BACKGROUND REPORT

National Reference Test Information

ofqual

NFER
National Foundation for Educational Research

Contents

1. Introduction

This document provides the background to the National Reference Test (NRT), and includes information about how the NRT was developed, the approach taken, and the way the test is administered, marked and analysed. Each year we will publish separately the results for the NRT.

Ofqual has contracted the National Foundation for Educational Research (NFER) to develop, administer and analyse the NRT in English and maths.

1.1. Overview of the NRT

The NRT was introduced to provide additional evidence to support the awarding of reformed 9 to 1 GCSEs in English language and maths. The NRT is a short test which reflects the sorts of questions in the GCSE examinations for English language and maths, and the questions largely remain the same each year. It is taken by a representative sample of students in year 11 who are taking their GCSEs in the same academic year.

Schools are selected to take part based on two variables – previous achievement in GCSEs in English language and maths, and school size. The questions in the NRT do not change from year to year so that, each year, we can compare students' results to those of students who took the test in previous years. This provides evidence of any changes over time in the performance of students in English language and maths in England. The tests focus on providing information about performance at the 7/6, 5/4 and 4/3 grade boundaries and they provide an additional source of evidence to support the awarding of GCSEs in English language and maths from 2019 onwards.

In 2015, Ofqual contracted NFER to develop the NRT and test development took place over several years. In autumn 2015, around 50 schools took part in trials of the questions to be used in the tests. In March 2016, year 11 students in over 300 schools took part in a Preliminary Reference Test to check that all aspects of the tests worked properly.

Students in 341 schools took part in the first live NRT in 2017. The outcomes of the 2017 NRT were benchmarked against the GCSE results for 16-year-olds in 2017, to establish a baseline for subsequent years.

From 2018, results from the NRT measure changes in performance compared to that 2017 baseline. NRT results will provide an additional source of evidence for awarding GCSEs in English language and maths from 2019 onwards, and that could mean that grade standards at grade 7 and/or grade 4 are adjusted to take account of changes in performance demonstrated by the NRT.

Tests take place in late February/early March each year. Each test takes an hour and is administered in schools by NFER's test administrators. Students are selected to take either an English or a maths test, and they are given one of eight different test booklets. The questions in the test booklets are designed to overlap so that each question is included in two booklets. Performance across all booklets in a subject is linked statistically. This approach provides NFER with reliable evidence on students' performance in each subject while minimising the test burden on students and schools.

After the tests are marked, the tests are first analysed to see how well they performed. The subsequent analyses link the results for the different test booklets and estimate the proficiency of all the students on a common scale for each subject. In 2017, the baseline

year, these proficiency estimates were benchmarked against the proportions of students achieving grade 7 and above, grade 5 and above, and grade 4 and above at GCSE. In subsequent years, we will compare the proportions of students at and above these proficiency levels with the baseline year, to see whether there are changes in the performance of students.

Each year we will publish a Results Digest which will include brief information on the actual samples achieved for English and maths, how representative they are, and the performance of the students on the tests. Ofqual will also publish a statement explaining how this evidence has been taken into account in the awarding of GCSEs in English language and maths.

1.2. Background to the NRT

In 2010 the Government set out its intentions to reform GCSEs and AS/A levels in the White Paper *The Importance of Teaching*.[1] Both the DfE and Ofqual then undertook a number of consultations on curriculum content and assessment arrangements, respectively.

In addition to the changes to GCSE, the Government also proposed a national test to monitor the performance of students over time, in a way that would not be affected by accountability pressures. In February 2013, the DfE consultation on secondary school accountability proposed the following:

> National standards can be tracked using different tests that are independent of the qualifications, and independent of government. There would be no incentives to reduce the rigour of these Tests. The Tests could be taken by a sample of pupils sufficiently large to make robust judgements about changes in national standards. Such Tests, in English, maths and science, would be similar to the well-known and well respected PISA, PIRLS and TIMSS tests, but would take place annually.

(DfE, 2013a)

In October 2013, the DfE response to the consultation noted that:

> The consultation asked for views about how to use and develop sample tests to track national standards at key stage 4. We sought views in particular from assessment experts on this proposal [...] They also said that the most useful purpose of such a test would be to provide independent evidence of each cohort's English and maths capabilities during year 11, to support the process of setting standards in external examinations, such as GCSEs. We have decided that this should be the primary purpose of the new sample tests. Ofqual are leading the development of sample tests for this purpose.

