Guidance

Guidance on Monitoring Access to National Curriculum Assessments

Regulatory Arrangements for National Assessments: National Curriculum and Early Years Foundation Stage (Ofqual, 2011)

Audience: The Responsible Body and Independent Research Providers.

Ofqual/12/5219
October 2012

Contents

Purpose of recommendations

Guidance on Monitoring Access to National Curriculum Assessments suggests ways in which the accessibility of a National Curriculum assessment, such as a key stage test, could be evaluated after the assessment has been released. This volume is designed to help the responsible body check that all the questions in the assessment were accessible to the widest possible range of pupils.

If you are seeking the principles which should guide the development of clear assessment questions, please refer to *Guidance on the Principles of Language Accessibility in National Curriculum Assessments* and *What Makes Accessible Questions. The Principles of Language Accessibility in National Curriculum Assessments: Research Background* (Ofqual, 2012).

The recommendations in this guidance suggest two approaches to post-release evaluation of assessments.

- The first proposal draws on the knowledge and expertise of teachers who work with pupils who have a range of special educational and assessment needs. Working closely with a small group of their own pupils, the teachers collect detailed, focused information about the way in which their pupils accessed the live assessment. This 'case study' methodology provides detailed, insightful information about the way in which pupils who actually took the assessments were affected by the language, layout and presentation of each question.

- The second proposal suggests the use of a statistical technique: Differential Item Functioning (DIF) analysis. This can be used to identify any questions, or parts of a question, on which pupils belonging to a particular group tend to perform less well than might be expected from their overall performance on the assessment. If most pupils in a particular group have difficulty with one specific question then it is possible that the wording or layout of the question is somehow misleading them, so the problem should be investigated further with those pupils.

The two proposals in this volume are complementary. The DIF analyses may provide useful independent statistical data relating to individual questions in an assessment, while the qualitative data obtained from the teachers working with pupils will provide a rich source of information which assessment developers will find most valuable.

Introduction

This guidance suggests ways in which the accessibility of a National Curriculum assessment, such as a key stage test, might be evaluated after the assessment has been released.[1] In England, the responsible body, and agencies appointed by such a body to develop a National Curriculum Assessment, already carry out pre-testing statistical analyses to check that all the questions in the assessment will be accessible to the widest possible range of pupils. To evaluate assessment questions after the tests have been released, the responsible body needs to be convinced that additional post-release statistical analyses are:

- a proportionate response to potential concern.
- a good complement to other work that is carried out to provide evidence on minimising bias and measurement error.
- feasible in terms of the limited amount of resource available.
- possible due to availability of whole-cohort, item-level data, which is hard to generate in a system which uses paper-based marking. However, item-level data will be generated automatically where the National Assessments are electronically marked.

The above points do not affect Proposal 1 but must be carefully considered in relation to Proposal 2:

- The appropriateness of carrying out DIF analysis may be affected by the disparity in numbers between groups of pupils with protected characteristics (the focal group) and the majority (the reference group). However, we believe for many of the focal groups there are large enough subjects to perform DIF analyses.
- Assessment experts may question the way in which pupils are categorised in the National Pupil Database and how they should be categorised in future. To illustrate the point, pupils categorised with particular special educational needs may have a wide variety of needs.

[1] Recommendations describing the principles which should guide the development of clear assessment questions can be found in *Guidance on the Principles of Language Accessibility in National Curriculum Assessments* and *What Makes Accessible Questions. The Principles of Language Accessibility in National Curriculum Assessments: Research Background* (Ofqual, 2012).

Nationally comparable evidence on the fairness of assessments

In future, it is possible that agencies with an interest in fairness may use freedom of information to request data from the responsible body in order to carry out post-release analysis using item-level data, when such data becomes available.

Provided that all the contentious points had been weighed, this document suggests uniform procedures for the analysis of assessment data. A diverse range of agencies, interest groups and academic institutions could produce nationally comparable analyses of assessment in England, when using the procedures described in this guidance

Even if the responsible body does not feel that post-release analyses are a priority due to the limited amount of public money available for work on fairness, this guidance may support independent research providers with an interest in educational assessment.

A history of good practice

Accessibility is clearly essential to ensuring the validity of any assessment for all of the different groups of pupils taking it. Since the inception of the National Curriculum assessments, there has always been a strong emphasis on the development of accessible questions. In part, this was because the original key stage tests were designed for the whole cohort of pupils and needed to be as inclusive as possible. Furthermore, key stage tests have always been statutory and intended for any pupil who was working at the appropriate level.

Even before the Disability Discrimination Act of 1995 it was not possible to argue that a pupil with a disability that might inhibit access should not be doing the Key Stage tests, so fair access had to be ensured as far as possible. Close attention and extensive resources have always been devoted to this, with a development cycle that includes the thorough scrutiny of all test questions by experts who have worked with pupils with a range of special educational and assessment needs. Modified assessments were produced for some specific groups of pupils, and it would be interesting to verify the effectiveness of these modifications. We must also make sure the modifications on the assessments do not alter the focal construct.

The proposals

This volume contains two proposals for ways in which the accessibility of a National Curriculum Assessment may be evaluated and assured.

- Proposal 1 relates to the creation of Accessibility Teacher Review Panels. These build on the sound practice already established of detailed input at the development stage of an assessment cycle from experts who have knowledge of pupils with different special educational and assessment needs. The proposed introduction of Accessibility Teacher Review Panels at the evaluation stage of the cycle would extend this by exploiting the insight and understanding of teachers who work with such pupils to check on the accessibility of the live assessments. This would provide a rich source of data based directly on their experiences and on those of their pupils. The data that would be collected through the Accessibility Teacher Review Panels in Proposal 1 would be mainly qualitative.

- In contrast, Proposal 2 is based on the use of DIF analysis, so it is quantitative in nature. This proposal builds on the current common practice of the agencies responsible for the development of National Curriculum Assessments such as key stage tests, which regularly use DIF analysis to assist the test development process. The results of DIF could help identify test items that might be biased against certain groups of test takers with certain background characteristics. This proposal makes use of the much larger sample sizes that are available from a whole cohort of pupils, which may allow evaluative DIF analyses to be conducted in relation to groups which constitute much smaller proportions of the whole.

