Ofqual Board

Paper 24/16

Date:
27 July, 2016

Title:
National Reference Test

Report by:
Dennis Opposs, Standards Chair
Tim Leslie, Director, National Reference Test Project

Responsible Directors:
Michelle Meadows, Executive Director, Strategy, Risk and Research
Marc Baker, Chief Operating Officer

Paper for decision

open paper

Issue

1. We have developed the new National Reference Test (NRT) to provide evidence for Ofqual on changes in performance standards over time in GCSE English language and mathematics in England at the end of Key Stage 4.

2. The development will be complete by the end of August. In March we held the Preliminary Reference Test (PRT), an operational trial to demonstrate whether the test, as designed, could provide the evidence on changes in performance standards. We formed a Sub-Group of the Standards Advisory Group who have reviewed the test's design and the results from the trial. They concluded that the quality of the National Reference Test is sufficient and that following further development work on issues raised by the Sub-Group, the tests have the potential to be used in GCSE awarding processes.

3. We now need to consider whether we proceed to introduce the NRT and to conduct the first annual test in schools in 2017.

Recommendation

4. The Board is recommended to approve that Ofqual introduces the National Reference Test and to conduct testing in schools for the next four years. The first tests will be held in schools in 2017.

5. In making this recommendation, the Board is asked to note that:

   a. the Board will receive an annual report each summer, setting out how the test performed and how results were used in GCSE awarding. This will provide an opportunity for the Board to consider whether any changes to the test or how results are used should be made. Regular updates will also be included in the COO's report to each Board meeting;

   b. detailed features of the test's design and how results will be used will continue to evolve, which will address the issues raised by the NRT Sub-Group of SAG. The Sub-Group will remain in place to provide advice on the design and use of the NRT.

6. The Board is also asked to confirm that the COO is authorised to approve variations to Ofqual's agreement with NFER subject to:

   a. Ofqual being able to meet any resulting changes in NFER's charges from the Programme funds provided by DfE;

   b. the contract variations not exposing Ofqual to materially greater risk.

Background and purpose of the National Reference Test

7. Over the last few years when the comparable outcomes approach to awarding has been used, GCSE and A level results have not increased markedly from year to year and we have moved away from the dumbing-down debate that was characteristic of most summers in the 1990s and 2000s. However, that debate has been replaced by secondary schools, for which GCSE results are a key accountability measure, calling for the removal of what they see as a cap on grades on the basis that they believe that the attainment of 16 year olds has been rising.

8. A new element in this picture is that we also have Government expectations that as a result of its reforms, as students take the new GCSEs, attainment will begin to rise. For example, the introduction of the Progress 8 accountability measure encourages schools to raise the performance of all their students. Additionally, the Government has raised a 'good pass' from a grade 4 to a grade 5.

9. There is no solid evidence available from GCSEs (or elsewhere) of any recent rise in attainment amongst 16 year olds in England. The idea of the NRT was to provide a key source of evidence for the independent regulator on changes in attainment at the end of Key Stage 4 which

could be used when awarding GCSEs. Without that source of evidence, it is difficult to engage meaningfully in a debate about whether attainment over time is rising, falling or remaining static.

10. The Board was originally advised in January 2014 (paper 103/13 Setting GCSE and A level Grade Standards), of the rationale for the introduction of a reference test (attached as Annex A). That rationale still holds today.

11. The stated purpose does not include monitoring of students' performance in specific areas of the curriculum. For example, the tests would not reliably tell us about how the performance of England's 16 year olds is changing year on year in using language creatively in writing or using algebra to create proofs. Nor will the tests reliably tell us how the same students are performing in, for example, science or history.

The Board's previous consideration of the NRT

12. The Board was advised at its meeting in July 2013 that DfE had consulted on the introduction of a national sample test as part of its proposals for changes to the school accountability measures at Key Stage 4. Following discussions between Ofqual and DfE, ministers had indicated that Ofqual should own a more focused reference test to help set and maintain standards for new GCSEs and that funding would be provided for this work.

13. This was confirmed in a letter on 19 July 2013, from a DfE director to the Chief Regulator:

"Ministers have now given their policy steer on the purpose and ownership of the tests. They agree with Ofqual that the sampling tests should be used as a reference test to help set and maintain standards for new GCSEs."

14. This was also confirmed by DfE in its response to the consultation, published in October 2013:

"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 agreed that a sample test should be introduced at Key stage 4. They also said that the most useful purpose of such a test would be to provide independent evidence of each cohort's English and mathematics 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."