(DfE, 2013b)

In November 2013, significant changes to GCSEs were announced. Reformed GCSEs have new and more demanding content, in line with the revised Key Stage 4 curriculum

[1] https://www.gov.uk/government/publications/the-importance-of-teaching-the-schools-white-paper-2010

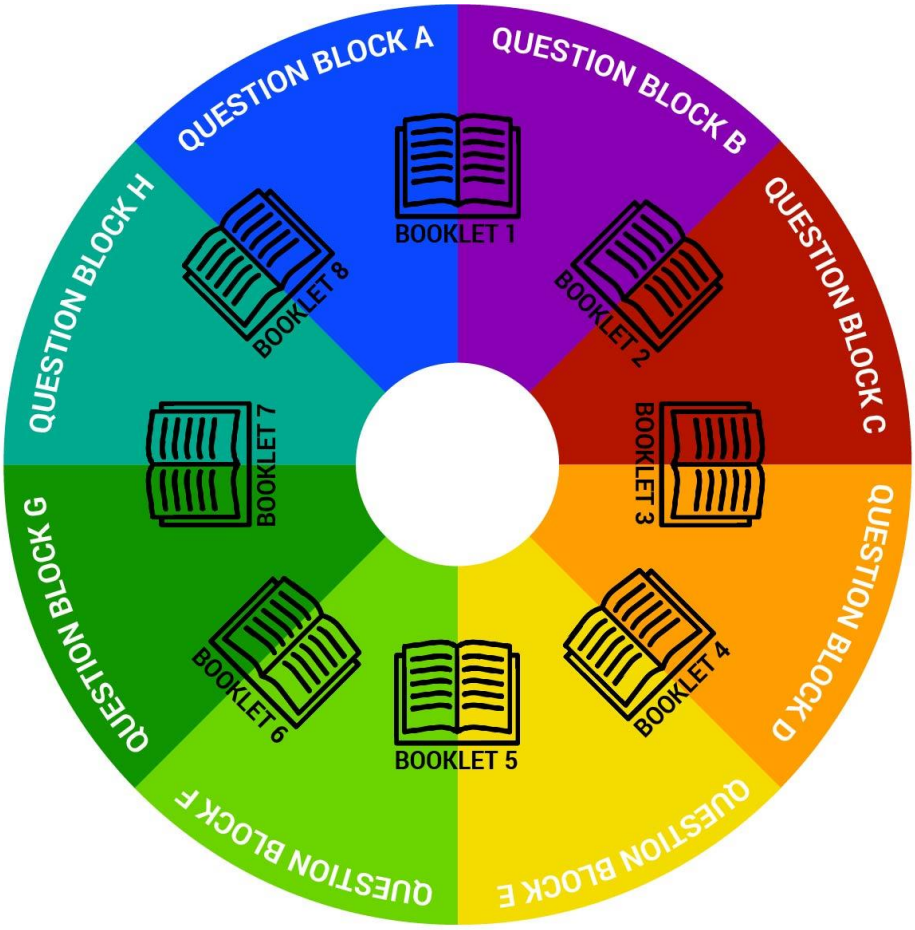
introduced by the DfE in 2014, and a new grade scale. The first reformed GCSEs were awarded in 2017 for English language, English literature and maths.

The NRT was developed to provide evidence on year-to-year and long-term changes in performance of year 11 students in England in English language and maths. It is intended to provide a national picture, and individual results are not reported to schools or students or to other organisations such as Ofsted. Test results each year will provide an additional source of evidence to be considered in awarding decisions for GCSE English language and maths each summer.

2. Test structure and development

2.1 General structure of the NRT

The NRT is made up of a series of test booklets for both the English and maths test. Test questions are grouped into blocks and each test booklet is made up of two blocks of questions. The blocks are organised in an overlapping design such that each block, and therefore each question, appears in two test booklets. This design provides a broad coverage of the curriculum and good measurement of the national population, while minimising the time each student is tested. It is the same basic structure as used in international surveys.



