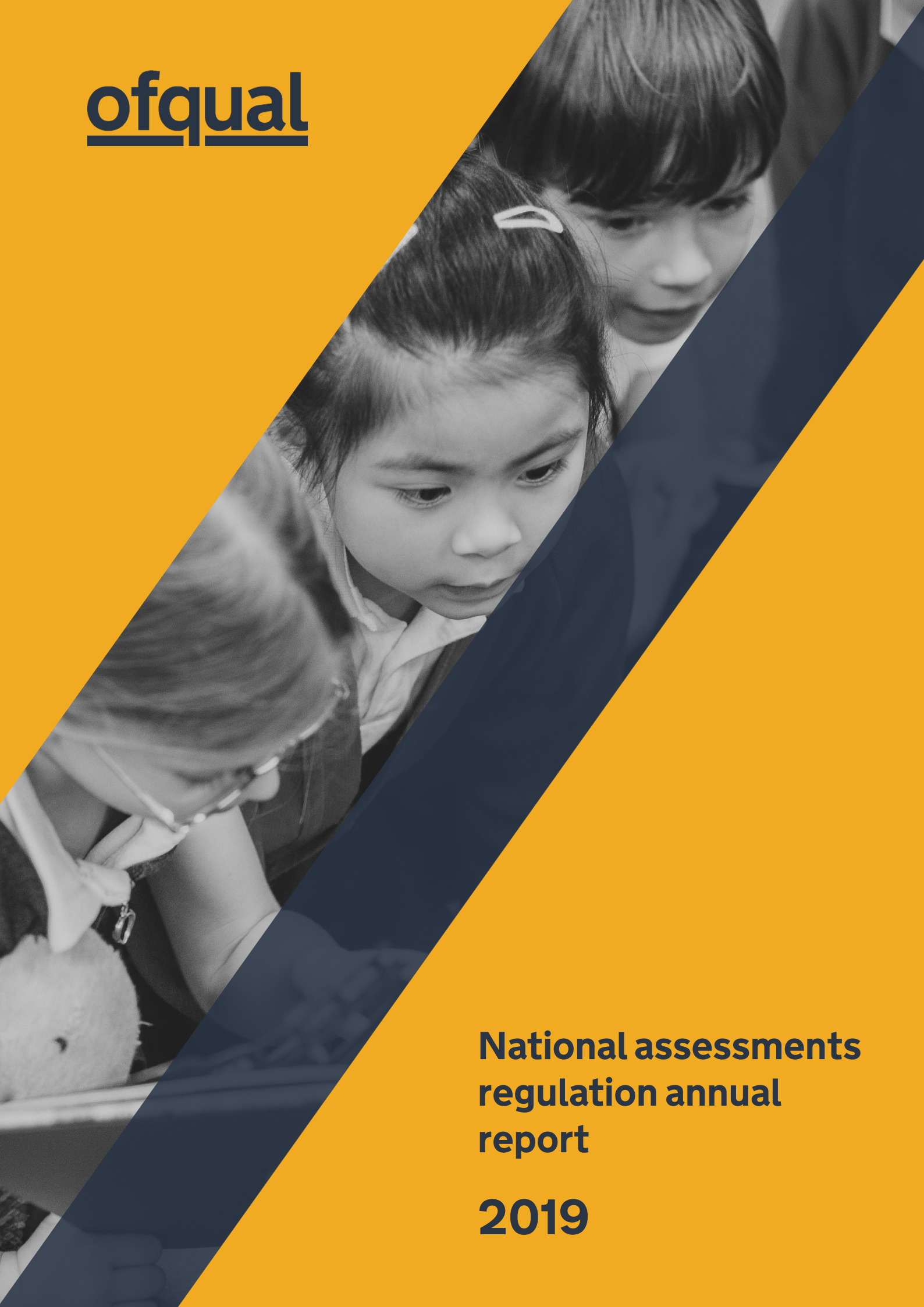
ofqual



National assessments regulation annual report

2019

Contents

Executive summary

This report explains how we have regulated national assessments in 2019. We continued to take a risk-based approach, focusing on key aspects of validity and high-stakes assessments, such as those at key stage 2 which underpin school accountability. Our report describes how we are meeting commitments made in our corporate plan for 2019/2022.

For national testing, this report provides assurance in a number of areas.

We continue to be satisfied that the Standards and Testing Agency (STA) took an appropriate approach to making sure that the new standards set in 2016 were effectively maintained through 2017 and 2018 to 2019.

Our analysis suggests that the consistency of STA's external marking remains very high, with 99.4% of markers agreeing with the definitive mark (set by senior markers) across 6.4 million marked items. We observed high-quality marker training materials being developed and used.

Our observations of test development meetings continue to suggest a strong focus within the STA on the validity of each national test produced.