  DIF analysis can provide a useful, objective strategy to identify possible problems with particular questions. However, DIF analysis is statistically valid only if the pupils affected by the issues form a homogenous group. Pupils with special educational needs are extremely varied – there may be a greater difference between two pupils with special educational needs than between two randomly selected mainstream pupils. So the use of DIF analysis as a statistical method needs careful consideration before it is adopted. It must also be noted that different statistical approaches for computing DIF may provide different results; therefore, care must be exercised to adopt the approach that yields the most reliable outcomes.

Between them these two approaches are designed to provide a balanced set of evidence to address the question: *How does the wording and presentation of an assessment impact on pupils in different groups?*

These groups would include pupils with a range of special educational or assessment needs, such as:

- dyslexia or specific learning difficulties.
- speech, language and communication impairment.
- hearing or visual impairment.

They might also include underperforming pupils with particular backgrounds, such as:

- those with English as an additional language.
- white English working-class boys.

- children from the traveller community.

The two proposals were put forward in a Consultation in 2011, and the recommendations made here are based on the outcomes of the studies. A more detailed summary and discussion of the Consultation is available online at www.ofqual.gov.uk/files/2011-06-16-accessibility-consultation.pdf

Proposal 1: Accessibility Teacher Review Panels

Current position

An essential step in the development of all externally designed and implemented summative tests that are used for standards monitoring purposes, such as Key Stage 2 mathematics, science and English tests, is the scrutiny of the questions by teacher panels and by panels of expert reviewers. Between them these teachers and reviewers have expertise relating to a wide range of issues. The responsible body may meet with the teachers and the reviewers to discuss the draft material and come to an agreed decision on the amendments that should be made to each question. In the course of this process the responsible body or agencies appointed to write and trial the questions become aware of the sorts of issues that the reviewers focus on, so that over time question writers may become better at creating accessible questions themselves.

This question review process is already well established at the development stage for key stage tests. Furthermore, once a set of tests has gone live and has been taken by a cohort of pupils, an evaluation test review may be carried out. This:

> involves a small team of subject and assessment experts, including teachers, reviewing the test materials independently…
>
> The resulting report provides a critique of the tests, any questions raised, validity of the materials, clarity of questions for children, design of pages and artwork, cognitive demands required by the tests and so on. The reports arising from these influence the development of future years' tests. (QCDA, 2010, p.20)

Recommendation to establish Accessibility Teacher Review Panels

The evaluation test review that is conducted on each set of key stage tests is very valuable in offering an overall picture of the way in which the tests operated in practice when the pupils took them in the live context. The evaluation test review process could be extended to include panels with a particular focus on issues relating to the accessibility of the materials for pupils with a range of special educational or assessment needs.

Proposal 1: Accessibility Teacher Review Panels

Quality assurance Accessibility Teacher Review Panels with a rolling programme focusing on different subjects and papers should be established to evaluate the accessibility of National Curriculum Assessments.

This evaluation process, which could be conducted by the responsible body, involves the following steps:

1. A select group of teachers who work with pupils who have a wide range of special educational or assessment needs and who took the assessment are invited to take part in the review. These teachers are selected to represent different types of school and different geographical areas.

2. Soon after the pupils have taken the assessment (within a week or so), each of the teachers talks to one or more small groups of up to four of their own pupils. The pupils offer a representative sample of the types of special educational or assessment needs of pupils in the school and have completed either mainstream or modified versions of the tests.

3. The pupils look at a copy of the assessment paper, and talk about what they did. A short list of question prompts is given below to help teachers to structure their discussions with the pupils.

4. The teachers consider their pupils' comments, and decide whether they indicate any issues relating to the accessibility of the questions.

5. Each teacher completes the Evaluation Questionnaire for teachers found at the end of this section.

6. The teachers are then invited to a meeting with the responsible body, to report on their own pupils' experiences with the assessment.

This evaluation test review, with a particular focus on accessibility issues, would not need to be conducted every year with every assessment. One paper in one subject – for example, Mathematics Paper A or English Reading – might be selected for this process one year, and a different subject and paper in the following year.

Questions that might be used on the prompt sheet include:

- Which was the most interesting question? Why was it interesting?
- Which was the most boring question? Why was it boring?
- Were any of the questions unfair? Why were they unfair?
- Were any of the questions difficult to understand? What made them difficult to understand?
- What could the people who write the tests do to make them better?

These prompts should be used flexibly, and adapted or amended as the teacher considers appropriate to enable pupils to best express themselves in discussing their experiences in the assessment. Teachers should be invited to discuss the prompts that they found effective with their pupils in order to guide members of the Accessibility Teacher Review Panel in subsequent years.

It should be noted that pupils might find it difficult to distinguish between questions that are challenging in terms of their subject demand and those that are hard to access. So, for example, a pupil might say that two of the questions were 'difficult to understand', but in one she may have been unable to access the language in which the question was phrased, while in the other she may not have had the subject-specific cognitive skills needed to solve the problem. The first of these issues might make the question invalid because the language, rather than the test construct, caused the difficulty, but the second question would be a valid assessment of the subject even though the pupil was unable to answer the question. Teachers will need to tease out issues relating to accessibility and distinguish these from problems arising from pupils' cognitive understanding of the subject. Issues relating to accessibility may be classified according to the categories indicated in the Evaluation Questionnaire for teachers included at the end of this section.

The meeting of an Accessibility Teacher Review Panel to discuss pupils' experiences in the live tests, informed by the completed Evaluation Questionnaires, will offer highly focused qualitative evidence on the accessibility of the assessment materials for pupils with a range of special educational and assessment needs. This will provide valuable qualitative data to help the responsible body evaluate the continuing accessibility of National Curriculum Assessments over time.

Recommendations based on the Consultation survey results

- Accessibility Teacher Review Panels should be established to evaluate the accessibility of National Curriculum Assessments such as key stage tests after they have been released and taken by a cohort of pupils.
- Different subjects and papers should be evaluated in different years.
- The teachers involved should have a wide range of experiences and expertise related to pupils with special educational or assessment needs.
- The teachers should work with small groups of pupils soon after they have taken the assessments, focusing on both modified and unmodified versions of assessments.
- Following the discussion, teachers should complete the Evaluation Questionnaire for teachers.
- The teachers should be invited to a meeting with the responsible body, to share and discuss their own pupils' experiences with the assessment.
- The findings of the Accessibility Teacher Review Panel should inform development of future test questions.