15. In April 2014, Ofqual included outline proposals for a reference test in its consultation on Setting Grade Standards for new GCSEs in England. The Board noted the outcome of the consultation at its meeting on 19 August 2014 and were advised that more detailed proposals had been included in a draft Invitation to Tender sent to prospective NRT suppliers. The Board paper also confirmed that, in line with the DfE's decision, the purpose of the NRT would be to provide evidence for Ofqual on changes in performance standards over time in GCSE English language and mathematics in England at the end of Key Stage 4.

16. The Board also confirmed that the NRT should only provide evidence on changes in the performance of the national cohort, not to provide information about changes in performance for individual schools, or by school type, geographical region or exam board entry. An individual student's performance in the NRT would not be used when awarding the GCSE to that student.

17. In December 2014 (Board paper 68/14), the Board approved that Ofqual should invest in the development of the NRT for the purpose stated in the Information to Tender, issued on 26 September 2014, subject to the outcome of the evaluation of responses received from potential suppliers. The Board approved the award of the contract to NFER at its meeting on 27 February 2015 (Board paper 103/14). That approval included a full-scale operational trial of the developed test in March 2016 (the PRT), based on field trials from October 2015, to inform Ofqual's decision whether to proceed with live testing in 2017.

The Preliminary Reference Test

18. Annex B provides the Board with a summary of the design of the NRT, how it has been developed and an assessment of the PRT, the trial we held in over 300 schools this March.

19. Recruiting schools for the PRT had been challenging, reflecting the voluntary nature of the trial. NFER had approached over 2000 schools, of which 326 finally agreed to take part in the test. NFER's administration of the test in schools went well. Student drop-out on the day of the test at around 20% was about twice as high as had been expected. Once GCSE results are known in the summer, we shall check the extent of the bias this has created in the results. Looking ahead, we will want to repeat such checks annually and monitor whether this bias changes over time.

20. The quality of NFER's marking was also acceptable but we do consider there is room for further improvement. The psychometric analysis used to link test versions worked satisfactorily but we will consider options to make the approach more straightforward if this does not significantly lessen the precision of the results. Overall, both the maths and English tests functioned well, discriminating students' abilities and all items in

each test contributed to the overall measure. The tests are ready to be used without replacing any item.

Advice received from the NRT Sub-Group of the Standards Advisory Group

21. To support the introduction and use of the NRT, we formed a Sub-Group of the Standards Advisory Group (SAG) to provide advice to the Board. The Sub-Group, chaired by Mike Cresswell, met on 20th June and the minutes of its meeting are included at Annex C. SAG endorsed the Sub-Group's advice at its meeting on 1st July. (Minutes included in Board paper 32/16.)

22. The SAG NRT Sub-Group concluded that the quality of the National Reference Tests is sufficient and that following further development work on the range of issues noted in the minutes, the tests have the potential to be used in GCSE awarding processes.

Using NRT outcomes in GCSE awarding

23. We have discussed with the exam boards how NRT outcomes might be used in GCSE awarding. They see no technical reason why that aim could not be achieved. Nevertheless, exam boards' views continue to differ on both the principle of the NRT and the effectiveness of using its results when awarding GCSEs.

24. Our working assumption is that comparable outcomes is taken as the default position and that NRT results would be used to confirm whether there have been any significant changes in attainment. If there were, that might imply that awarding using comparable outcomes alone would be unfair, and the NRT outcomes would then provide a means of adjusting GCSE results.

25. There are different ways in which this task could be achieved. The most likely approach involves using KS2-based statistical predictions as now in the awards but adjusting the predictions in advance using decisions based on NRT outcomes.