Each test takes an hour and follows the curriculum for the respective new GCSE. The tests are designed to be accessible for the same range of students as currently take the GCSE.

2.2 Development of the tests

NFER followed a systematic procedure involving a series of stages leading up to the first live tests in 2017.

- Test/content framework – the structure and content of the tests
- Question writing
- Field trial of tests and questions
- Preliminary Reference Test

All of these contributed to ensuring the quality of the questions used in the final tests.

Test/content framework

The content of the tests reflects the content and assessment objectives of the reformed GCSEs in English language and maths.

Question writing

The questions were developed by teams of writers who were also current and former secondary teachers with experience of teaching GCSE. All questions underwent an extensive review process during development. First, the questions were subjected to team review by other expert question writers. Some questions were revised and some questions were removed. There is more detail about the specific design of the English and maths tests in sections 2.3 and 2.4 below.

Following that initial review, the test developers conducted ‘cognitive laboratories’ to validate a selection of the questions. This involved the test developer taking a set of questions into a school and conducting a group interview with a small number of students. The students explained what they thought the questions meant, and this allowed test developers to identify any areas of ambiguity or poorly drafted questions.

Field trial

A full field trial took place over two weeks in autumn 2015. The main purpose of the field trial was to provide information about the technical functioning of the test booklets as a whole and each question, in readiness for the Preliminary Reference Test in March 2016. Eight test booklets were trialled for maths and 12 for English. These included twice the number of marks needed, to increase the likelihood of being able to select questions which performed satisfactorily for the next phase. For reading, two sets of questions were prepared for each text, which meant there were more test booklets.

Following the administration of the field trial in schools and the marking of the completed booklets, detailed statistical analyses were undertaken of each booklet and each question.

Preliminary Reference Test (2016)

The Preliminary Reference Test (PRT) was intended as a full-scale trial of the processes planned for the NRT as well as a final check on the functioning of the tests and questions before their use in the first live NRT in 2017.

The PRT was administered in schools in March 2016 by NFER's visiting test administrators, to ensure the security of the materials. Participation in the PRT was voluntary but over 300 schools took part. A total of 4,044 students in 226 schools sat the English test and 5,833 students in 312 schools sat the maths test.

The marking and analysis procedures for the PRT followed closely those to be used in the NRT. This allowed a full evaluation of the systems and the tests themselves. NFER concluded that the tests and individual questions had functioned well.

First live National Reference Test in 2017

The first full run of the National Reference Test took place in 2017. In total, over 7,000 students in 339 schools took part in the English test and a further 7,000 students in 340 schools took part in the maths test. Analysis of the functioning of both tests suggest they performed well and as expected.

2.3 The English test

The subject content for the English test includes questions on:

- literary fiction;
- literary non-fiction such as biography;
- other writing, such as essays, reviews and journalism.

The majority of the marks (all of the writing marks and 16 of the 25 reading marks) are for extended response questions. The remaining nine marks on each reading test are for short-answer responses. These are generally questions assessing straightforward understanding of information from the text or making simple inferences, although some are more challenging.

The reading blocks reflect the content domain for English language by presenting candidates with a range of genres and text types drawn from high-quality, challenging texts from the 19th, 20th and 21st centuries. Each text studied represents a substantial piece of writing, making significant demands on students in terms of content, structure and the quality of language. When selecting texts, NFER were careful to avoid bias and the use of any material that might cause offence.

In the English test, the first block is a reading test and the second is a writing test. Each block is marked out of 25 marks and students are advised to spend broadly equal time on each component.

Each reading block in the test booklets includes either one or two stimulus texts, comprising approximately 500 words. In some booklets, there are two shorter texts rather than one longer one, to allow comparison between the texts. Students are asked a number of questions that refer to the extracts. Some questions require short responses or require the student to select a response from options provided.

Reading texts were selected for the NRT based on analysis of how the questions performed at the field trial and at the PRT, as well as feedback from panels of subject experts. Other factors taken into consideration included the requirements of the test framework, and feedback about how easy it had been to mark the responses consistently.

The writing test is a single 25-mark task. Following detailed analysis and review, four prompts (questions) were selected for inclusion in the NRT. As with the reading texts, the statistical data and expert feedback were considered in making these selections.