Appendix A to Proposal 1: accessibility of National Curriculum Assessments – Evaluation Questionnaire for teachers

Please describe in detail any accessibility issues your pupils found in the National Curriculum assessments. Identify assessment questions that may have caused problems for your pupils.

Please complete this form electronically or on paper. You will need a separate copy for each group of four pupils with whom you discussed an assessment.

Section 1: the school[2]

Name of school:

Address of school:

Number of pupils on roll:

Type of school (please click or circle):

- Maintained school
- Voluntary school
- Community school
- Community special school
- Foundation school
- Foundation special school
- Trust school
- Maintained nursery school
- Pupil referral unit
- Grammar school
- Free school

[2] 'To find out more about types of schools, visit Department for Education: www.education.gov.uk/schools/leadership/typesofschools

- Technical Academy
- City Technology College
- Specialist school
- Boarding school
- University technical college
- Studio school
- Faith school
- Independent
- Other (please specify)

Your role (for example, class teacher, classroom assistant, Senco):

Section 2: the pupils

Date of assessment:

Date of discussion:

Subject and paper discussed:

What are the special educational or assessment needs of the four pupils in the discussion group? (Please tick or click.)[3]

Assessment need	Pupil 1	Pupil 2	Pupil 3	Pupil 4
Dyslexia or specific learning difficulties (SpLD)				
Speech, language and communication needs (SLCN)				
Hearing impaired				
Visually impaired				
Other (please explain below)				

Did any of the pupils use a modified version of the paper? (Please tick or click.)

Modified paper	Pupil 1	Pupil 2	Pupil 3	Pupil 4
Large print				
Modified large print				
Signed				
Other (please explain below)				

[3] Further guidance on these categories of assessment need may be found at www.ofqual.gov.uk/files/2011-06-15-general-and-vocational-consultation-on%20principles-for-language-modification.pdf

Section 3: the issues

Please identify any questions which raised an issue on readability of test items. Indicate which of the four pupils were affected by it.

Issue 1: long sentences

Please read about *Issue 1: long sentences* in the accompanying *Notes*.

Check the assessment for long sentences. Are they too complex for pupils to understand them easily? Could they be divided into two or more shorter, more accessible sentences?

Long sentences	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 1: long sentences*:

Issue 2: sentence difficulty

Please read about *Issue 2: sentence difficulty* in the accompanying *Notes*.

Check the assessment for short, complex sentences. Could they be made simpler, even though this might make them longer?

Sentence difficulty	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 2: sentence difficulty*:

Issue 3: passive voice constructions

Please read about *Issue 3: passive voice constructions* in the accompanying *Notes*.

Check the assessment for passive voice constructions. Could the same idea be presented as a direct, active statement instead?

Passive voice	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 3: passive voice constructions*:

Issue 4: conditional clauses

Please read about *Issue 4: conditional clauses* in the accompanying *Notes*.

Check the assessment for conditional clauses. Could they be replaced with direct sentences?

Conditional clauses	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 4: conditional clauses*:

Issue 5: relative clauses

Please read about *Issue 5: relative clauses* in the accompanying *Notes*.

Check the assessment for relative clauses. Could the information they contain be separated off from the question itself?

Relative clauses	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 5: relative clauses*:

Issue 6: cohesion and coherence problems

Please read about *Issue 6: cohesion and coherence problems* in the accompanying *Notes*.

Check the assessment for problems of cohesion. Could information be phrased more cohesively?

Cohesion and coherence	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 6: cohesion and coherence problems*:

Issue 7: legibility and layout

Please read about *Issue 7: legibility and layout* in the accompanying *Notes*.

Check the type face and point size used in the assessment. Is it as clear as possible?

Check the overall layout of the pages. Is it clear and uncluttered?

Legibility and layout	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 7: legibility and layout*:

Issue 8: prior knowledge

Please read about *Issue 8: prior knowledge* in the accompanying *Notes*.

Check the contexts in which the questions are set. Are any of them likely to be unfamiliar to certain groups of pupils?

Prior knowledge	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 8: prior knowledge*:

Issue 9: word familiarity

Please read about *Issue 9: word familiarity* in the accompanying notes.

Check the assessment for words that are likely to be unfamiliar to pupils. If they have been used, are they explained?

Word familiarity	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 9: word familiarity*:

Issue 10: use of graphics

Please read about *Issue 10: use of graphics* in the accompanying *Notes*.

Check the diagrams and pictures in the assessment. Do they help to convey information efficiently? Are there any questions which would have benefitted from a graphic?

Use of graphics	**Pupil 1**	**Pupil 2**	**Pupil 3**	**Pupil 4**
Question number(s)				

Comments on *Issue 10: use of graphics*:

Appendix B to Proposal 1: notes on the Evaluation Questionnaire for teachers

Introduction

This checklist is offered to help members of an Accessibility Teacher Review Panel to identify any problems relating to the accessibility of a National Curriculum Assessment after it has been released. It is designed to capture the good practice currently employed by the responsible body and its agencies when National Curriculum Assessments such as key stage tests are developed, and it reflects the template used to review questions during the development process. It builds on Ofqual's guidance for qualifications regulators and awarding bodies on designing inclusive qualifications, which is provided in the guidance *Fair Access by Design* (Ofqual, 2010), and it is based on the guidance given in Section 1 of this guide on ways to formulate and present accessible questions. The information and comments that the panels provide will help to inform the future development of assessment materials.

A number of features can make it difficult for pupils to understand a question. Dahlia Janan and David Wray (2011) discuss these in some detail. Problems that can arise include:

- long sentences.
- sentence difficulty.
- passive voice constructions.
- conditional clauses.
- relative clauses.
- cohesion and coherence challenges.
- legibility and layout.
- unfair advantage due to prior knowledge.
- word familiarity.
- use of graphics.