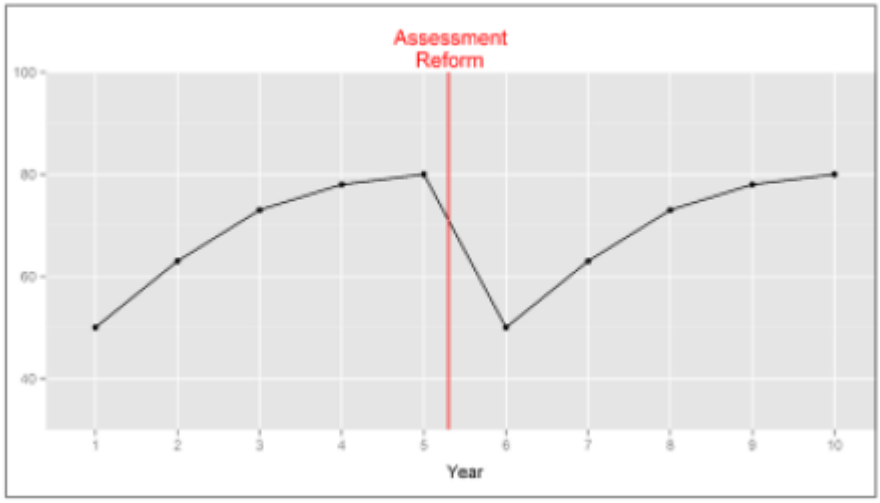
26. Before any adjustment could be made to the predictions, there would have to be careful consideration of what the NRT outcomes meant and so what changes, if any, are most appropriate. We are assuming that the decision to make an adjustment would be one for Ofqual to take, the exam boards then being required to implement it.

27. A high level proposal regarding how we could approach determining the application of the NRT outcomes to GCSE awards is attached at Annex D. Governance arrangements would have to be developed and documented.

28. The 2017 NRT is the first year of full testing so the outcomes cannot be used in GCSE awards – there is no previous set of outcomes to use as a comparison to see whether there has been any change over time.

2018 would provide the first opportunity for action to be taken. However, there would be just two years of data available at that time and the advice from SAG is that we ought not to use those NRT outcomes to affect GCSE results.

29. That advice is based partly on the need to pay careful attention to the Sawtooth effect. This is a pattern observed in the US where outcomes for high-stakes assessments dip suddenly when that assessment undergoes reform, followed by a period of relatively rapid improvements as students and teachers gain familiarity with the new assessment.



30. Our recent research based on GCSEs and A levels suggests that a similar pattern is evident in our exams. The implication of this – which relates to the justification for using the comparable outcomes approach in the first year of new exams – is that an apparent rise in performance between the first and second years on a new exam may represent more an increased familiarity with the new syllabus and its assessments rather than a genuine rise in attainment. GCSE awarding in 2018 (and perhaps beyond) will therefore have to factor in the Sawtooth effect on both the GCSEs and the NRT.

31. We are confident at present that NRT outcomes can properly be applied to GCSE English language and mathematics, the two subjects assessed in the NRT. It is less clear how those outcomes might be applied to other GCSE subjects. They would not provide the same sort of direct evidence and yet may be indicators of changes of a construct sometimes called “general academic application “for the same 16 year olds. Exam boards will need to be mindful of the NRT results when awarding other GCSE subjects. Exactly how we will operationalise this will require further consideration.

32. Expanding the subjects covered by the NRT may not be feasible as the costs for Ofqual and the burden on schools could be too great. Parallel with the work on the NRT, we are, though, continuing to explore with

the exam boards and others ways in which GCSE awarding may be strengthened so that it may be able to take more account of evidence other than statistical predictions based on KS2 prior attainment. That work involves both seeing how awarders' judgements of the quality of candidates' work may provide more reliable evidence and comparing our awarding methods against alternatives used in other countries.

Value for money of the NRT and its burden on schools

33. In proposing to introduce the NRT, we have considered whether this would represent an appropriate use of public funds and that the burden on schools and exam boards is appropriate in the context of the intended purpose for the NRT. It is not feasible to estimate direct financial benefits but the following assessment provides our justification to introduce the NRT.

34. Each year, about 600,000 students in year 11 take GCSEs in maths and English language, virtually the entire cohort. We have been careful to keep the sample size we require for the NRT to the minimum that can provide results that will be sufficiently precise to be able to be used in GCSE awarding. We will ask about 30 students in each of 300 schools to take the maths test, and another 30 in each school to take the English test – a total of about 18,000 students, 3% of the national cohort. NFER will try to avoid asking the same schools to take part in the test in two consecutive years, although to maintain the representativeness of the sample, this may not always be possible. The test will last an hour so students will be out of their lessons for about an hour and a half. Schools will be able to agree with NFER the day and time when the test will take place within the two-week test window. We consider that this burden on schools and students is appropriate.

35. We might compare the additional cost that the NRT will introduce to the 'exam system' with what is already incurred. We do not have information on the costs that exam boards and schools incur from students taking GCSEs. So to illustrate the scale of the change, we can use GCSE entry fees as an indicator for current costs. (Clearly, this does not include costs that schools incur in teaching, or in exam preparation.) Exam boards charge about £30-35 for each student that a school enters for a GCSE. The total year 11 entries for GCSE maths and English language are about 1.2m, resulting in entry fees of about £36-42m pa. The annual cost for the NRT, including Ofqual's own costs as well as those for our test supplier are about £2m pa, in the region of about 5% of entry fees, and about £3.30 per student.