We have continued to monitor and provide feedback on key areas of risk to validity, for example, including in relation to the moderation of teacher assessments, and the approach to the prevention and detection of malpractice.

This year we also agreed a Memorandum of Understanding with STA. This aims to support effective working between Ofqual, as the regulator of National Assessments, and STA, as the regulated provider of National Assessments. The Memorandum was published in September 2019 to bring greater public transparency to the relationship.

2020 will see the introduction of the Multiplication Tables Check in Year 4 and the Reception Baseline Assessment for children entering the reception year. STA will also be using a new supplier to deliver 2020's summer test operations, including key stage 2 marking. Our strategic focus during this period will remain the validity of assessments, including monitoring for risks to validity that could arise as a result of any changes. We will also continue to provide technical advice where relevant and report on processes critical to supporting national assessment validity.

Introduction

About National Assessments regulation

Ofqual regulates statutory early years foundation stage profile (EYFSP) assessments and statutory national curriculum assessments (some of which are also known as 'SATs'), which, together, we refer to as 'national assessments'. Ofqual's national assessment objectives, duties and powers are set out in law. We are responsible to Parliament, primarily via the Education Select Committee, rather than to government ministers.

Our objectives are to promote standards and confidence in national assessment and our primary duty is to keep all aspects of national assessments under review.

We focus on validity, that is, the quality of assessment. We also have a duty to report to the Secretary of State if we believe there is, or is likely to be, a significant failing in national assessment arrangements.

We fulfil our objectives primarily by observing, scrutinising and reporting on key aspects of assessment validity. We take a risk-based approach, which includes focusing on those assessments which have the 'highest-stakes', such as those relied upon within school accountability measures. As well as identifying risks to validity that can be addressed by responsible bodies to improve the quality of assessments over time, our regulation also seeks to provide independent assurance as to whether evidence suggests that processes are robust.

Ofqual can provide advice to support government decisions about future assessments, but we do not decide what national assessments there should be; nor are we responsible for the curriculum or school accountability policy. These things are determined by the Secretary of State for Education.

The primary body responsible for national curriculum assessments is the Standards and Testing Agency (STA). STA is an executive agency within the Department for Education (DfE) and may contract with suppliers to help develop, deliver or monitor national assessments. Other organisations also have responsibilities for aspects of national assessments, including local authorities, schools and other parts of DfE, for example, teams responsible for early years assessment.

Context for 2019

2019 was the fourth year of reformed assessments based on a new primary national curriculum. Assessments include teacher assessment and tests at key stages 1 (KS1) and 2 (KS2). 2019 was the second year of revised teacher assessment frameworks for writing at KS1 and KS2. This year saw the introduction of revised teacher assessment frameworks for reading, mathematics and science at KS1 and for science at KS2. There was no KS2 science sample test taken in 2019, however results for 2018's science sample test were published in July 2019 (see section B below). The science sample test is next due to be administered in 2020.

Section A: Priorities for 2019

During the 2019 assessment cycle, we monitored processes critical to maintaining test validity, including test development, standards maintenance and marking procedures. We continued to focus on KS2 assessments, as the highest-stakes national assessments which underpin both progress and attainment school accountability measures.

In March 2019, we published research on international approaches to the assessment of writing at the end of the primary stage. We also monitored STA's ongoing response to previous research and reviews including in relation to moderation and malpractice, and began a wider piece of work to consider the impact of increasing test familiarity on outcomes in a range of contexts.

We continued to monitor the introduction of our revised Regulatory Framework for national assessments (published in March 2018). In September 2019 we published a Memorandum of Understanding between Ofqual and STA to provide greater clarity and transparency in our national assessments role and working relationship with STA. We also began to monitor preparations for a new test operations supplier for possible impacts on assessment validity, as set out in our letter to the Education Select Committee (published in September 2018).

We monitored the development of new assessments, such as the reception baseline and multiplication tables check, in line with our response to the primary assessment consultations (published in June 2017), providing technical advice where appropriate.

This report summarises our activities and provides a view on key aspects of the validity of national assessments in 2019.

2019 priorities: summary

1. Monitoring processes supporting the validity of summer 2019 tests, in particular test development, marking and standards maintenance (Section B)

2. Publishing research on international approaches to writing assessment at the end of the primary stage, monitoring STA's ongoing response to previous research and beginning work to explore the impact of increasing test familiarity on outcomes (Section C)

3. Continuing to monitor the introduction of our revised regulatory framework for national assessments, new assessments in development and preparations for a new test operations supplier for possible impacts on assessment validity (Section D)

Section B: Monitoring assessments in 2019

Test development

National curriculum test development is a complex, technical process, with test items being developed over approximately 3 to 4 years. Tests must meet all the requirements of the relevant Test Framework, including sampling appropriately from the national curriculum, providing effective differentiation across the range of pupil performance within the national cohort, providing appropriate accessibility and meeting diversity and inclusion requirements.