When a National Curriculum Assessment paper is reviewed to check for accessibility, each of these features should be considered in turn. Other points which are not covered in this checklist but which teachers feel are relevant should of course also be noted.

Notes on Issue 1: long sentences

Long sentences can be difficult to follow. It may be possible to divide a long sentence into two shorter ones which are easier to understand. The total number of words may be the same or even greater, but because each sentence is shorter and less complex it may be more accessible to pupils.

Example: 2010 Key Stage 2 Reading answer booklet

The instructions at the front of the answer booklet explain:

> *Some questions are followed by a short line or box.*
> *This shows that you need only write a word or phrase in your answer.*

These two shorter sentences are used instead of the longer:

> *Some questions are followed by a short line or box to show that you need only write a word or phrase in your answer.*

The one long sentence contains too much information and is too complex. The two short sentences are clearer.

Notes on Issue 2: sentence difficulty

As Janan and Wray explain, "The common belief regarding sentence structure is that the longer sentences are, the harder the text is to read" (2011, p.15). But this is not necessarily the case because even a short sentence may have a complex structure. As an example, we can examine the following two sentences:

1. *The girl standing beside the lady had a blue dress*.

2. *The girl had a blue dress and she was standing beside the lady*.

Sentence 1 is 10 words long and sentence 2 is 13 words long. Yet Reid, in her classic study of children's comprehension of syntactic features (Reid, 1972), found that 59 per cent of her sample of 7-year-olds understood sentence 1 to mean that both the girl and the lady had a blue dress. All of them, however, understood sentence 2 perfectly. Thus the longer sentence was easier to understand than the shorter.

Notes on Issue 3: passive voice constructions

Passive ('a square *was drawn*'; 'a book *was read*') rather than active ('Amy *drew* a square'; 'Yusuf *read* a book') constructions can be more difficult to understand. This is often why names are used in questions.

Example: 2010 Key Stage 2 Mathematics Test B, q5

Liam takes his dog to the clinic on Saturday.

This construction is used rather than the passive:

A dog is brought to the clinic on Saturday.

Notes on Issue 4: conditional clauses

Conditional function words such as if may be difficult for pupils to understand. Separate, direct sentences may be easier.

Example: 2010 Key Stage 2 Mathematics Test A, q12

Liam spends £14 altogether on the Big Wheel and the Rollercoaster.

Big Wheel	Rollercoaster
£2.50	£1.50

He goes on the Big Wheel twice.

How many times does he go on the Rollercoaster?

These separate direct sentences are used rather than the more complex conditional phrasing:

If Liam spends £14 altogether on the Big Wheel and the Rollercoaster, and if he goes on the Big Wheel twice, how many times does he go on the Rollercoaster?

Notes on Issue 5: relative clauses

A relative clause may give information with fewer words, but it may be more difficult to unpack than two or more sentences.

Example: 2010 Key Stage 2 Mathematics Test A, q3

This table shows six different types of cat and where they are found in the world.

Which type of cat is found only in Africa?

The description of the table is given first, and then the question itself is given separately. This avoids the use of the relative clause ‘shown in the table’:

Which of the six different types of cat shown in the table is found only in Africa?

Notes on Issue 6: cohesion and coherence problems

As Janan and Wray (2011, p. 7) explain:

> one of the key features of a text is that it is not just a group of words and sentences. Instead, there is a structure in a text which glues the various text components together…
>
> Yet problems of cohesion can easily cause difficulties for pupils reading assessment questions. The beginning of the Key Stage 2 English (2009) Reading answer booklet, for example, has the following:
>
> □ You have now had 15 minutes to read No place like home and The Earthship leaflet. In this booklet, there are different types of question for you to answer in different ways.
>
> It may well be that some pupils reading this thought initially that there should be questions for them to answer in the Earthship leaflet. The reference 'this booklet' might well be interpreted to refer to the previously mentioned leaflet, instead of the booklet the pupils are actually reading..

Notes on Issue 7: legibility and layout

Legibility and layout can have a significant effect on pupils' ability to access the questions. Large, non-serif font such as 14-point Arial can help many pupils, not just those with visual impairments, to read and understand what is being asked. Uncluttered pages, with clear diagrams and graphics, can increase access.

Notes on Issue 8: prior knowledge

The National Curriculum requires pupils to be able to use and apply their knowledge in a meaningful way. Many questions, including those both in mathematics and in English, are set in contexts. But as Janan and Wray (2011, p.16) explain:

> prior knowledge influences what is understood from text. This means that two individuals with different prior knowledge but equal in reading comprehension still would exhibit different levels of comprehension of the same text.

If the context in which a question is set is likely to be unfamiliar to certain groups of pupils then they may have greater difficulty accessing it.

Notes on Issue 9: word familiarity

Mathematical, scientific or other technical terms may be part of what the assessment is testing, in which case they should not be changed. But other words that present the context in which the question is set may be unfamiliar to pupils, and these should be replaced or explained.

Example: 2010 Key Stage 2 Mathematics Test B, q5

This table shows the opening times of a pet clinic.

In this question the phrase ‘ pet clinic’ is used, rather than the more formal, less familiar ‘veterinary surgery’.

Example: 2010 Key Stage 2 Reading booklet

On 16th August 1896 a group of prospectors located gold in the Yukon, near a town called Dawson.*

** People who search for gold are called prospectors.*

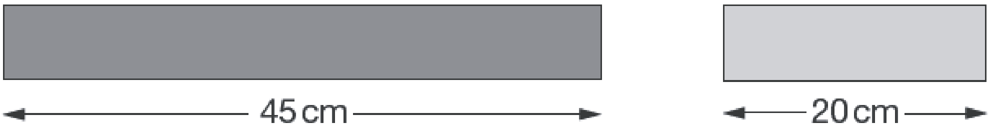
In this reading booklet article about the Yukon Gold Rush the potentially unfamiliar term ‘prospector’ is used, but it is explained in a footnote.

Notes on Issue 10: use of graphics

Graphics can be used to present information in a way that makes it more accessible to many pupils. For example, a diagram may convey key information, or a picture can help to explain the context in which the question is set.

Example: 2010 Key Stage 2 Mathematics Test A, q13

Liam has two different sizes of rectangle.