36. The illustrations in the previous two paragraphs relate to a 'steady state', once NRT results are taken into account in each year's GCSE awarding. This will not occur in 2017, when GCSE results will be used to establish the baseline performance in the NRT. It is also unlikely that we will consider the NRT results will lead to any adjustments in GCSE outcomes in 2018, the first time that we will potentially be able

to compare performance between years. The NRT costs in 2017 and 2018 may more appropriately be thought of as a continuing investment to establish the NRT for the longer term.

37. The NRT adds to the cost of Ofqual's role in maintaining standards and public confidence. DfE has agreed to provide additional funds for this purpose and the introduction of the NRT is not diverting Ofqual's other resources from commitments we have set out in our corporate plan.

38. The benefits that arise from the NRT are to enhance public confidence in the grades that exam boards award and to have evidence to justify whether performance in the GCSEs is changing over time. GCSEs in maths and English language are the two most important subjects taken at GCSE.

39. We will keep the value of the NRT under close review and will report each summer to the Board. We shall assess the quality of the evidence obtained from the NRT and how useful this has been in maintaining standards in GCSE awarding.

    a. In summer 2017, following the first live NRT, we will confirm how the test performed operationally and technically. This will take into account the impact of enhancements suggested by the NRT Sub-Group, the willingness of schools and students to take part in the test, and more detailed proposals for the use of NRT results in GCSE awarding (including how the Sawtooth effect will be taken into account).

    b. In summer 2018, following the first GCSE awarding where NRT results will have been considered. At this time, we will also need to decide whether Ofqual should start the procurement of a replacement test supplier, which might succeed NFER at the end of the initial term of the current contract in autumn 2020. Alternatively, Ofqual could exercise the option to extend the current contract with NFER for an additional two years to 2022.

    c. In summer 2019 it is more likely that NRT results will have played a part in GCSE awarding. If the current contract with NFER has not been extended, Ofqual will be at the point of awarding a new contract so will need to decide whether to continue with the NRT.

Implementing the National Reference Test

40. The tests in both maths and English language are now ready to use. We therefore recommend that Ofqual proceeds to implement the NRT and that the Board agrees to the NRT being held each year for the next four years. This period aligns with our existing contract with NFER and also with DfE's commitment to provide funding for the current CSR period.

41. Detailed features of the test's design and how results will be used will continue to evolve, which will address the issues raised by the NRT Sub-Group of SAG.

42. We will hold the first annual live tests in schools in 2017. We will ask a minimum of 300 schools to take part, the same number that we had planned for the PRT. At each school, for each subject, 24 students will take a live test and another 6 will be asked to trial questions that are being developed for use in future years. (Students will not know whether they are taking live questions or those being developed.)

43. NFER will start to recruit schools to take part in the 2017 test in September. We will support the recruitment through a programme of communications and stakeholder engagement. The window for testing in schools has been brought forward slightly from the PRT and will take place between 20 February and 3 March 2017. We require NFER to report the outcome of the test to Ofqual in May and having now trialled processes in the PRT, a little more time needs to be made available for the marking, analysis and reporting.

44. DfE has introduced legislation that makes it mandatory for most state schools to take part in the test[1]. The legislation does not apply to a small number of schools that became academies before 2010, or to independent schools. However, schools from these groups will be asked to take part, albeit on a voluntary basis. The legislation also gives school heads the authority to withdraw individual students from taking the test but guidance from NFER will explain that this discretion should be used only in exceptional circumstances.

Stakeholders and communications

45. Our priority is to establish the NRT as a significant element of the annual approach to setting and maintaining standards at GCSE. We have already discussed our plans with many stakeholders before and during the NRT's development. We have taken a cautious line, not wanting to be over-optimistic about how well the NRT would perform before we had the results from the PRT. We can now be more confident about the test's design. However, it will be essential that the schools and students that are included in the sample do take part. This must be a key message. The test will only be successful if schools play their full part in it.