2.4 The maths test

The maths test covers the full range of content for the reformed GCSE specifications in maths: number, algebra, ratio and proportion, geometry and measures, and statistics and probability. Students have access to the same mathematical formulae as they have in the GCSE maths exams. Unlike GCSE, which includes a non-calculator paper, calculators are allowed for all test booklets. One of the reasons for this was to make the NRT as straightforward as possible for schools and students.

Unlike GCSE, the tests are not tiered. There are several reasons for this, including practical and timing challenges. Students can change their tier of entry for GCSE right up until the day before the GCSE exam, so a student might not be certain in February/March, when the NRT is taken, whether they will sit foundation or higher tier GCSE. More importantly, the primary aim of the NRT is to measure national performance across the test as a whole, and tiering is not necessary to do that with the overlapping booklets design.

The proportion of marks allocated to each assessment objective matches that of higher tier in the GCSE papers. The distribution of marks across the mathematics content areas is targeted to be intermediate between the allocations for higher and foundation GCSE for number and algebra topics and matches the higher tier percentages for the remaining content areas. These patterns allow the NRT to derive most precision in the range of grades 4-7.

To ensure the tests assess reliably at the 7/6, 5/4 and 4/3 grade boundaries, about three-quarters of the total marks are targeted at these grade boundaries (roughly a quarter of the marks per boundary) as this is where the measurement precision is required. The remaining marks are divided between the 2/1 and the 9/8 boundaries. These items are included for accessibility with respect to the 2/1 items and to provide challenge and fuller curriculum coverage with respect to the 9/8 items.

For the NRT, maths questions are arranged into eight test booklets using the same overlapping booklet design as for English. Following the field trial, the tests were assembled using feedback from the analysis of how well the questions had performed, and feedback from subject experts. Initially eight blocks of items totalling 25 marks were constructed, so that they were reasonably comparable in terms of content areas, target grades and assessment objectives. These blocks were paired to make booklets. Hence, each item appears in two test booklets, and this overlapping design allows the entire set of questions to be linked together.

2.5 Continuing question development

If we are to measure changes in performance over time, it is essential to keep the test questions as consistent as possible, re-using these each year. While there is no routine replacement of any questions each year, it is essential to have the possibility of changing a proportion of questions for a range of reasons, as follows:

- contingency question bank in the event of a security breach

- replacement of poorly performing questions
- extending the range of curriculum coverage
- gradual shifts in curriculum and delivery

For these reasons, each year about a fifth of the students taking the tests are given a booklet which contains half ‘live’ NRT questions and half new or ‘refresh’ questions. Students do not know whether they are taking the live test or a refresh booklet, and published results of the NRT are based on the live test booklets only.

The design of the refresh booklets makes it possible to link the new questions to the existing questions, as we can see how the same students perform on both. The processes for writing and developing the refresh questions for English and maths are essentially the same as those set out above for the test booklets of the NRT.

3. The survey sample

In order to obtain a representative sample for the NRT a two-stage procedure is adopted: first a sample of schools is selected and then a random sample of students is selected from within each school.

School sample

The target sample is at least 330 responding schools per subject, with 30 students taking each of the English and maths tests (24 students taking the live test and 6 students taking refresh booklets). In total, that means just under 20,000 year 11 students would take one of the tests.

Across the eight booklets, this means that if we achieve a full sample, then 990 students would have completed each test booklet and 1,980 students would have attempted each question (since each question appears in two test booklets). In practice, the actual numbers are likely to be lower where students are not in school on the day of the test, or do not answer all the questions.

The student population is year 11 students taking English language and/or maths GCSE. Each year, all year 11 students who are registered at maintained schools, academies (including free schools) or independent schools in England and who will take their GCSEs in English language and/or maths later that academic year are eligible to take part in the NRT. Year 10 students taking GCSE a year early are not eligible and nor are older students.

Students registered at Special Schools, Studio Schools, Pupil Referral Units and FE Colleges are not part of the target population. Schools with 15 or fewer students are not included, due to the relatively small number of students.

In order to achieve maximum sampling precision, the sample is stratified on two variables: previous GCSE achievement and school size. First, schools are categorised into five strata based on previous achievement in GCSEs in English language and maths. Previous performance at school level is highly correlated with future GCSE performance. Second, each of the five strata is divided into a relatively large number of sub-strata by school size. It is within those substrata (referred to as the 'ultimate strata'), that the random sampling of schools takes place.