In this question key information about the rectangles is shown in a diagram.

Example: 2010 Key Stage 2 Reading booklet

The reading booklet article about the Yukon Gold Rush is illustrated with historical photographs taken at the time. These help to convey the context in which the events took place.

PHOTO REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

PHOTO REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

These are just some of the issues that you should consider when you are checking the accessibility of a National Curriculum assessment for your pupils. You may have other points that you know from experience can cause particular difficulties for some groups of pupils, and these of course should also be noted. Your observations as a practitioner with direct experience of working with pupils who have taken the assessment is invaluable, and is greatly welcomed by those who are responsible for the development of the assessments.

Appendix C to Proposal 1

Table 1: Overview of proposals for quality assurance of National Curriculum Assessments

* Quality assurance DIF analyses could not take place until the completion of marking and data collection.

	Before assessment administration	**Assessment administration**	**Within two weeks of assessment completion**	**Within three weeks of assessment completion**	**Within one month of assessment completion**	**Within three months of assessment completion**
Proposal 1: Accessibility Teacher Review Panels	Subject and paper for focus selected. Teachers recruited by research provider contracted out by responsible body.		Teachers work with small groups of pupils and complete the Evaluation Questionnaire for teachers.	Teachers meet with responsible body and/or relevant assessment development agency.	Responsible body reports on outcomes. This may lead to further discussion.	
Proposal 2: evaluative DIF analyses	Focal groups identified (for example gender, English as an Additional Language , Special Education Needs , Race/ethnicity, Free School Meals , Date of Birth).	Pupil background data collected.				* Responsible body carries out DIF analyses on selected focal groups. Responsible body reports on outcomes.

Appendix D to Proposal 1:

Accessibility Teacher Review Panels should be geographically spread across counties, metropolitan districts and boroughs. The following table lists all relevant authorities in England:

Type		Total
Two-tier 'shire' counties	Buckinghamshire, Cambridgeshire, Cumbria, Derbyshire, Devon, Dorset, East Sussex, Essex, Gloucestershire, Hampshire, Hertfordshire, Kent, Lancashire, Leicestershire, Lincolnshire, Norfolk, Northamptonshire, North Yorkshire, Nottinghamshire, Oxfordshire, Somerset, Staffordshire, Suffolk, Surrey, Warwickshire, West Sussex, Worcestershire	27
London borough	Barking and Dagenham, Barnet, Bexley, Brent, Bromley, Camden, Croydon, Ealing, Enfield, Greenwich, Hackney, Hammersmith and Fulham, Haringey, Harrow, Havering, Hillingdon, Hounslow, Islington, Kensington and Chelsea, Kingston upon Thames, Lambeth, Lewisham, Merton, Newham, Redbridge, Richmond upon Thames, Southwark ,Sutton, Tower Hamlets, Waltham Forest, Wandsworth, Westminster	32
Metropolitan district	Greater Manchester: Bolton, Bury, Manchester, Oldham, Rochdale, Salford, Stockport, Tameside, Trafford, Wigan Merseyside: Knowsley, Liverpool, Sefton, St Helens, Wirral South Yorkshire: Barnsley, Doncaster, Rotherham, Sheffield Tyne and Wear: Gateshead, Newcastle upon Tyne, North Tyneside, South Tyneside, Sunderland West Midlands: Birmingham, Coventry, Dudley, Sandwell, Solihull, Walsall, Wolverhampton	36

	West Yorkshire: Bradford, Calderdale, Kirklees, Leeds, Wakefield	
Unitary authority	Bath and North East Somerset, Bedford, Blackburn with Darwen, Blackpool, Bournemouth, Bracknell Forest, Brighton and Hove, Bristol, Central Bedfordshire, Cheshire East, Cheshire West and Chester, Cornwall, County Durham, Derby, Darlington, East Riding of Yorkshire, Halton, Hartlepool, Herefordshire, Isle of Wight, Kingston upon Hull, Leicester, Luton, Medway, Middlesbrough, Milton Keynes, North East Lincolnshire, North Lincolnshire, North Somerset, Northumberland, Nottingham, Peterborough, Plymouth, Poole, Portsmouth, Reading, Redcar and Cleveland, Rutland, Slough, Southampton, Southend-on-Sea, South Gloucestershire, Stockton-on-Tees, Stoke-on-Trent, Shropshire, Swindon, Telford and Wrekin, Thurrock, Torbay, Warrington, West Berkshire, Wiltshire, Windsor and Maidenhead, Wokingham, York	55
sui generis	City of London, Isles of Scilly	2
	Total	152

Proposal 2: Differential Item Functioning (DIF) analyses

Current position

The purpose of National Curriculum Assessments is to ascertain what pupils have achieved in relation to attainment targets (content standards) for the relevant key stage. Such measurements must be valid and reliable.

Tests are supposed to gather evidence on educational achievement, differentiating pupils with basis on their knowledge, skills and abilities. Educational achievement is an intangible and abstract concept inferred by means of assessments, such as written tests over some well-defined domain of knowledge. When educational assessment items carry DIF they will not differentiate pupils only on the '*construct*'' being assessed. That is to say, *construct irrelevant* aspects may be undermining the validity of questions.

DIF analysis is a well-established statistical procedure that is often used to identify individual questions that may differentially perform across groups of pupils with different academic and personal background variables. Such differential performance may be due to the influence of cognitive sources of construct-irrelevant variance such as unnecessary linguistic complexity of the assessment and specialised knowledge that is not related to the aim of the test. When a test requires "construct-irrelevant knowledge or skills to answer an item and the knowledge or skill is not equally distributed across groups, then the fairness of the item is diminished" (ETS, p.8).[4]

Construct irrelevant variance can occur when performance on a test is affected by sources of knowledge that are different from those that the test is intended to measure, causing test scores to be less valid for a particular group of pupils. For example, if the test is designed to assess pupils' knowledge in mathematics, but the pupils in a particular group are unable to understand the complex linguistic structure of the questions, then those questions will not be a valid assessment of their mathematical understanding. If pupils in a certain group, which is referred to as the focal group (for example, linguistic, cultural and gender), perform less well on a specific item when compared to pupils in the reference group (after controlling for the overall differences in their ability scores), then it is possible that the item is not effective as a measure of achievement for members of that particular focal group. An

[4] Educational Testing Service (2009) ETS Guidelines for Fairness Review of Assessments.p,8. www.ets.org/Media/About_ETS/pdf/overview.pdf

‘item’ can constitute a whole question, or it can form part of a longer question which has several parts and carries a number of marks.