46. We propose to announce our decision to proceed with the NRT early in September, as soon as the schools return from the summer holiday and after the GCSE results have been announced. NFER will start to recruit schools for testing in 2017 later that month. Prior to the announcement, we will brief key stakeholders and ask them to

[1] The Education (National Curriculum) (Key Stage 4 Assessment Arrangements) (England) Order 2016 http://www.legislation.gov.uk/uksi/2016/476/contents/made

contribute in making information about the NRT widely available. Our announcement will:

- state the purpose of the NRT and explain it will help to identify if there is a change in the performance over time nationally in the GCSE, and therefore to reflect this in the grade that each student is awarded;
- confirm that the test design is suitable for this purpose, using evidence from the PRT;
- explain why it is essential that schools that are selected to take part in the test agree to do so, to achieve results that are nationally representative and statistically valid;
- also explain that, although results in 2017 will not be used for awarding GCSEs in that year, they will provide the baseline from which changes in performance in future years can be compared;
- make clear that this is no longer a research project. The NRT is being introduced as an annual sample test and most schools are required through new legislation to take part if selected;
- provide context for using NRT results in GCSE awarding. They will be used together other sources of information that are already used as well as the judgement of exam boards' senior examiners. We will act cautiously and it may not be before 2019 that the NRT outcomes first have an impact on GCSE grade standards.

Finance and Resources

47. The Board has already approved Ofqual entering into a contract with NFER for the design, development and delivery of the NRT. This includes four annual live testing cycles (2017 to 2020) and the option to extend for a further two years. The total contract value is £13.04m. The contract is based on the Government's Model Services Contract.

48. Following the Government's Spending Review 2015, DfE confirmed that Ofqual's settlement for financial years 2016/17 to 2019/20 includes programme funds to enable Ofqual to deliver the NRT.

49. When the Board approved Ofqual entering into a contract with NFER, we did not set out how variations to the contract should be approved. Variations are likely from time to time. The Board is therefore asked to confirm that the COO is authorised to approve variations to Ofqual's agreement with NFER subject to:

    a. Ofqual being able to meet any resulting changes in NFER's charges from the Programme funds provided by DfE;

    b. the contract variations not exposing Ofqual to materially greater risk.

50. Variations that either could not be funded from the allocated Programme funds or are considered to increase risk materially to Ofqual will be brought to the Board for consideration.

Impact Assessments

Equality Analysis

51. The NRT is not a qualification. It is not assessing the performance of each individual student, nor will it award a result to each individual. Nevertheless, we have designed the NRT to be accessible to most students with disabilities. NFER does not exclude students with disabilities from the sample of those who are asked to take the test. NFER also provides similar arrangements to those that are available to students where they take their GCSE including additional time, enlarged papers, a version in braille, the use of readers and scribes, and allowing the use of word processors to respond.

52. The legislation making it mandatory for schools to take part gives head teachers the discretion to withdraw individual students from taking the test. If NFER cannot meet the needs of a particular student, the head teacher can withdraw the student and therefore the student would not face responding to a test using inappropriate methods.

53. During development of the test, NFER held a workshop with experts in various areas of disability and school exam officers with particular experience in special educational needs. Their advice was helpful to ensure test questions were accessible to such students and that test arrangements were consistent with those available for GCSEs.

54. NFER monitored and reported how access arrangements were used in the PRT. They are required to do this each year and to report their findings to Ofqual.

Risk Assessment

55. This paragraph has been redacted as its publication would be prejudicial to the effective conduct of public affairs.

Regulatory Impact Assessment

56. We have discussed with exam boards how results from the NRT can be used in GCSE awarding. More detailed discussions are planned but the initial conclusion is that the NRT results could be applied quite straightforwardly in the current pre-award process by taking them into account in the predication matrices that are based on key stage 2 results. This should not increase the regulatory burden on the exam boards. They will be required to engage in a review of each year's NRT results in May and early June, prior to Ofqual deciding whether GCSE performance has changed.

Paper to be published	Yes
Publication date (if relevant)	Not before when Ofqual announces its decision, planned for early September

	2016
If it is proposed not to publish the paper or to not publish in full please outline the reasons why with reference to the exemptions available under the Freedom of Information Act (FOIA), please include references to specific paragraphs	Annex C is closed because it relates to the development of public policy and its publication would be prejudicial to the effective conduct of public affairs.

ANNEXES LIST:-

ANNEX A **The rationale for introducing the NRT**

ANNEX B **The design and development of the National Reference Test**

ANNEX C **NRT Sub-Group minutes (closed)**

ANNEX D **Proposal for the use of NRT outcomes in GCSE awarding**

Annex A

The rationale for introducing the NRT

The Board was advised in January 2014 (paper 103/13 Setting GCSE and A level Grade Standards) of the following rationale for the introduction of a Reference Test.