For each school in the original sample, up to three replacement schools are also selected. These are used when the original school has students taking GCSE but does not participate. Since the NRT is statutory for the majority of schools, this is unusual, except in exceptional circumstances. Schools who have no students taking GCSE do not take part, and they are not replaced.

Student sample

In the next stage, all schools that have agreed to take part are asked to submit details of all year 11 students who are about to take a GCSE in English language and/or maths.

In schools with 60 or more eligible students, 60 of them are randomly selected to take part (30 for English and 30 for maths[2]). Hence, 30 students would take an English test and 30 students would take a maths test. Within schools with fewer than 60 eligible students, all eligible students take part, divided randomly across the English and maths tests.

In each school, six of the students in each subject take the tests which are part of the refresh exercise (see Section 2.5). Students do not know if their test is a live or a refresh booklet. Hence for the eight live NRT booklets, each is completed by three students in each school (provided the school has full attendance). The refresh booklets are each taken by one or two students in each school.

Each year, when we publish the results of the tests, we will include full details of the achieved sample.

4. The testing process

In order to ensure the security of the NRT, so that items and test booklets may be used in future years and to minimise the administrative burden on schools, the NRT is administered by trained staff visiting schools for that specific purpose. The test administrators take all the test materials with them and take them away after the test sessions.

All test administrators are current or recently retired teachers with experience of the classroom environment and have enhanced DBS checks. Most have previous experience of the role and are used to supporting schools in preparation for tests. There are around 180 administrators in total each year. All receive training shortly before the administration of the NRT in late February and early March.

The role of the administrator is to administer the tests consistently, in accordance with the instructions, so that the evidence derived from the NRT can be relied on. This involves making arrangements with the schools allocated to them, establishing a professional relationship with the relevant staff, issuing test booklets to students, ensuring that every aspect of the administration is carried out meticulously and in line with the instructions provided, dealing with queries and problems and ensuring that all materials are completed and returned intact.

The administrators are responsible for the security and confidentiality of all of the testing materials from the time they receive them to the time they return them. They are also

[2] Schools with entries for only one of GCSE English language or GCSE maths will only take part in the test in that subject, so only 30 students will be selected.

responsible for ensuring they are available to receive the materials, have a suitable place to store materials before and after testing, have a suitable method of transport to and from schools, and the arrangement of the return of materials. These roles and responsibilities are communicated to administrators in their handbook and reiterated in the training session.

NFER monitors the quality of the test administration in approximately 10 per cent of the schools taking part. This involves an experienced administrator or a member of NFER staff member observing the test administration and providing feedback about the quality of the administration. Separately, Ofqual may also make school visits to observe the test administration.

Access arrangements similar to those used at GCSE are used in the NRT, so if students are going to be supported in their GCSEs by a scribe or reader, or given any other support, the school can use this approach in the NRT. Braille or modified enlarged print versions of the test are available for students who require this. Enlarged tests and coloured paper formats are also supplied if required.

The access arrangements for the NRT also make provision for students who would normally use a word processor to type their answers during exams. For the GCSE exams, students view a paper version of the test and are allowed to type their answers directly into a Word document. To maintain security of the NRT, students use an online tool to enter their responses to the test questions. Students access the test via a specified web-link launched from a web browser. Students are also provided with the paper version of the test and are allowed to complete it on paper, or type their answers into the online tool, or a combination of both methods.

5. The marking process

The processes for marking have been designed to mirror those used for GCSE marking and build on processes used extensively within NFER for marking of national curriculum tests, international surveys and other assessment projects.

Scripts are scanned into the marking system and students' responses are marked onscreen by trained markers and the marks are also captured electronically.

As soon as booklets have been scanned, script images and data are transferred to the onscreen marking system (OMS). Questions are broken down into item groups and individual items before being allocated to markers.