DIF analyses compare the performance of two groups of pupils adjusted for their overall level of ability in order to disentangle the effects of construct irrelevant variance and ability level. Consistent differences between results on a particular item for two groups of pupils of the same ability level suggest that DIF may be present. However, it is important to note that results of DIF analyses can suggest only that a degree of differential performance between the two groups is present, not necessarily that the item is biased.

To establish that an item distorts achievement it is also necessary to determine the non-target constructs that lead to the between-group differences in performance. Thus, DIF only indicates the differential performance of certain test items: it does not indicate validity issue or bias.

DIF analysis first adjusts for pupils’ overall performance, and only then does it identify certain specific items as showing differential functioning across the focal and the reference group. So, to illustrate this point, two groups of pupils might be selected by their language status with English language speakers as the reference group and pupils with English as an additional language as the focal group. Both groups of pupils took a mathematics test with 50 one-mark questions, some of which had a complex linguistic structure. The mean score for the reference group – the first language English speakers – on all 50 items was 35, but for the English as an additional language group it was 25, giving a performance-gap of about 30 per cent. This clearly shows that the first language English speakers performed substantially better on the questions in this test than their English as an additional language peers. However this overall group difference does not constitute DIF. If all items present the same performance gap between the two groups then there is no DIF. But if the pupils who are identified as English as an additional language did much better on one particular item (perhaps one with a less complex linguistic structure) then that item will be identified as showing DIF, favouring pupils who are in the English as an additional language category. Similarly, if English as an additional language pupils perform much lower than their overall gap (say 50 per cent lower) on a different item then that item will be identified as showing DIF, favouring first language English speakers, the reference group.

It is important to note that items identified as showing DIF are not necessarily biased against any group. DIF simply means an item performed differently across the two groups when pupils’ overall score differences are adjusted across the focal and reference groups.

As stated, a DIF analysis compares the performance of pupils in the focal group with the performance of pupils in the reference group. It might show that Question 4, say, distorts the performance of members of a particular group in comparison to the other questions. But a DIF analysis cannot show whether the whole test is invalid for members of the affected group. So, for example, if all the members of a particular group tend to do worse than mainstream pupils on every question in the test then the mean item score of the group will be lower than that of the mainstream pupils. But the DIF analysis itself will not suggest that there are any validity issues in any particular item or in the entire test. For this reason, if a focal group has problems accessing a test as a whole – for example, because of poor reading proficiency or limited language skills – DIF analysis will not suggest that distortion might exist. A DIF analysis cannot say 'This test is biased against pupils in this group'. It can only say 'Questions X, Y and Z differentially perform across the focal and reference groups.'

The developers of the key stage tests have extensive experience of carrying out some types of DIF analysis during the development phase of their work. The outcomes of these analyses are used to identify any items which may cause differential performance for particular groups, such as boys in comparison to girls (or vice versa), or pupils with English as an additional language in comparison to first language speakers, and to guide amendments to these questions or parts of a question.

Methodological issues in DIF analyses

Literature on DIF analyses and related methodologies emphasises the importance of sample size. Effective DIF analysis relies upon there being enough pupils in the category which is the focus of the analysis to allow for statistically robust conclusions to be drawn. Petersen (1987) recommends that: "The minimum number of subjects suggested for DIF analysis is 100 subjects in the smaller group (the focal group) and the reference group should have a total of at least 500 subjects."

Many sources of distortion affect subgroups with small numbers of pupils. For example, research suggests that many pupils in low-incidence disability groups, such as those who are hearing or visually impaired, or who have learning difficulties, have difficulty with some types of test item such as those with crowded pages or complex charts and graphs (Abedi and others, 2008). But there may not be enough pupils in each of these categories to form focal groups, so DIF analyses cannot be conducted on these subgroups as focal groups.

The system is now changing to item banking and item validity trialling. But in the past, questions being developed for National Curriculum assessments such as key stage tests have traditionally undergone at least one technical pre-test, in which they were trialled by a sample of several hundred pupils. There would normally be enough pupils involved in a pre-test to allow a gender DIF analysis to be carried out because about half of them would be boys and half would be girls. There might have been enough pupils who had English as an additional language at a different level of

English proficiency to offer a viable focal group. But where only a relatively small number of pupils belonged to a particular group there might not have been enough to provide an adequate sample on which to base any meaningful statistical conclusions. So, for example, while a focal group of all pupils with English as an additional language might have been large enough, there were unlikely to be enough pupils from each specific ethnic or language group to allow a DIF analysis which distinguished between, say, Polish speakers and Chinese speakers. Similarly, there might not have been enough pupils with some specific categories of disability, such as hearing or visual impairment or cerebral palsy, to form a focal group.

It should also be noted that even the relatively small number of pupils who share one particular condition or disability, such as hearing impairment or dyslexia, may not form a homogeneous group. There may be so much variation between individuals that classing them together is not useful because it ignores their very significant differences. Putting all the pupils who have any registered special educational or assessment need into a single focal group is likely to be even less effective. There is thus a tension between the formation of statistically robust focal groups which contain at least 100 pupils, and the need to ensure that these pupils really do form a homogenous group.

For this reason, although it offers a useful objective measure of possible distortion caused by individual items in a test, DIF analysis is of limited value for the evaluation of the accessibility of an assessment for pupils who have a range of special educational and assessment needs. It should be supplemented by methods based on the collection of qualitative data as described above in Proposal 1.