National curriculum tests are psychometrics-based assessments. Two major strands of evidence inform the test-development process: qualitative professional judgement, generated through expert reviews, and statistical evidence, generated through pre-testing processes (including Technical Pre-Testing).

For tests at both KS1 and KS2, expert reviews take place regularly during the development process, often at three key points:

- prior to the first item (question) trial (Item Validation Trialling, or IVT)
- prior to the second trial (Technical Pre-Test or TPT)
- prior to the live test.

The statistical evidence which informs test-construction comes from a Technical Pre-Test (TPT), where test items are trialled by a carefully-selected sample of pupils and the outcomes are subject to detailed statistical analysis.

The test development process followed by the STA means that live tests (both question papers and mark schemes) are constructed on the basis of a wide range of evidence, including real pupil answers to questions (from trialling); item-functioning data (from the Technical Pre-Test); the views of practising teachers, inclusion and disability experts, expert markers and professional test developers.

During the past year, we have observed a number of test development meetings (9) across key stages 1 and 2 for a range of subjects (reading, mathematics and grammar, punctuation and spelling). We did not observe all types of meetings held, but sampled across the range. The meetings we observed included:

Expert review: Test Review Group (TRG), where teachers and subject experts provide advice on items in development.

Expert review: Inclusion, where experts in different types of special educational needs and disabilities, SENCOs and cultural inclusion experts provide views on items and texts in development.

Resolution: where the views of expert reviewers are considered and any issues arising are resolved.

Item Finalisation: where the outcomes of trialling are given detailed consideration and decisions are made about what needs to happen to each item in the light of the evidence gathered.

Test construction: where final decisions are made about the items to be included in a particular test, with careful attention given to ensuring that the test meets the requirements of the relevant Test Framework and items are appropriately ordered.

Project Board: governance meetings where senior staff are presented with validity evidence and outcomes of the test development process at key milestones, to approve for trialling or live test use.

In last year's report, we noted that the STA had reviewed its expert review process to improve consistency between subjects and ensure that a wider range of voices is heard. At the test development meetings we attended, we observed expert discussions, clearly informed by a range of evidence. The Test Frameworks, which set out the requirements for each test, were key to decision-making. There was regular and careful checking of National Curriculum references, to ensure that the questions reflected the breadth and depth of that curriculum, and close reference to

the cognitive domains, to ensure that the items were targeting the required skills. We observed good use of statistical evidence to support judgements about the effectiveness/validity of the items. Once again, we observed effective questioning of expert reviewers by STA test developers, ensuring that the STA had a clear understanding of the points being made. STA test developers were particularly keen to ensure that test items reflected and would support good practice in the classroom.

During trialling, the STA uses 'coding' instead of marking. Instead of being marked simply correct/incorrect, trained coders give each specific response a different code. STA test developers can then identify the number of pupils in the trial who gave that response. Coding is particularly useful because it provides STA with evidence of the range of responses that pupils give, allowing them to consider well in advance of the test which answers should count as credit-worthy and which should not. This enables detailed and evidence-based mark schemes to be produced.

It is, of course, not always possible to specify in a mark scheme every possible pupil response. During test development meetings, we observed careful attention being given to how mark schemes could be used in marker training to support markers to 'think in the right way' about how to mark responses not necessarily covered in the mark scheme. Across meetings where mark schemes were being considered, we observed a high degree of professionalism, attention to detail, care and thoroughness.

At test development meetings, regular consideration was given to how the process could be further refined. There was a strongly collaborative culture, where senior members of STA staff both supported and (where necessary) challenged less experienced colleagues. Governance meetings demonstrated senior managers providing effective oversight.

In conclusion, our observations of test development meetings indicated that STA retains a keen focus on assessment validity in every test produced.

Marking quality

KS1 tests are marked by teachers, to inform teacher judgements, while KS2 tests are externally marked. External marking allows for a greater degree of control over marking quality and is a key process supporting the validity of KS2 testing. The quality of marking is an important consideration when making judgements about the validity of National Curriculum Tests, so we have continued to carefully monitor the quality of STA's marking.

For 2019, the same measures were in place to assure the quality of marking for externally marked tests as there had been in previous years. These included:

- training all markers using the same script and training materials
- requiring markers to pass a training exercise prior to live marking
- testing accuracy during marking against 'validity items' that have already been marked by senior markers
- maintaining a marking hierarchy to provide oversight and ensure that items that markers are unclear about can be 'escalated' to a more senior marker
- stopping markers from marking particular items if their marking is not of sufficient quality and re-marking relevant items.

As in previous years, this year we observed KS2 marker training for reading, mathematics and grammar, punctuation and spelling tests (GPS). Across each year we have noted the high quality of marker training and our analysis of the STA's marking data has indicated that marking is highly reliable across all subjects. Nonetheless, reading items are, by their nature, more challenging to mark, and for 2019, we carried out additional monitoring on the marking of reading, to provide further information about how markers were trained in this subject.

