	.62	12.09	.017	9.18	.010	.004	.003	I1
MA_Q14	.66	.62	5.52	.238	1.15	.56	.002	.001	N
MA_Q15	.77	.70	11.50	.021	7.45	.024	.003	.002	I1
MA_Q16a	.73	.71	3.98	.409	.12	.941	.001	.000	N
MA_Q16b	.55	.55	4.67	.323	2.86	.239	.001	.000	N
MA_Q16c	.63	.60	1.59	.810	.96	.618	.000	.000	N
MA_Q17	.72	.71	2.41	.661	2.21	.332	.001	.001	N
MA_Q18	.55	.48	8.23	.084	6.03	.049	.002	.001	N
MA_Q19a	.69	.66	1.68	.79	1.22	.54	.000	.000	N

MA_Q19b	.50	.48	3.62	.459	1.47	.480	.001	.000	N
MA_Q20	.60	.55	2.96	.565	1.43	.489	.001	.000	N
MA_Q21i	.55	.56	6.98	.137	6.87	.032	.002	.002	N
MA_Q21ii	.44	.46	6.86	.143	5.88	.053	.002	.002	N
MA_Q22a	.44	.45	1.49	.685	1.21	.546	.000	.000	N
MA_Q22b	.58	.56	1.73	.629	.63	.731	.001	.000	N
MA_Q23a	.41	.43	3.99	.136	3.11	.078	.001	.001	N
MA_Q23b	.51	.52	3.57	.168	1.21	.272	.001	.001	N
MA_Q24i	.59	.57	.63	.732	.02	.886	.001	.000	N
MA_Q24ii	.41	.42	4.56	.102	1.89	.169	.001	.001	N
MA_Q25i	.32	.30	.43	.808	.34	.562	.000	.000	N
MA_Q25ii	.24	.25	.35	.839	.11	.741	.000	.000	N

MB_Q1	.91	.84	26.37	.000	25.22	.000	.015	.010	I2
MB_Q2	.91	.92	3.01	.556	2.06	.357	.002	.001	N
MB_Q3i	.92	.90	3.71	.446	1.34	.513	.002	.001	N
MB_Q3ii	.84	.81	4.08	.395	4.02	.134	.001	.001	N
MB_Q4i	.97	.95	6.66	.155	6.64	.036	.008	.008	N
MB_Q4ii	.90	.88	5.33	.255	1.48	.477	.002	.000	N
MB_Q5a	.80	.79	3.97	.410	.052	.974	.001	.000	N
MB_Q5b	.86	.85	4.16	.385	.36	.834	.001	.000	N
MB_Q6a	.93	.94	5.44	.245	3.82	.148	.003	.002	N
MB_Q6b	.86	.90	13.18	.004	11.67	.003	.005	.005	I2
MB_Q7a	.81	.74	13.24	.010	9.96	.007	.004	.003	I1
MB_Q7b	.74	.72	3.46	.484	3.41	.182	.001	.001	N
MB_Q8	.77	.76	3.85	.426	2.38	.304	.001	.000	N

MB_Q9	.78	.75	3.31	.507	.05	.973	.001	.000	N
MB_Q10a	.66	.68	3.59	.310	2.76	.252	.001	.001	N
MB_Q10b	.71	.63	10.76	.029	7.79	.020	.003	.020	I1
MB_Q11	.82	.85	9.10	.059	5.79	.055	.003	.020	N
MB_Q12	.50	.48	4.35	.360	.78	.677	.001	.000	N
MB_Q13i	.74	.75	7.42	.115	6.65	.036	.002	.002	N
MB_Q13ii	.72	.74	8.62	.071	7.92	.019	.002	.002	N
MB_Q14a	.66	.68	7.47	.113	2.55	.279	.002	.001	N
MB_Q14b	.67	.62	9.53	.049	1.40	.495	.003	.000	I1
MB_Q15i	.82	.80	2.21	.531	1.94	.379	.001	.001	N
MB_Q15ii	.80	.78	3.38	.337	1.68	.431	.001	.001	N
MB_Q16	.57	.51	7.42	.115	4.93	.085	.001	.001	N
MB_Q17	.73	.73	1.40	.497	1.14	.286	.000	.000	N