1. While the use of comparable outcomes at A level has been little criticised of late, the same cannot be said for GCSE. This seems to have been a consequence of the different contexts within which these qualifications operate and their different purposes.

2. A level results are primarily used for selection into higher education courses. From the universities' perspective, keeping the national A level grade outcomes broadly constant from year to year serves them well.

3. At GCSE the position is different. Schools in the state sector feel under great pressure from the Government's targets, particularly expectations that proportions of 16 year olds having achieved grade Cs in high profile subjects will rise year on year. There are currently no similar pressures for schools and colleges in relation to 18 year olds. A clear tension has arisen between Government expectations for 16 year olds' attainment and the application of the comparable outcomes approach at GCSE beyond the first year of new exams. The implication of keeping the comparable outcomes approach in years 2, 3, 4 etc of the GCSE exams is that national grade C outcomes will remain broadly constant from year to year despite schools' increasing efforts to improve their performances.

4. The Ofqual public position has not been that national grade C outcomes will necessarily remain the same from year to year. We say on our website:

   > *We believe that grade inflation – year-on-year increases in results without any real evidence of improvement in performance – should be avoided. It undermines confidence in the qualifications and in students' achievements. Our approach aims to control grade inflation, but to allow genuine improvements in performance to be recognised.*

5. The problem lies in how the comparable outcomes approach squares with allowing "genuine improvements in performance to be recognised". This is not an issue about the first year of new exams. It is in the use of comparable outcomes in the following years.

6. If there is a genuine improvement in performance of students in the second year of an exam it is likely that is largely because their teachers are more familiar with the requirements of the course and

the nature of the exams and so are better able to prepare students. It is unlikely that this improved performance indicates that the latter cohort is substantially better in terms of, for example, their capacity for future learning. If we don't want unfairly to advantage the second cohort, the use of comparable outcomes appears appropriate. In doing so we should acknowledge that any small increase in, for example, students' capacity for future learning would not be recognised by increases in greater proportions of higher grades.

7. Suppose though that in the fourth year of a GCSE examination there are *genuine improvements in performance* of the students. The Ofqual position is that this can be an acceptable justification for the proportion of students awarded a higher grade that year to rise.

8. The challenge arises from the nature of that evidence we require. Using only their judgement of scripts, awarders are unable to make the fine judgements necessary to decide whether a grade boundary should be put at, for example, 62, 63 or 64 marks. If that is the case, without an improbable great hike in performance from one year to the next, it is demanding for awarders' judgements to provide persuasive evidence. Indeed that is the justification for the use of a reference test to help us maintain grade standards – in a performance sense – in the new GCSEs.

9. As the Chief Regulator said to the then Secretary of State in her 22 August 2012 letter, our comparable outcomes approach can make it harder for genuine improvements in performance to be fully reflected in the results. It is important though that we remain open to the possibility that an exam board could present us with evidence in this regard which, after careful consideration, we concluded did indeed justify an out of tolerance award.

10. The challenge described above is the justification for the use of National Reference Tests to help us maintain grade standards – in a performance sense – in the new GCSEs. The purpose of the Reference Tests is to provide Ofqual and the exam boards with that credible, persuasive evidence that will allow us to justify a change from one year to another in the proportion of 16 year olds getting a high grade in their GCSEs.

Annex B

The design and development of the National Reference Test

Summary of NRT test design

1. The content and cognitive domains are based on the subject content and assessment objectives issued by DfE for the reformed GCSEs in maths and English language.

2. Test items are designed to strike a balance between making them accessible to students who are studying to take the GCSE and ensuring that they are not too close in style to those used in the GCSE question papers, or to those of any specific awarding organisation. In this way, performance in the National Reference Test should be resistant to the possible increase over time in formulaic approaches to question answering.

3. Test results focus on measuring changes in performance at three key grade boundaries: 3 / 4 (the current D / C boundary), 4 / 5 (reflecting government's expectation of what will constitute a 'good pass'), and 6 / 7 (the current B / A boundary).

4. Both the maths and English tests use a spiralled booklet design: eight overlapping booklets with each test item appearing in two booklets – each student will take 25% of the complete test. Booklets are of a broadly similar level of demand. Items in maths booklets increase in demand through the booklet with a few items at the end of each booklet that draw on content that only students studying for the higher tier GCSE will have studied; however, the maths NRT is not tiered and all students are encouraged to attempt all questions. The English booklets all contain a reading section followed by a writing section with most items being of an extended response design which are double-marked.