We observed the range of meetings relating to marker training for reading in 2019. At initial meetings, senior markers reviewed and confirmed the final draft of the initial suite of training materials that had been drafted by the Marking Programme Leader, prior to User Acceptance Testing (UAT). Following this, materials were further reviewed in light of how they performed at UAT. Throughout the process, we observed careful consideration of how training points would be delivered and experienced, with all participants contributing. There was detailed finessing of materials to maximise the effectiveness of the training.

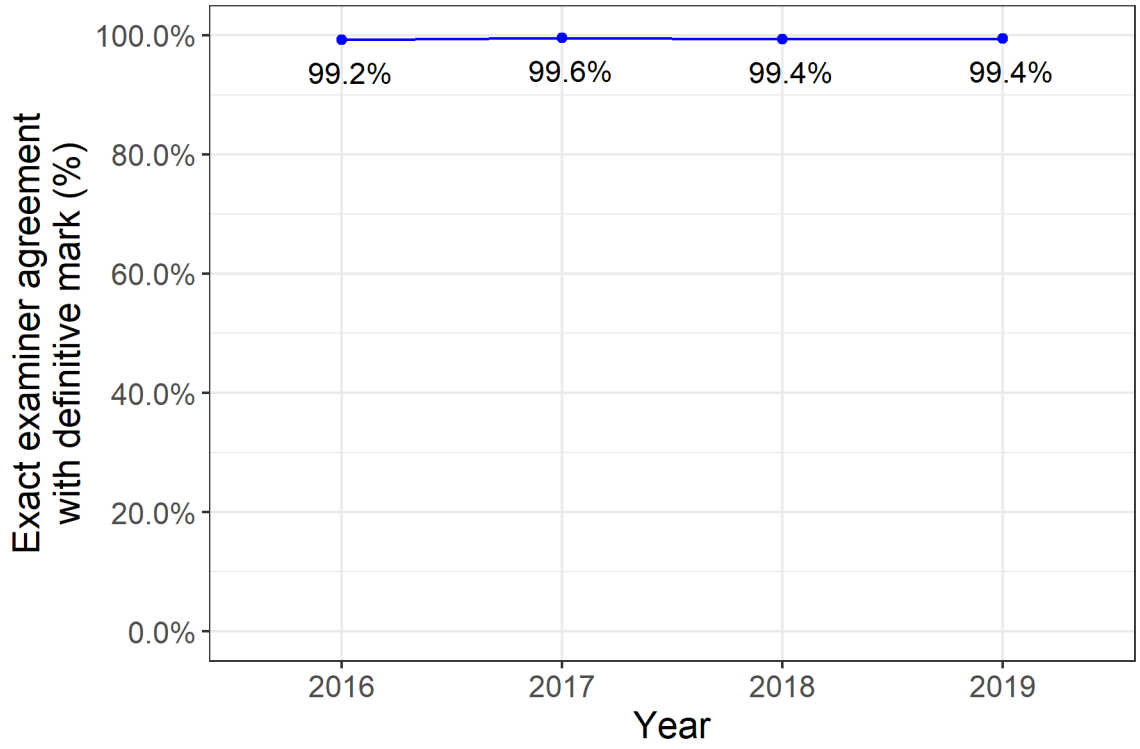
We observed the training of marking team leaders, where there was, similarly, a strong culture of professionalism and maintaining relationships. This allowed an open approach, where people appeared comfortable to talk about where they had made mistakes and why. Senior markers were taking notes and were responsive to feedback. This supported the development of a clear, common understanding of the mark scheme. There was a shared understanding that high quality training materials underpinned the effectiveness of the marking process.

Marker training is delivered through a 'cascade' model, with the most senior markers being trained first and then delivering training to more junior markers. This meant that there were some differences in the style of presentation, but no evidence that these differences impacted on the quality of the training. Markers received training on each item that they would be marking, with detailed explanations of which points were credit-worthy, so could be given marks, and which were not. Markers completed training exercises, which were checked by team leaders. Again, there was an open and professional atmosphere, with markers happy to raise questions and seek further clarification.

Validity items (or 'seeds') are items which have been marked by expert senior markers, working with STA Test Developers, so the definitive mark for them is known. These are used to check that markers are marking accurately, so their quality is key to marking reliability. We observed validity items being selected carefully across subjects, to ensure that they covered the full range of the mark schemes, including marking principles.