MB_Q18	.51	.47	6.86	.144	1.31	.520	.001	.000	N
MB_Q19i	.59	.58	4.57	.206	1.38	.503	.001	.001	N
MB_Q19ii	.58	.57	5.06	.168	1.31	.520	.001	.000	N
MB_Q20a	.65	.64	6.71	.152	4.12	.127	.002	.001	N
MB_Q20b	.54	.53	3.53	.474	.80	.672	.001	.000	N
MB_Q21	.66	.59	5.56	.135	4.76	.093	.002	.001	N
MB_Q22i	.58	.53	4.76	.190	3.84	.147	.001	.001	N
MB_Q22ii	.40	.34	2.34	.505	2.33	.312	.001	.001	N
MB_Q23	.51	.49	3.53	.317	3.13	.209	.001	.001	N
MB_Q24	.50	.43	4.57	.206	3.75	.153	.001	.001	N
MB_Q25i	.57	.60	8.26	.016	8.25	.004	.002	.002	I1
MB_Q25ii	.35	.36	4.79	.091	3.83	.050	.001	.001	N
MB_Q26i	.53	.52	2.62	.270	.94	.333	.001	.001	N

MB_Q26ii	.29	.31	10.80	.005	2.83	.092	.003	.001	I2

** Spring 2011. Reference group, non-English as an additional language or language group major 'English' = 22,550 (84 per cent). Focal group, English as an additional language, language group major or 'other' = 4,266 (15.9 per cent).

Table 3: Number of items identified as No-DIF, I1-DIF, I2-DIF and I3-DIF

Test	**Gender**				**English as an additional language**			
	No DIF	I1	I2	I3	No DIF	I1	I2	I3
MMM	9	2	9	0	10	2	8	0
MA	23	5	12	0	35	2	3	0
MB	25	5	10	0	33	4	3	0

We also computed DIF based on the free school meals variable using pupils who receive free meals as the focal group and those who are not eligible as the reference group. We used all test items in Key Stage 2 Science 2011Test A and Test B. We have not found any items in either of those tests to have serious DIF issues (significant DIF) in terms of logistic regression multiple R-squared.

Table 4: DIF analyses for the 2011 Key Stage 2 Science Test A by free school meal eligibility****

Item	MR	MF	χ^2/ANOVA U/N	Sig U/N	χ^2/ANOVA U	Sig U	R^2 U/NU	R2 U	DIF
SA_Q1a	.98	.97	784.26	.000	783.21	.000	.000	.000	N
SA_Q1b	.95	.91	1861.00	.000	1861.09	.000	.000	.000	N
SA_Q1c	1.03	0.98	3481.75	.000	5219.40	.000	.001	.001	I1
SA_Q1di	.77	.66	4270.83	.000	4266.73	.000	.000	.000	N
SA_Q1dii	.90	.84	2538.23	.000	2538.02	.000	.000	.000	N
SA_Q2a	.82	.73	4166.11	.000	4161.03	.000	.000	.000	N
SA_Q2b	.72	.61	5131.04	.000	5130.98	.000	.000	.000	N
SA_Q2c	1.90	1.95	631.75	.000	901.47	.000	.001	.000	N
SA_Q3a	.98	.96	1426.03	.000	1425.92	.000	.000	.000	N
SA_Q3b	.89	.85	2975.04	.000	2958.27	.000	.000	.000	N