There are three separate marking teams, for reading, writing and maths. Each is made up of qualified teachers, who have experience of teaching and marking at GCSE level in the relevant subject. The teams are also balanced in terms of experience of the different exam boards. Each of the three marking teams is led by an experienced lead marker supported by a group of team leaders. The lead marker and team leaders are collectively referred to as the lead marking team. The lead marker for each subject is responsible for ensuring consistent marking across their marking team, including the consistency of judgements made by the senior markers. Each team leader monitors the work of around ten markers. Marking teams for maths and reading are structured to focus on specific questions and individual markers mark clusters of questions across multiple scripts to maximise consistency and quality of marking. Writing markers mark complete scripts.

Markers are trained in consistent application of the mark schemes, access and use of the onscreen marking system (OMS), escalation of queries, security requirements and quality expectations and control methods. The marker training also highlights the important elements of the mark scheme.

Following their training, markers are required to complete a standardisation exercise. This involves marking a small number of common responses and submitting their marks to their senior marker who checks that they are within the specified tolerance for the question. Markers whose mark exceeds the acceptable tolerance margin for a question receive further training and are subject to further monitoring. Markers cannot proceed to live marking until they have satisfied the marking team that they can perform acceptably.

The marker training and associated exercises ensure that markers are aligned in their understanding and interpretation of the mark scheme and they can apply it consistently to the required standard.

Double marking is undertaken for all reading extended responses and all writing responses. Each marker has no knowledge of the previously given mark, in order to prevent this affecting the outcome. If the two allocated marks are within an agreed tolerance then the mean (average) mark (rounded up to the nearest whole number) is used as the final score for the response. Where the difference between the marks awarded by the first and second marker is outside the agreed tolerance, a third marker marks the response. The third marker is unaware of any previous marks awarded. A ‘true’ mark is then calculated for the response based on an average of the third marker’s mark and the closer mark from the first or second marker (rounded up to the nearest whole number).

During the marking process, student responses that have received an agreed mark from the group of expert markers are used at random points. These ‘seed’ responses are given to all markers, who have no means of identifying them from the broad mass of responses. The rate at which these ‘seeds’ are allocated to markers and the associated tolerance values are set in advance for each question across both subjects. Markers who deviate from the agreed marks for these seed responses (beyond the agreed tolerances) are either given further training or do not receive their full marking allocation. Where necessary, markers will be stopped from marking and their allocation is redistributed to other markers.

There are many factors which contribute to the quality of the marking. These include:

- development of clear and concise mark schemes which have been previously trialled
- selection of experienced and capable markers and a team structure that allows close supervision, with team leaders reviewing batches of markers’ marking
- comprehensive training of the markers immediately before marking begins
- a standardisation procedure to check each marker’s application of the mark scheme, before they begin marking
- seeding of responses into the markers’ allocation as a check on their ongoing consistency and performance
- for English, double marking of some items that require greater levels of academic judgement.

Some of these processes lead to quantifiable outcomes. Each of these is reviewed as part of each year’s NRT process.

6. Analysis

The key outcome of the NRT is an annual estimate of any change in the performance of 16-year-old students in English and maths at key grade boundaries: 7/6, 5/4 and 4/3. This is reported as an estimate of the proportion of students (percentiles) that would achieve these grades (and above) in the respective GCSEs. For English, each test booklet consists of a reading and a writing component, but for analysis, these are combined into a single measure for English.

For each of English and maths, the test consists of eight booklets. These are randomly assigned to students within participating schools to minimise any systematic differences in the groups of students sitting each of the test booklets thereby enhancing the precision of the estimation of performance at the national level.

The first step in the analysis uses Classical Test Theory (CTT) techniques to examine overall test quality and the performance of the items. At test booklet level, the average score (mark) along with the spread of scores achieved ('standard deviation') are calculated, as well as indicators of test reliability such as Cronbach's coefficient alpha, which measures the degree to which the questions of a booklet all measure the same construct. Indicators of item functioning such as facility rates (indicating on a scale of 0 to 1 how easy a question is) and discrimination indices (indicating on a scale of –1 to +1 how well a question differentiates between high- and low-ability students) are also calculated. For questions that are multiple marked, measures indicating the extent of mark agreement between markers for the same responses are computed. Information about missing data is examined, for instance, relating to the impact of time pressure. Basic checks on data quality are also carried out.