We analysed operational marking data from 2019, using the same methodology as in previous years. Details about how these metrics are calculated are set out in Ofqual's report 'Marking metrics' (2016). In brief, metrics are created from the data arising from the operational monitoring of quality of marking during live marking sessions. We assume that the most appropriate measure of consistency of marking is based on the difference between two marks given for a single response. Thus the data used is the mark–re-mark data from validity items (ie an analysis of the difference between the 'definitive mark' set by the senior markers and the actual mark awarded). Our analysis suggested that the consistency of STA's external marking in 2019 remained very high, with 99.4% of markers agreeing with the definitive mark across approximately 6.4 million marked items. This analysis

provides evidence that the quality assurance measures currently in place for KS2 marking (summarised above) are effective.



The graph above provides evidence that the quality of STA marking has remained high across all years since new testing arrangements were introduced in 2016. According to our analysis, exact agreement with the definitive mark is above 99% throughout 2016-2019.[1]

Standards maintenance

Each time a new test is produced, the exact level of difficulty is likely to be slightly different from previous tests. So any test developer, such as STA, must ensure that there is a technically appropriate process for maintaining standards, to ensure that the meaning of the test result remains consistent between different tests.

To maintain test standards each year, STA use a psychometric (statistical) process called 'equating'. This is supported by significant pre-testing of items used in live tests over a number of years, alongside 'anchor' items, which are test questions for which the standard is already known.

The equating process is explained in lay terms in our national assessments report for 2017, available here. Technical detail on the equating process is set out in the STA's Test Handbook.

The process for maintaining test standards in 2019 was based on the same assumptions and professionally recognised techniques as in 2017 and 2018. We reviewed these assumptions in 2017.

[1] Figures are rounded to one decimal place.

We observed the standards maintenance meeting for both KS1 and KS2 tests in 2019. Evidence from test equating indicated that the 2019 mathematics and GPS tests were slightly more difficult than the 2018 tests and the 2019 reading test was of a similar level of difficulty to the 2018 test. We are content that 2019's standards maintenance meetings were carried out in-line with the procedures set out in the STA's Test Handbook, and took due account of data indicating small differences in the difficulty of the tests, aiming to ensure that results reflected the attainment of pupils, rather than the level of difficulty of the particular tests.

Results

Key Stage 2

Following the marking window, in which scripts from more than 600,000 pupils (circa 3.5 million papers) were marked in approximately 3 weeks, KS2 test results were made available to schools on 9 July 2019, alongside national results and raw-score to scaled-score conversion charts.

This year, 79% of pupils nationally met the expected standard in mathematics, which represented an increase of 3 percentage points from 2018. 78% of pupils met the expected standard in the GPS test, which represented no change from 2018. 73% of pupils met the expected standard in reading, which represented a fall of 2 percentage points from 2018.

After the introduction of the new national curriculum tests in 2016, results increased in all subjects in 2017 and in 2018. Increases in the early years of a new test can often be attributed to pupils and teachers becoming more familiar with the content and style of the new tests[2]. Such effects are complex and contested and it is very difficult to understand when these types of effects are no longer active (see also Section D below).

Turning to teacher assessment, revised assessment frameworks for writing were introduced in summer 2018; 2019 was the second year of their use. As we noted last year, significant changes were made for 2018, including the introduction of a "more flexible" approach to the 'secure-fit' framework. The revised frameworks removed some of the previous criteria and added other elements; some criteria also changed significantly in their focus. We also noted that changes, no matter how small, to 'secure-fit' or 'mastery' assessment criteria will make at least some change to the overall assessment standard. For these reasons, pass rates cannot be directly compared between the old and new frameworks. In addition, teacher assessments are not subject to the same controls as externally set and marked tests. The national proportion of pupils reported to have reached the expected standard for KS2 writing, based on teacher assessments, was 78% in both 2018 and 2019. The combined figure for pupils achieving the expected standard in reading, writing and mathematics for 2019 was 65%, a small change from 2018 (64%).

For KS2 science, 83% of pupils were reported to have met the expected standard in 2019 based on teacher assessments. As the frameworks were changed for 2019, results cannot be directly compared between 2018 and 2019.

In 2019, results were also reported from the KS2 science sample test taken in 2018. This is administered biennially by STA in a selection of schools to inform a national

[2] Ofqual's research on similar effects in relation to qualifications can be found here.

view of standards in primary science. Individual results are not provided to schools or pupils. June 2018 saw the third live administration of this test (the first took place in June 2014 and the second in June 2016). In line with other KS2 assessments, reporting arrangements changed in 2016. Previous national curriculum levels were removed and replaced with new scaled scores and a new expected standard of attainment was set. To allow for effective estimation of outcomes for the 2018 cohort, there was a large overlap of items between that cohort and the 2016 cohort. The key stage 2 sample test is subject to a similar development process as other key stage 2 tests.

