Standards & Testing Agency

National curriculum test handbook: 2016 and 2017

Key stages 1 and 2

December 2017

Contents

1 Introduction

The government introduced a new national curriculum for schools in England in 2014. This curriculum was assessed in mathematics, English reading and English grammar, punctuation and spelling for the first time in 2016. This handbook describes how the tests of the new curriculum at key stages 1 and 2 were developed, and presents validity and reliability evidence related to the tests.

1.1 Purpose of this document

This handbook has been produced by the Standards and Testing Agency (STA) to explain how the tests of the national curriculum are designed, developed and delivered. STA is an executive agency of the Department for Education (DfE) and is regulated by the Office of Qualifications and Examinations Regulation (Ofqual) using Ofqual's regulatory framework[1] for national assessments.

This document will be of interest to those involved in assessment, including in schools.

[1] Available at: www.gov.uk/government/publications/regulatory-framework-for-national-assessments

2 The tests

All eligible pupils in England who are registered at maintained schools, special schools or academies (including free schools) are assessed at the end of key stages 1 and 2 in mathematics, English reading and English grammar, punctuation and spelling[2]. These tests are renewed annually and are taken during a specified period in the summer term. Details of the key stage 2 science sample test will be provided in a separate handbook.

2.1 Development of the tests

The new national curriculum test models were developed in English and mathematics for pupils at ages 7 and 11 to align with the aims, purposes and content of the 2014 national curriculum.

STA used the relevant curriculums in English and mathematics to develop suitable test frameworks[3] which outline the content and specifications of the tests used at the end of key stages 1 and 2.

The tests reflect the content and cognitive domains detailed in the test frameworks. Care is taken within each test to ensure only the skills necessary to that test are assessed, so the tests are fair and valid.

2.2 Purpose and uses of the tests

The main purpose of statutory assessment is to ascertain what pupils have achieved in relation to the attainment targets outlined in the national curriculum (2014).

The intended uses of the outcomes, as set out in the Bew Report and the government's consultation document on primary assessment and accountability, are to:

- hold schools accountable for the attainment and progress made by their pupils
- inform parents and secondary schools about the performance of individual pupils
- enable benchmarking between schools and monitoring of performance locally and nationally.

[2] The key stage 1 test in English grammar, punctuation and spelling is non-statutory.

[3] Available at: www.gov.uk/government/collections/national-curriculum-assessments-test-frameworks

2.3 The test frameworks

The purpose of the test frameworks is to guide the development of the tests. By providing consistent parameters for the development and construction of tests, the test frameworks allow valid, reliable and comparable tests to be constructed each year.

The test frameworks were written primarily for those who write test materials. They have been made available to a wider audience for reasons of openness and transparency. STA developed the frameworks in consultation with the DfE curriculum and assessment teams, panels of teachers and subject experts to refine their content, confirm their validity and make them fit for purpose.

The test frameworks provide information pertaining to what the tests will cover. The frameworks do not provide information on how teachers should teach the national curriculum or assess pupils' progress.

Each test framework contains:

- a content domain, setting out which parts of the national curriculum can be assessed through the test
- a cognitive domain, outlining the demands of the test and the cognitive skills required for the subject
- a test specification, which gives details of test format, item types, response types, marking and the balance of marks across the content and cognitive domains. It also explains how the test outcomes will be reported
- the performance descriptors for each subject.

2.3.1 Development of the cognitive domain

The cognitive domains make explicit the thinking skills and intellectual processes required for each test. Each item is rated against the relevant components of the cognitive domains. By taking this information into account during test construction, the tests will be comparable in terms of cognitive skills and demand of the items from year to year. This contributes to the reliability of the tests.

The cognitive domains were initially developed through a literature review. The existing domains within the literature were broad in style and some were more suited to specific

subjects than others[4][5][6][7], for example the CRAS scale for mathematics[8] and the PISA reading item difficulty scheme[9] for English reading.

These existing models were synthesised and amended to take account of the specific demands of each subject and the cognitive skills of primary-aged children, with resulting models for each subject that allow items to be rated across different areas of cognitive demand.

To validate the cognitive domains, panels of teachers reviewed the test frameworks. They were asked to comment on the extent to which the cognitive domain set out the appropriate thinking skills for the subject and age group. Also, pairs of test development researchers independently classified items against the cognitive domain and their classifications were compared.

Refinements were made to the cognitive domains based on both the inter-rater consistency between test development researchers and the comments gathered from the teacher panels, ensuring the cognitive domains published in the test frameworks were valid and usable.

2.3.2 Development of the performance descriptors

Performance descriptors describe the typical characteristics of pupils whose performance is at the threshold of the expected standard. Performance descriptors for each test were created by subject specialist test developers in conjunction with teachers and curriculum experts, using a variety of sources. The descriptors were reviewed and validated by a panel of teachers. The performance descriptors were published in test frameworks in

---

[4] Bloom's cognitive taxonomy (Bloom, B., Engelhart, M., Furst, E., Hill, W., and Krathwohl, D. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company) formed a useful starting point for thinking about cognitive demands such as knowledge, understanding and analysis and this was further refined by referring to footnotes 3, 4 and 5 below.

[5] Edwards, J. and Dall'Alba, G. (1981). Development of a scale of cognitive demand for analysis of printed secondary science materials. Research in Science Education, Volume 11, Issue 1 pp 158-170.

[6] Hughes, S., Pollitt, A. and Ahmed, A. (1998). "The development of a tool for gauging the demands of GCSE and A-level exam questions." Presented and published at BERA meeting August 27-30 1998.

[7] Lumley, T., Routitsky, A., Mendelovits, J. and Ramalingam, D. (2012). The revised PISA reading item difficulty scheme, a framework for predicting item difficulty in reading tests. Available at http://research.acer.edu.au/pisa/5.