5. Most items will be re-used each year to provide the anchor in standards but the test design allows for some items to be replaced, for example, if they no longer appear to be functioning well. Each year some new items will be developed and trialled alongside the live tests to provide a reserve bank of items for test re-fresh.

School and student sample

6. A two stage, stratified sampling approach is used to achieve a nationally representative student sample.

7. A stratified school sample is created based on schools' historic performance in GCSE maths and English language. A main sample of schools is drawn from each stratum using the number of students enrolled in year 11 at each school to produce a distribution reflecting school size. Replacement schools are also selected for

each school in the main sample which can be used if a school in the main sample does not agree to take part in the NRT. Each school in the sample is asked to take part in both the maths and English NRT.

8. Schools provide a list of all students in year 11 who will be entered for the GCSEs in maths and/or English language and from this list, students are selected at random to take the test in one of the subjects.

Measuring performance

9. Classical test theory and IRT approaches are used to evaluate how test items perform. IRT models are used to establish the ability scale, to estimate the percentage of students that have achieved the level of performance in the NRT representing the three key grade boundaries of 3/4, 4/5 and 6/7.

10. The NRT ability measure relating to each of these grade boundaries will be established in 2017 after the GCSE results are available. In subsequent years, the NRT will indicate whether there has been a change in the percentage of students achieving these abilities representing each key grade boundary.

11. It is important to note that the NRT Is designed to measure the performance of the cohort, not to predict individual students' performance at GCSE.

12. The NRT is specified to achieve a level of precision at the three key grade boundaries of +/- 1.5% at the 95% confidence level. However, it was recognised that, until the test had been trialled, the level of precision that could be achieved could not be confirmed.

Test development

13. NFER developed items for the NRT in preparation for a field trial, held in autumn 2015. 175 schools took part and each item was trailed with about 200 students. Twice as many items were trialled as were required for the completed test. NFER recruited and trained a team of markers, therefore enabling the mark schemes also to be trialled. Item performance data was prepared using both classical test theory and IRT approaches.

14. At this stage, NFER reviewed the items and mark schemes with a team of subject experts and also with a panel of experts who advised on the accessibility of items for students with disabilities and special educational needs. Ofqual also reviewed the items with the support of our own subject experts. They had also advised Ofqual during DfE's development of the reformed GCSEs' subject content and the accreditation of exam boards' qualifications.

15. Using data from the field trial and feedback from subject experts, NFER amended some items or their mark schemes before selecting the better performing items to construct the test booklets. Ofqual completed a final review of the completed booklets and mark schemes to confirm that the test materials were suitable to be used in the PRT and conformed to the overall test design.

Preliminary Reference Test – to provide a full-scale operational trial

16. The PRT was held between 7 and 15 March 2016. The drawn sample aimed to achieve 7,500 students taking each test across a sample of 300 schools. The PRT took place in 326 schools, of which 312 schools (5388 students) took part in the maths test and 226 schools (4044 students) in the English test. Only students who were in year 11 and being entered to take the GCSE were eligible to take the test.

17. The smaller number of schools taking part in the English test was a consequence of many schools entering students for the international GCSE in English language who were therefore not eligible to take the NRT.

18. The lower number of students taking the test was also a consequence of a higher than expected absence rate on the day of the test; about 20% of students that had been selected to take the test did not turn up although this percentage varied considerably across the schools, some achieving a full turnout but a few with about half of the selected students missing. This highlights the practical challenges of testing a sample of students drawn from across the entire year group in each school.

19. NFER struggled to recruit schools to take part in the test. It had to approach all the replacement schools as well as those in the main sample. The PRT was a voluntary exercise for schools and many were unwilling. Government has now introduced legislation that will make it mandatory for most schools to take part which should improve future recruitment. Also, schools may respond more positively when they know that the test is 'for real', rather than a trial during the test's development. We also expect that virtually all state schools will enter students for the reformed GCSEs in 2017, which will increase the proportion of schools in the main sample that have students who are eligible to take the NRT. This will help particularly with recruitment for the English test. Achieving the school sample will contribute significantly to improving the precision of the NRT results.

20. Feedback from schools on NFER's administration of the tests was very positive. Ofqual also attended 10 schools to observe the tests. There was evidence in some schools that communication between NFER's test administrator and school staff before the day of the test could have been better although this also clearly depends on how

effective communication is within the school team.