The results of the 2018 science sample test estimated 21.2% of pupils to be performing at the expected standard. This represents a slight decrease, from 22.3% in 2016. It should be noted that the proportion of pupils not sitting the test also rose, from 10% in 2016 to 14% in 2018. Since these pupils are counted as not performing at the expected standard, this is likely to have affected the outcomes.[3]

Some differences between teacher assessments and testing may be expected, due to differences in assessment methods, for example, that testing is based on a single performance and that the sample test covers more of the curriculum than the teacher assessment criteria. The gap between the percentage reaching the expected standard for science in 2018 according to teachers' assessments (82%) and the percentage reaching the expected standard for science that same year according to the sample test (21.2%) is notable. Testing is a highly controlled process compared to teacher assessments (which for science is also not moderated). The disparity between these two figures makes it difficult to draw reliable conclusions about KS2 science attainment; this area would benefit from further investigation.

Key Stage 1

KS1 outcomes are teacher assessed, informed by externally set, but internally marked, tests in reading and mathematics. 75% of pupils were reported to have met the expected standard for reading and 76% in maths. Changes made to the 2018/2019 KS1 reading, mathematics and science teacher assessment frameworks mean that 2019's results in these subjects are not directly comparable to those made in previous years.

Writing is teacher assessed but not informed by an externally set test. For KS1, as noted above for KS2, revised assessment frameworks for writing were introduced in summer 2018; 2019 was the second year of their use. 69% of pupils were reported to have reached the expected standard in writing in 2019. This represents a small (1 percentage point) drop on 2018. Under the new teacher assessment framework for science introduced for 2018/19, 82% of pupils were reported to have reached the expected standard.

The Year 1 phonics check has been in place since 2012. After an initial period of increase, which may have been partly related to increasing test familiarity, outcomes have remained broadly similar since 2016. Outcomes of the 2019 check slightly decreased (down 0.6 percentage points on 2018), with the rounded figure showing that 82% of pupils were reported as meeting the expected standard in Year 1.

[3] This is standard practice for national assessments which are used to measure school performance and, unlike qualifications, are not primarily used to certify the attainment of individual pupils.

Early Years Foundation Stage Profile

The EYFSP is an observational teacher assessment at the end of the reception year. After an initial period of increase following the introduction of a revised Profile in 2013, the rate of increase has slowed in recent years. Outcomes for 2019 showed that 71.8% of children achieved the 'good level of development' standard, up very slightly (0.3 percentage points) on 2018.

Section C: Research and reviews

International approaches to writing assessment

In March 2019, we published our review of international approaches to assessing writing at the end of primary education. This followed from our previous research on the consistency of local authority moderation of KS2 writing.

Our aim was to provide evidence that could both support stakeholder debate and inform government's ongoing exploration of potential alternatives to the current model. The research focused on large-scale assessments around the age of 10/11 and covered jurisdictions which are English speaking or use English for assessments. It also provided a summary of the different approaches taken to the assessment of writing in England since national testing was first introduced at KS2 in 1995.

The report begins by considering how writing might be conceptualised. It reviews the history of writing assessment under the National Curriculum in England, focussing on Key Stage 2 assessments, showing how both external testing and teacher assessment have been used and setting out the variety of approaches that have been taken. It then reviews 15 international assessments of writing used at the end of primary education, looking at both the assessments themselves and how they are marked and graded. The research also considers more innovative approaches, such as comparative judgement methods and automated (computer) marking. Finally, it draws on the evidence to discuss the advantages and disadvantages of different approaches. It did not seek to conclude which approach was 'best', because judgements should be made in the light of different policy contexts and the purpose of any particular assessment.

Updates on previous research and reviews

Moderation of teacher assessment

In 2018, we published research based on small-scale observations of the moderation of key stage 2 writing in summer 2017. This recommended that STA take steps to reduce the risk of inconsistency both between local authorities and individual moderators. We have continued to monitor moderation processes and STA's response to this research.

Having made improvements to guidance and communications for 2018 (including introducing new writing frameworks), for 2019, STA focused on further developing moderator training and materials. This year's training saw a significantly greater focus on supporting the validity of moderation and the consistency of moderator judgements than in previous years. For example, the meaning of specific individual statements within the teacher assessment frameworks was unpicked and

moderators were specifically trained on the detail. There was a clear focus on statements where evidence from 2018 had indicated inconsistent interpretation.

Scripts to exemplify relevant training points were provided and commentaries referred directly to teacher assessment framework statements, rather than providing general analysis. There was a greater focus on the technical (subject-specific) language moderators would use to explain the rationale for their judgements.

STA also made some improvements to administrative processes for 2019, including providing training earlier in the academic year to allow greater time for local authorities to cascade training, and increasing the number of training events to allow greater access. Refinements to the training schedule meant that there was more effective engagement with the materials from both trainers and moderators, with time for training exercises and discussion. Facilitators were better-prepared and more able to answer questions.