Recommendation to use quality control DIF analyses for evaluation purposes

The bodies responsible for the development of National Curriculum Assessments such as key stage tests, or the agencies appointed to develop and trial the tests, routinely use the pre-test results to carry out at least gender and English as an additional language DIF analyses as part of the process of developing draft tests.[5] If the responsible body routinely collected item-level data according to a rigorous sampling methodology after the tests had been taken then DIF analyses could also be used for evaluation purposes. With sound sampling procedures, a robust data collection programme in England could provide much larger sample sizes drawn from a whole cohort taking a national test. This would allow quality control DIF analyses to be conducted with focal groups of pupils who make up much smaller proportions of

[5] One must ask whether there are enough pupils in the different categories of English as an additional language. If data sets are unsuitable for carrying DIF analyses, detailed data collection covering the different categories of English proficiency will be required.

the entire year group. Such analyses would enable the responsible body to monitor possible item impact for pupils in a wider range of categories.

It should be noted that, for evaluation purposes, a large sample containing all school types would be required. Careful sampling of school types would produce more reliable results for the purpose of identifying items that perform differently across homogeneous categories of pupils who have English as an additional language type 1, 2, 3, 4 or 5 or for pupils with different categories of special educational or assessment need.

Proposal 2: DIF analyses

Provided that a business case for the routine collection of item-level data in England is accepted, the bodies responsible for the development of National Curriculum Assessments could conduct quality control DIF analyses. The analyses would be conducted with a range of homogeneous focal groups using data from the live tests to provide statistically robust sample sizes. These analyses would be used where possible to identify construct irrelevant variance and to evaluate possible sources of DIF for a wider range of pupils.

Possible focal groups for the proposed quality control DIF analyses could include:

- gender.
- English as an additional language.
- free school meals.

In addition, there might be enough pupils in the whole cohort to allow focal groups to be formed with pupils with some specific types of special educational or assessment need, such as learning or reading disabilities. The total number of pupils with English as an additional language might be large enough to allow them to be split into groups with different levels of fluency, from ‘New to English’ through to ‘Fluent in English as an additional language’. The different categories of English as an additional language should never be collapsed; that is to say, dichotomising data should be avoided. However, adjacent categories of English as an additional language may carefully be reviewed for aggregation. Other possibilities for focal groups might be identified with further research (see ‘Appendix D: suggestions for future research’). However, it would be essential to ensure that the sample sizes were large enough in each case to allow the analysis to provide meaningful results, and that all the members of the group had enough in common to ensure the group’s coherence. The issues discussed above should be borne in mind.

For quality control DIF analyses to be carried out, one would need to collect item-level data (the number of marks awarded for each part of each question to each pupil) from the live tests. The relevant pupil background data would also be obtained during the test administration for formation of additional focal groups when relevant.

The collection of DIF data would serve a number of purposes. These include:

- guiding future item development by providing robust data relating to previously developed items.
- exploiting the much larger sample sizes available to explore a wider range focal groups.
- confirming pre-test DIF analysis outcomes.
- establishing a bank of items which show differential functioning in order to inform item writers and further research.

The DIF approach offers very useful information in the measurement of the validity of National Curriculum Assessments for pupils with a range of special educational and assessment needs. However, as discussed above, it has a number of limitations, so more focused qualitative methods should also be used to provide a balanced measure. Suggestions for these were presented in Proposal 1. For a detailed discussion of statistical models used for computing DIF, please refer to *Did It Work? Evaluating Access to National Curriculum Assessments: Research Background* (Ofqual, 2012). The research background provides examples of the application of DIF to National Curriculum assessments.

Recommendations based on the Consultation survey results

- DIF analyses should be used to evaluate the validity of national assessments, using samples drawn from the whole cohort of pupils.
- DIF analyses should be carried out with a range of focal groups, including:
    - gender.
    - English as an additional language (such groups must be subdivided into different focal groups depending on their level of fluency).
    - free school meals.
    - certain categories of disabilities.
- Further focal groups for DIF analysis may be identified, including:
    - looked-after children.

- some specific special educational needs conditions such as dyslexia.

However, a DIF analysis can be conducted only if the focal group is homogeneous and contains at least 100 members.

Suggested methodology

Logistic regression is a version of multiple regression in which the criterion variable is dichotomous. A multiple regression is an extension of a simple regression – a linear model in which one variable as the outcome or criterion variable is predicted from a single predictor variable. The simple linear model regression takes the form $Y_i =(b_0+b_1X_1)+ \varepsilon_i$ in which Y is the outcome variable and, X is the predictor, b_1 is the regression coefficient associated with the predictor and b_0 is the value of the outcome when the predictor is zero. A multiple regression model predicts the outcome by a linear combination of two or more predictors. The form of the model is $Y_i =(b_0+b_1X_1+ b_2X_2+\ldots..b_nX_n)+ \varepsilon_i$ in which the outcome is denoted as Y, and each predictor is denoted as X. Each predictor [of for example assessment performance] has a regression coefficient b_i associated with it, and b_0 is the value of the outcome when all predictors are zero.

Logistic regression supports psychometricians in their need to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. A logistic regression approach for detecting DIF has many capabilities that justify the use of this technique. Intensive efforts are required to carry out the analyses, but logistic regression is recommended (along with the more general approach of linear regression) because it can handle both dichotomous and polytomous item responses. Logistic regression can detect both uniform and non-uniform DIF, providing a wealth of information on items that other approaches may not be able to provide.

Table X below provides an example of the type of outputs that can be obtained from a logistic regression approach to DIF. This approach can be used in several steps. The first step, which is referred to as the 'base' model, provides the base information for the comparisons when you enter into the regression equation the focal and reference group membership, the group membership interaction and the total test score.

- Column 1 in Table X shows item number.
- Columns 2 and 3 reports the mean item scores for the reference and focal groups respectively. These two columns provide information on how the two groups perform similarly or differently.
- The columns labelled as chi-square and significance of chi-square for the non-uniform DIF (Columns 4 and 5 respectively) provide information on the

significance of the differences between the performance of the focal and reference group in a particular item when examining both uniform and non-uniform DIF.

- Similar information is provided for uniform DIF in Columns 6 and 7 for the uniform DIF.

- Column 8 refers to the coefficient of determination R^2 (R-square) for the uniform/non-uniform DIF and Column 9 presents similar information for the uniform DIF.