Then the data from the eight test booklets are combined and subjected to Item Response Theory (IRT) analysis. IRT scaling has been the approach of choice for many decades in other tests with a similar overlapping booklet structure and purpose, such as PISA. An IRT analysis creates a scale of latent trait on which test takers and test items can be placed. In educational assessment, the latent trait can be thought of as subject attainment. A test taker's position on the scale indicates their level of subject attainment ('proficiency' or 'ability') and a test question's position on the scale indicates the level of subject attainment required to answer it successfully ('difficulty'). In IRT, scale creation relies on the assumption of a model which postulates a mathematical relationship between a student's score on a question, on the one hand, and their proficiency level and the question's difficulty level, on the other. Under the constraint of an assumed model, the IRT analysis infers the level of proficiency/difficulty each student/question must have for the observed pattern of students' scores on individual questions to occur. The IRT analysis of the NRT test data creates a scale on which all test questions (regardless of the groups of participants who have answered them) and all test participants (regardless of the booklets they have sat) can be placed and compared and on which the grade boundaries targeted by the NRT can be specified.

From 2018 onwards, IRT analyses are also used to link results between years. The approach used is concurrent item parameter calibration and is the most efficient method to construct a common scale. In this method, item parameters for each question are

estimated, using the data for all the years of the NRT, and this enables the final outcomes to be estimated more precisely.

Before that analysis takes place, there is an investigation of any differential item functioning (DIF) comparing the current year with all the previous years, to judge whether any item functions differently in the current year compared to previous years. This process identifies questions which have apparently become easier or more difficult over time, relative to the rest of the test. Questions which show DIF are identified statistically, and are then subject to a formal review by subject experts to establish if there is an identifiable reason for the DIF. Such reasons may be physical such as misprints in the test booklet, or a change in the marking of a question, or changes in the environment, such as the question being very similar to a GCSE question. If such a reason is reliably identified and supported by evidence, the question is not used in the link between years. Where questions are removed from the link, this is reported in the results report, along with the outcomes of the detailed analyses described above. Any question removed from the link is still included in the estimation of student proficiency in the respective years.

Linking test results between years means that test participants from different years can be placed and compared on the same proficiency scale. The percentages of students in each year's national cohort reaching the same proficiency thresholds represented by the percentages of students in 2017 achieving at the key grade boundaries can be estimated and compared. These estimated percentages will be reported along with a measure of precision, which is half of the 95% confidence intervals around these estimates (a 95% confidence interval is, loosely speaking, the range of values that contains the true value being estimated, with a 95% probability). The reported percentages in previous years may change very slightly from one year's estimate to the next, as a result of using the concurrent item analysis, which each year will include an additional year's data. However, these changes are likely to be very small and are unlikely to affect the comparisons between years.

The precision measure combines statistical uncertainty arising from the fact that NRT only tests a sample of schools and students, and uncertainty in estimating student proficiency in the IRT analysis. The target for the NRT is to achieve a 95% confidence interval of plus or minus not more than 1.5 percentage points from the estimate at each proficiency threshold. NFER will report changes between years as statistically significant if the probability of them occurring by chance is less than 5%. They will also report whether any of those changes were also statistically significant at the 1% level (ie if the probability of them occurring by chance is less than 1%).

A key question arising from the NRT results in a given year is whether there is a statistically significant difference from the results in previous years. There are various ways of estimating this, such as a two-sample t-statistic. However, since there are three grade comparisons across multiple years, there are a number of comparisons that can be made. As the number of simultaneous comparisons grows, the probability increases that some of them are statistically significant by chance. To ensure that the chosen level of significance is achieved overall, NFER have implemented an adjustment for multiple comparisons. Using this method, NFER will advise on the outcomes of the multiple comparisons at the chosen levels of statistical significance each year. It is possible that the data may show that there is a statistically significant change over time ie since the first NRT results in the 2017 baseline year but the change is not statistically significant when comparing with the previous year.

7. References

Department for Education (2013a). *Secondary School Accountability Consultation* [online]. Available: https://www.gov.uk/government/consultations/secondary-school-accountability-consultation [15 May, 2018].

Department for Education (2013b). *Reforming the Accountability System for Secondary Schools. Government Response to the February to May 2013 Consultation on Secondary School Accountability* [online]. Available: https://www.gov.uk/government/consultations/secondary-school-accountability-consultation [15 May, 2018].

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

August 2023

Ofqual/19/6527/3