Moderation is a challenging activity and, as might be expected, there remained some areas where further improvements could be made. For example, while there were clear attempts to define and clarify the expected standard, delegates would have benefitted from a greater focus on the boundary between the expected standard (EXS) and the greater depth standard (GDS). A clearer focus in the training materials on work at the borderline would have been helpful here. Similarly, guidance would have been helpful on the extent to which each of the GDS statements have to be met for work to be judged GDS and the extent to which GDS has to be sustained throughout a piece/portfolio. For example, would it be sufficient for each of the GDS standards to be met, but only in (say) one piece of writing? Similarly, while the 'secure-fit' nature of the framework was mentioned at training, again, this would have benefitted from greater attention to promote a more consistent understanding. Some attendees also felt that there was a lack of clear guidance around the place and purpose of the professional discussion within the moderation process.

STA has committed to continuing to focus on the quality of support and training for moderators. This includes carrying out research into the consistency of the training cascaded by local authorities to their moderators. To address concerns about the predictability of moderation, STA is considering the feasibility of analysing data over time to understand the extent to which schools' judgements change in moderated versus unmoderated years.

We continue to recommend that STA considers what more may be done to improve consistency in application of the framework, both in terms of how judgements are made and the administrative processes used by different local authorities. We also continue to recommend that STA keeps under review the approach taken to the assessment of writing; it is helpful that STA is continuing to explore the potential of comparative judgement as a methodology for assessing writing. We will continue to keep these areas under review.

Malpractice review

The number of malpractice and maladministration[4] complaints made to the Standards and Testing Agency represents only a very small fraction of assessments

[4] 'Malpractice' is used by Ofqual and in this report to mean intentional maladministration. STA uses the term 'maladministration' and defines it as "any act that could jeopardise the integrity, security or confidentiality of the national curriculum assessments and could lead to results that do not reflect the

taken. Nonetheless, maladministration, whether deliberate or not, can lead to test results that do not accurately reflect the unaided work of the pupils and have significant impacts on public confidence. It is very important that all schools can be confident that tests are administered fairly. So in 2018 we carried out a review of the available documentation relating to STA's approach to malpractice prevention and detection. We provided feedback to STA on areas which had the potential to be strengthened, including in relation to test administration and independent monitoring, teacher assessment, safeguarding of confidential assessment materials and the use of special considerations and access arrangements.

In 2019 we have continued to monitor STA's response and commitments made in respect of our review. Since the review, STA has strengthened the requirement in local authority monitoring guidance that conflicts of interest must be managed, and updated test administration guidance to (a) proscribe smartwatches from being worn during testing and (b) strengthen the recommendation that schools arrange for tests to be independently observed. STA has also reviewed and strengthened its investigation process by making provision for schools to make written representations in response to investigation findings, prior to a final determination being made. However, STA has not made significant improvements in response to other aspects of our review. For example, we provided feedback on the need to improve data monitoring and analysis, however data monitoring was reduced in 2019 due to unplanned resourcing constraints, although STA plans to reinstate this for 2020.

Figures recently released by the Standards and Testing Agency have shown an increase in the number of allegations of maladministration[5] in 2018 compared to 2017. The change was mainly at KS2 where allegations increased by 34% (from 344 to 461). The number of annulments or amendments made to KS2 results due to maladministration rose by 56% (from 78 to 122). However, it remains the case that amendments and annulments took place for less than 1% of schools using the assessments.

STA has committed to focusing on the detection and prevention of malpractice for this year, for example, in relation to the analysis of relevant datasets. We will continue to monitor and report on this area.

Content validation study and reading review

In 2017 we published a content validation study, looking at how the new primary national curriculum was translated into testing and a review of evidence on the accessibility of the 2016 reading test. We found that the approach to curriculum sampling was robust, comparing favourably to international approaches for similar tests. In response to our review of the 2016 reading test, STA committed to a number of actions, the majority of which are either on track or completed (see 2018 report). However, while STA remains committed to researching data to understand more about why the 2016 reading test was not finished by 25% of pupils, regrettably, it has not yet been able to carry this out due to lack of sufficient resource. We will continue to monitor and report on STA's progress towards this commitment.

unaided abilities and achievements of pupils." – this definition includes both accidental maladministration and intentional malpractice/cheating.
https://www.gov.uk/government/collections/maladministration-reports

[5] STA do not currently categorise maladministration incidents by intent – ie whether they were intentional/cheating (malpractice) or accidental.

Exploring test familiarity

During 2019-20 we are exploring the wider phenomenon of the impact of the 'test familiarity' effect, where outcomes change in response to significant changes, such as to the syllabus or curriculum being tested. This work is looking broadly at the effect in a number of contexts, not just in relation to national testing, but more widely in relation to qualifications and examinations. It considers how the impact of such changes can be effectively managed, and what might be done to minimise the likelihood of results being misinterpreted following changes. We look forward to publishing this work later this year to aid stakeholders' understanding of how test outcomes can be interpreted following change.