SA_Q3c	.77	.69	3586.59	.000	3586.59	.000	.000	.000	N
SA_Q3d	.45	.32	4634.66	.000	4633.10	.000	.000	.000	N
SA_Q3e	.58	.43	5364.92	.000	5352.16	.000	.000	.000	N
SA_Q4a	.83	.76	2726.98	.000	2726.28	.000	.000	.000	N
SA_Q4b	.70	.60	4776.51	.000	4739.00	.000	.002	.000	N
SA_Q4c	.92	.88	2304.69	.000	2304.57	.000	.000	.000	N
SA_Q4d	.83	.74	3973.46	.000	3973.45	.000	.000	.000	N
SA_Q5a	1.14	1.31	3300.62	.000	4951.07	.000	.000	.000	N
SA_Q5b	.71	.57	5800.92	.000	5781.74	.000	.001	.000	N
SA_Q5c	.66	.54	4972.71	.000	4961.65	.000	.000	.000	N
SA_Q5di	.88	.80	4327.14	.000	4324.94	.000	.000	.000	N
SA_Q5dii	.71	.66	2870.28	.000	2869.72	.000	.001	.001	N
SA_Q5e	.58	.49	4025.00	.000	4023.59	.000	.000	.000	N

SA_Q6a	.87	.81	3300.97	.000	3296.56	.000	.000	.000	N
SA_Q6b	.70	.65	3085.79	.000	3084.86	.000	.001	.001	N
SA_Q6c	.44	.31	5124.71	.000	5116.05	.000	.001	.000	N
SA_Q6d	1.43	1.58	2861.82	.000	4281.63	.000	.000	.000	N
SA_Q7a	.63	.48	7019.42	.000	7011.54	.000	.000	.000	N
SA_Q7b	.82	.71	6033.32	.000	6029.24	.000	.001	.000	N
SA_Q7c	.85	.79	2898.67	.000	2898.65	.000	.000	.000	N
SA_Q7d	.82	.71	6033.32	.000	6029.24	.000	.001	.000	N
SA_Q7e	.77	.66	4491.36	.000	4484.05	.000	.000	.000	N
SA_Q7f	.43	.30	5753.32	.000	5745.11	.000	.001	.001	I1
SA_Q8a	1.63	1.57	2787.52	.000	4167.59	.000	.007	.006	I2
SA_Q8b	.46	.34	5970.94	.000	5962.10	.000	.000	.000	N

*** Spring 2011. Reference group, not eligible for free school meal = 21,878 (81.5 per cent). Focal group, free school meal eligible = 4,965 (18.5 per cent).

Table 5: DIF analyses for the Key Stage 2 Science (2011) Test B by free school meal eligibility***

Item	MR	MF	χ^2/ANOVA U/N	Sig U/N	χ^2/ANOVA U	Sig U	R^2 U/NU	R^2 U	DIF
SB_Q1a	.79	.68	5878.58	.000	5877.87	.000	.000	.000	N
SB_Q1b	.73	.67	2350.31	.000	2348.49	.000	.000	.000	N
SB_Q1c	1.20	1.35	2385.46	.000	3576.87	.000	.000	.000	N
SB_Q1d	1.42	1.62	3115.64	.000	4672.66	.000	.000	.000	N
SB_Q2a	.93	.88	2958.56	.000	2957.23	.000	.000	.000	N
SB_Q2b	.95	.91	2366.41	.000	2366.02	.000	.000	.000	N
SB_Q2c	.85	.75	5091.53	.000	5091.04	.000	.000	.000	N
SB_Q2d	.61	.49	6310.03	.000	6305.99	.000	.000	.000	N
SB_Q2e	.73	.60	4668.26	.000	4652.14	.000	.000	.000	N
SB_Q3a	.94	.91	2601.46	.000	2596.72	.000	.000	.000	N

SB_Q3b	.87	.80	4022.75	.000	4022.39	.000	.000	.000	N
SB_Q3c	.50	.43	3817.64	.000	3795.16	.000	.001	.000	N
SB_Q3d	.53	.47	2489.13	.000	2476.42	.000	.001	.001	I1
SB_Q4a	.63	.48	5777.10	.000	5773.51	.000	.000	.000	N
SB_Q4b	.48	.34	7404.60	.000	7404.48	.000	.000	.000	N
SB_Q4c	.70	.60	3134.89	.000	3133.76	.000	.000	.000	N
SB_Q4d	.94	.91	1537.28	.000	1536.53	.000	.000	.000	N
SB_Q4e	.83	.76	3409.47	.000	3400.98	.000	.000	.000	N
SB_Q5b	1.39	1.44	1868.36	.000	2802.16	.000	.000	.000	N
SB_Q7c	1.39	1.44	1868.36	.000	2802.16	.000	.000	.000	N
SB_Q7e	1.06	1.16	2146.49	.000	7279.37	.000	.000	.000	N

Spring 2011. Reference group, not eligible for free school meal = 21,878 (81.5 per cent). Focal group, free school meal eligible = 4,965 (18.5 per cent).