21. No security incidents were identified during the administration of the tests. Both NFER and Ofqual monitored social media and other on-line forums, and no test items appear to have been published.

22. NFER completed the marking of the scripts on schedule. There is scope to improve control of marking through more effective use of seeds and this will be addressed before the test in 2017 is held.

23. Cito, NFER's sub-contractor, completed the analysis of the test data on schedule and this enabled NFER to report draft results to Ofqual as planned. Ofqual, NFER and Cito reviewed the draft results enabling their completion by early June. From 2018, it is important that NFER can report results on schedule to enable Ofqual and the exam boards to consider how they should be taken into account when GCSEs are awarded.

24. The analysis of the test data confirmed that all items in both the maths and English tests functioned well. No item had to be excluded from the analysis to create the ability scale. All items are considered to be suitable to use in live testing. Mark schemes are also considered fit for purpose. A few small amendments may be made to the mark schemes and opportunities to improve the training of markers in how to use the mark schemes have been noted.

25. The precision of test results in the PRT was not as good as had been specified in the test design. This was mainly caused by the difficulty in recruiting schools in the main sample to take part, schools not having students that were eligible to take the test, and a higher than expected level of student non-response. Cito has estimated that if we can overcome these problems, we might expect the level of precision to improve by about a third. This would bring the precision of the maths test to about the specified level of +/- 1.5% at the 95% confidence level and to just over 2% for the English test. Other steps are being considered that might also improve the level of precision.

Annex D

Proposal for the use of NRT outcomes in GCSE awarding

This is a high level proposal regarding how we could approach determining the application of the NRT outcomes to GCSE awards in England. Detailed governance arrangement including detail of the decision making processes would have to be developed and documented.

1 At the start of May 2018, Ofqual will receive from NFER the draft technical report on the 2018 NRTs, providing statistical analyses of NRT data and their implications. This report would be shared in confidence as soon as possible with the responsible officers of AQA, OCR, Pearson and WJEC.[2]

2 In mid-May, Ofqual would convene a technical panel to scrutinise the report. The panel would have available NFER's draft technical report from that year's NRT. The terms of reference of the panel would focus on it providing advice to the Grade Standards Forum (see step 3 below) on what actions to take in that summer's GCSE awards on the basis of the NFER report on the NRTs and other pertinent evidence such as research into the Sawtooth Effect. The membership would be:

- Chair: Executive Director of General Qualifications, Ofqual
- Two Ofqual representatives from Strategy, Risk and Research
- One technical representative from each of AQA, OCR, Pearson and WJEC.
- Two external, independent representatives appointed by Ofqual. These might be members of Ofqual's Standards Advisory Group (SAG), or from the NRT Sub Group of SAG that we have established.

NFER would be invited to send one or two technical advisers. The timing of the forum allows for NFER to be asked to provide some limited additional analysis of the NRT results before finalising its Technical Report by the end of May.

3 In early June, Ofqual would convene the Grade Standards Forum, chaired by the Chief Regulator and including senior members of Ofqual staff together with the responsible officers of AQA, OCR, Pearson and WJEC (plus their technical experts). To help logistics, the meeting might take place on the morning of the June Maintenance of Standards meeting.

The purpose of the Forum would be to discuss what action if any would be appropriate to take in that summer's GCSE awards on the basis of the advice from the technical panel. This would be informed by the advice from the technical panel. The decision would be Ofqual's to take. The role of the

[2] We would have the opportunity to use the 2017 NFER report to run an induction session with the boards in early 2018 so they were clear what to expect in summer 2018.

boards at the Forum is so that they can make representations. There might also be benefits in inviting observers, for example, from the head teacher and teacher associations, and from the Wales and Northern Ireland regulators but their status and purpose would have to be made very clear.

4 Following the Forum, Ofqual would have to make a regulatory decision about exactly how that year's NRT outcomes would be applied in that summer's GCSE awards. The process for that decision will have to be carefully developed, agreed and documented.

5 The actions would then be built into Ofqual's data exchange processes, to be issued to the boards in the second half of June with the NRT-related part not published before GCSE results day. (NFER is contracted to provide a General Report of that year's NRT results that is suitable for publication. These data will be Official Statistics.). Additional conditions of recognition or regulatory requirements underpinned by conditions might also be required.

6 Assuming the process above described is successful, it would be repeated on an annual basis.