Section D: Regulating through change

Memorandum of Understanding

As part of the ongoing monitoring of our revised Regulatory Framework for National Assessments (published in 2018), in 2019 we agreed and published a Memorandum of Understanding (MoU) between Ofqual and the Standards and Testing Agency. This MoU supports, and is underpinned by, our Regulatory Framework. It aims to clarify and codify our day-to-day regulatory relationship with STA.

The MoU is not intended to detail every aspect of our working relationship, but is a statement of principles to guide relations. It covers areas such as our approach to engagement, how we exchange information, and how we should escalate any concerns or disputes. Clarity in these areas is important as we come up to a period of change, with new supplier arrangements and new assessments being introduced in 2020 (see below).

We published this MoU in September 2019, to provide greater public transparency to our day-to-day regulation and interaction with STA.

New and reformed assessments

During 2018/2019, STA has continued to research and develop assessments, carrying out further work on the reception baseline assessment (RBA), due for introduction in autumn 2020, and the multiplication tables check (MTC), due to be introduced in the summer of 2020 for pupils in year 4. The Department for Education has also been carrying out ongoing development work on the Early Years Foundation Stage Profile, with a revised Profile due to be used statutorily from 2021/2022.

We have continued to engage productively with STA on issues relating to the validity of these assessments in line with our response to the primary assessment consultations, published in 2017. For example, we have monitored and provided feedback to STA in relation to the development of the RBA, with our focus being on the validity of the assessment. In autumn 2018 we observed early trials, in two schools, of items to be included in the assessment and we reviewed trialling data. We mainly observed items designed to assess children's ability to self-regulate. Such items appeared time-consuming and difficult to consistently deliver. They are

not now to be included in the live assessment based on STA's own analysis of trialling data[6]. We welcome the trialling, piloting and evaluation of the RBA prior to its introduction and have observed STA giving detailed consideration to statistical evidence, stakeholder views and teacher feedback from trialling; data suggests high levels of reliability for items proceeding to the pilot phase. As set out in our response to the primary assessment consultations, we continue to encourage government to take a cautious approach to introducing this assessment. The RBA is designed to underpin school accountability measures and, in line with our consultation response, we continue to recommend that all reasonable steps are taken to make sure that a fair and equitable approach is achieved, for example, by considering how to reduce pressures on the assessment and how contextual information may be used alongside assessment outcomes.

We have also monitored and provided feedback in relation to the multiplication tables check. We invited the STA to demonstrate to our Access Consultation Forum the ways in which this on-screen test could be adjusted to make it accessible for different learners. This meant that views on its accessibility could be given by a range of experts in special educational needs and disabilities. We also observed the governance meeting at which the evidence to inform the construction of the Multiplication Tables Check pilot was considered. The process followed was the same as for all other STA assessments: we observed detailed consideration of the statistical evidence from trialling so that STA could aim to ensure that every test form used for the pilot was at an equivalent level of difficulty as possible. Again, in line with our response to the primary assessment consultation, we welcome the ongoing trialling, piloting and evaluation of this assessment prior to its introduction. Such developmental work, plus clear guidance and effective communication of the assessment's purpose and how data from the check will be used, will be important to support its effective administration as the first online national assessment.

Finally, we have monitored and provided feedback on the Department for Education's work on revisions to the EYFSP assessment criteria. Following feedback, the Department commissioned additional research during the pilot phase to consider in more detail the clarity and interpretation of assessment criteria.

We look forward to continuing to monitor and report on these areas as the new assessments are introduced.

Operational change

STA has contracted with a new supplier to deliver test operations for 2020. While we are not responsible for quality assuring test delivery, as this rests with the Secretary of State for Education, we are interested in the validity of assessment under the new operational arrangements. To that end, in 2019 we began to monitor STA's approach for potential risks to validity. Our focus is and will continue to be the quality of marking carried out by the new supplier; test development, standards maintenance and moderation remain largely unaffected by the change. We will continue to monitor marking quality closely, and look forward to analysing data from the 2020 assessments. To find out more about our approach to regulating through this period of supplier change, please see our exchange of letters with the Secretary of State and the Chair of the Education Select Committee, published in September 2018.

[6] For more information, see STA's Reception Baseline Assessment Framework documents

Section E: Looking forward

Our regulation during the 2020 assessment cycle will reflect commitments in our Corporate Plan and will take account of the findings of our work in previous years, stakeholder views and changes to assessment arrangements. While our focus may change in response to events or new information, our key priorities for 2020 and beyond are likely to include:

- continued monitoring of those processes which can support validity, such as the test development process, standards maintenance model and marking processes
- monitoring changes to assessments, the development of new assessments and potential validity impacts of operational changes, particularly in relation to the quality of KS2 marking
- a continued focus on areas where we see risks to validity, such as teacher assessments used for high-stakes purposes

We look forward to reporting on our regulation and reflecting on national assessment validity in 2020.

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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

January 2020

Ofqual/20/6584