  R-square is used in the context of statistical models such as the DIF logistic regression (for binary questions) and ordinal or linear regression (for questions with more than one response). The main purpose of R-square is the determination of outcomes or item behaviour on the basis of group membership (for example, gender, English as an additional language and free school meals eligibility) and not on the basis of pupils' ability. R-square is the proportion of variability in a data set that is accounted for by the statistical model DIF. R-square provides a measure of how well outcomes of the assessment series are likely to be predicted by the model. The most informative piece of information in Table X is therefore the R-square gain due to inclusion of focal/reference group membership and the interaction of the total score and group membership (uniform and non-uniform DIF) which is presented in Column 8 of Table X. This information, along with other information, is used to classify test items according to DIF designation(Column 10, Table X) as will be elaborated below.

  Focal group is the group of interest to the investigator (by way of example, boys or girls, learners with English as additional language, those receiving free school meals or the members of an ethnic minority). Reference group is the one to which all groups are compared. The R-square will inform how much group membership (and not pupils' ability) influenced the way in which items discriminated pupils. In Column 8 one will enter both uniform and non-uniform DIF. Uniform DIF refers to cases where there is significant difference in difficulty of an item between groups with the same ability. Non-uniform DIF refers to significant difference between focal group and reference group as a result of interaction between group membership and interaction of group membership and total test scores. So R-square in Column 8 will combine the two types of DIF – uniform and non-uniform.

- Column 9 will include information on uniform DIF only; that is to say, R-square will only refer to significant difference in the difficulty of an item between a focal group and the reference group that was due to group membership and not to pupils' ability. To summarise, Column 8 combines uniform and non-uniform DIF and Column 9 shows DIF due to the focal/reference membership alone (referred to as group membership). One will need to subtract Column 8 from Column 9 and find out how much in the overall DIF projection was purely due to R-square gain or information gain due to group membership.

- Column 10 provides the most important piece of information: whether or not the item is DIF. We identify an item as DIF (refer to as 'Information 3' or simply 'I3', which is equivalent to 'C' DIF in the DIF literature) when the R-square gain due to both group membership and interaction of membership DIF is significant and is 0.130 or higher based on the recommendation by Zumbo (1999).

  If the R-square gain is less than 0.130 then we propose a label as 'I2' or 'I1' DIF. The 'I1' DIF items are those with p-values of Chi-square between 0.05 to 0.01. The 'I2' DIF are those items with p-values of Chi-square less than 0.01. While these categories 'I2' and 'I1' do not constitute DIF and do not have the same meaning in the literature, 'I2' and 'I1' are observations that can provide useful information for item writers. Therefore, the 'I2' and 'I1' DIF categorisation is for information only. Instructions on how to classify items in any of these three categories are given below.

Table X: DIF analyses by DIF variable, a generic format

Col. 1 Test item number	Col. 2 Mean of test items reference group	Col. 3 Mean of the focal group	Col. 4 Chi-square /ANOVA for the uniform/ non-uniform DIF	Col. 5 Chi-square significance for the uniform/ non-uniform DIF	Col. 6 Chi-square/ANOVA for uniform DIF	Col. 7 Chi-square significance for the uniform DIF	Col. 8 R-square for the uniform/ non-uniform DIF	Col. 9 R-square for the uniform DIF	Col. 10 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**
Item 1									
Item 2									
Item K									

Information type 1: ‘I1’ DIF items are those with p-values of Chi-square between 0.05 to 0.01.

Information type 2: ‘I2’ DIF items are those with p-values of Chi-square less than 0.01.

Information type 3: ‘I3’: the R-square gain due to both group membership and interaction of membership DIF is significant and the total score is 0.130 or higher (Column 8).

Further recommendations

- Regular post-hoc DIF analyses can assist test development yet to come.
- A responsible body should use post-hoc DIF analysis to understand why, despite careful pre-testing development, certain items display differential functioning in live tests.
- Post-hoc DIF analyses can help test developers refine items in particular topic areas.
- A responsible body should choose a particular procedure which is the most suitable to each case.
- When conducting DIF, make sure that the overall ability index is reliable, valid and unidimensional.
- Explore both uniform and non-uniform DIF.
- Use a multiple DIF procedure to cross-validate findings
- Ensure that the background variables that define each focal group create a genuinely homogeneous set of pupils.
- Items identified as showing DIF should be subject to the bias/sensitivity review process to identify the cause of the DIF and establish whether indeed the item is biased against the focal group.
- If the number of pupils with English as an additional language is large enough, and there is a specific issue to explore, then conduct DIF analyses by specific language group.

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 1 December 2011).

Janan, D. and Wray, D. (2011) Principles of Language Accessibility for Test Developers in Charge of Writing Test Items for National Curriculum Assessments. Ofqual/11/4855 (Available online at www.ofqual.gov.uk/files/2011-06-15-principles-of-language-accessibility-test-developers-in-charge-of-writing-test-items.pdf

Ofqual (2012) *Monitoring Access to National Curriculum Assessments: Research Background*. Coventry: Ofqual.

Ofqual (2012) *Guidance on the Principles of Language Accessibility in National Curriculum Assessments.* Coventry: Ofqual.

Ofqual (2012) *What Makes Accessible Questions. The Principles of Language Accessibility in National Curriculum Assessments: Research Background.* Coventry: Ofqual.

Petersen, N. S. (1987, September 25). DIF procedures for use in statistical analysis [ETS internal memorandum]. 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).

Reid, J. (1972) Children's Comprehension of Syntactic Features Found in Extension Readers. In Reid, J. (ed) *Reading: Problems and Practices*. London: Ward Lock, pp. 394–403.

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

Bolt, S. E. (2004) *Using DIF Analyses to Examine Several Commonly-held Beliefs about Testing Accommodations for Students 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 20 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 & 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.

Fair Access by Design (Ofqual, 2010)
http://www.ofqual.gov.uk/files/fair_access_by_design.pdf

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). Available online at www.cresst.org/Reports/TECH498.pdf (accessed 28 June 2006).

Michaelides, M. P. (2008) An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating. *Practical Assessment, Research and Evaluation*, European Volume 13, Number 7, September.

Petersen, N. S. (1987, September 25). DIF procedures for use in statistical analysis [ETS internal memorandum]. Princeton, NJ: Educational Testing Service.

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, Philadelphia.

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.

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.

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

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

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

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