References

Abedi, J., Leon, S. and Kao, J. (2008) *Examining Differential Item Functioning in Reading Assessments for Students with Disabilities*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing. Available online at www.cse.ucla.edu/products/reports/R744.pdf (accessed 1st December 2011).

Abedi, J and Clausen-May, T. (2012) Guidance on Monitoring Access to National Curriculum Assessments. Ofqual.

Bolt, S. E. (2004) Using DIF Analyses to Examine Several Commonly-held Beliefs about Testing Accommodations for Pupils with Disabilities. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA, April. Available online at http://education.umn.edu/NCEO/Presentations/NCME04bolt.pdf (accessed 20th June 2011).

Clauser, B. E. and Mazor, K. M. (1998) Using Statistical Procedures to Identify Differentially Functioning Test Items. Educational Measurement: Issues and Practice, 17(1), 31–44.

Cohen, A. S., Gregg, N. and Deng, M. (2005) The Role of Extended Time and Item Content on a High-stakes Mathematics Test. Learning Disabilities Research and Practice, 20(4), 225–233.

Cortina, J. M. (1993) What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78(1), 98–104.

Hauser, C. and Kingsbury, G. (2004) *Differential Item Functioning and Differential Test Functioning in the 'Idaho Standards Achievement Tests' for Spring 2003*. Lake Oswego, OR: Norwest Evaluation Association.

Holland, P. W., & Thayer, D. T. (1988) Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer and H. I. Braun (eds) *Test Validity.* Hillsdale, NJ: Erlbaum (pp. 129–145).

Koretz, D. and Hamilton, L. (1999) *Assessing students with Disabilities in Kentucky: The Effects of Accommodations, Format, and Subject.* (CRESST Tech. Rep. No. 498). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Peterson, N. S. (1988) *DIF Procedure for Use in Statistical Analysis.* Princeton, NJ: Educational Testing Service.

QCDA (2010) Test Development – Level Setting and Maintaining Standards. Available online at http://webarchive.nationalarchives.gov.uk/20110813032310/qcda.gov.uk/resources/publication.aspx?id=3750a7cb-1ec5-450a-9186-a8161f1e7ddf (accessed 22 August 2012).

Roussos, L, Schnipke, D. L. and Pashley, P. (2000) *A Formulation of Mantel-Haenszel Differential Item Functioning Parameter with Practical Implications*. Law School Admission Council, Newtown, Pennsylvania.

Snetzler, S. and Qualls, A. L. (2000) Examination of Differential Item Functioning on a Standardized Achievement Battery with Limited English Proficient students. *Educational and Psychological Measurement*, 60(4), 564–577.

Zenisky, A. L., Hambleton, R. K. and Robin, F. (2004) DIF Detection and Interpretation in Large-scale Science Assessments: Informing Item Writing Practices. *Educational Assessment*, 9(1–2), 61–78.

Zumbo, B. D. (1999) *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Available online at http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf (accessed 10 March 2006).

Bibliography

Further reading

DIF analysis

Clauser, B. E. and Mazor, K. M. (1998) Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*, 17(1), 31–44.

Ofqual (2012) *Guidance on Monitoring Access to National Curriculum Assessments:* Coventry: Ofqual.

Penfield, R. D. and Lam, T. C. M. (2000) Assessing Differential Item Functioning in Performance Assessment. *Educational Measurement: Issues and Practice*, 19(3), 5–15.

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 2012.

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