[8] Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H. and Bramley, T. (1998). The effects of structure on the demands in GCSE and A level questions. Report to Qualifications and Curriculum Authority. University of Cambridge Local Examinations Syndicate.

[9] Kirsch, I., deJong, J., Lafontaine, D., McQueen, J., Mendelovits, J., and Monseur, C. (2002). Reading for change: Performance and engagement across countries: Results from PISA 2000. Paris: Organisation for Economic Co-operation and Development.

June 2015, meaning that schools were able to take account of them in advance of the first year of the new tests of the national curriculum in 2016.

An exercise was sent to attendees of the standard setting meeting for the 2016 tests, in which between 72% and 95% of participants indicated that they agreed or strongly agreed that the performance descriptors contained sufficient detail for mathematics, English reading and English grammar, punctuation and spelling. These agreement ratings provide evidence that the performance descriptors are fit for purpose.

3 The test development process

The test development process is based on an item-banking model. Each item is taken through a series of phases to establish whether it is valid, reliable and meets the purposes of the test. Items meeting these criteria are considered for possible inclusion in a live test. Any suitable items not used in a live test are kept in the item bank for potential use in future live tests. Live tests must meet the test specification in the test framework for each subject.

3.1 STA test development staff

Test development is conducted by STA's in-house test development division. This team comprises assessment experts, psychometricians, test development researchers and project staff. The team also involve staff from other parts of STA as required, such as STA's design team. STA's technical specialists have a range of experience both in the UK and internationally. This includes experience of working on a number of assessment programmes, classroom experience and expertise in the theories and techniques of assessment.

3.2 The process

The test development process takes approximately two and a half to three years to complete. An overview is provided in the diagram overleaf and full details of each stage are provided in chapters 4 to 13.

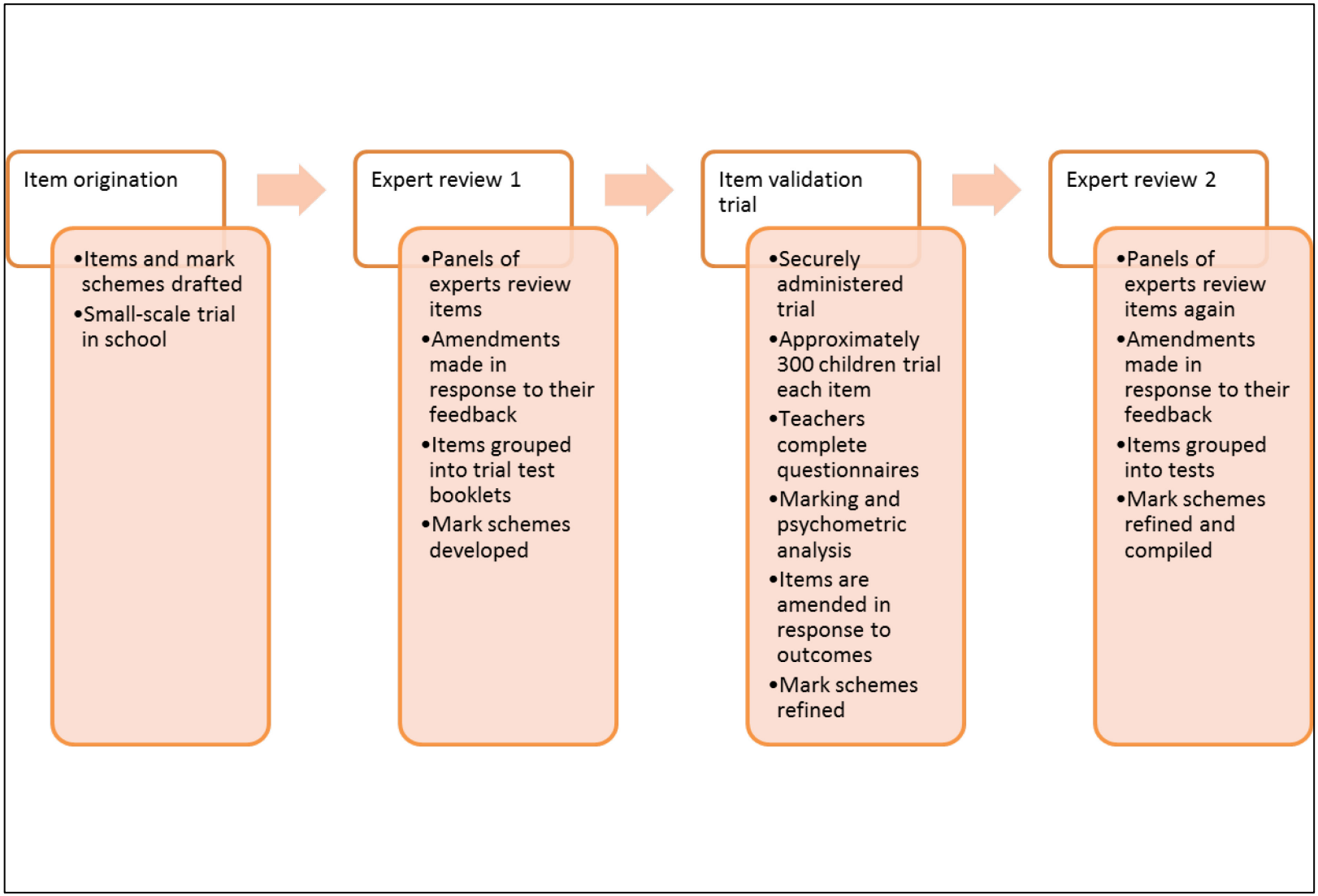

Figure 1: Overview of test development stages

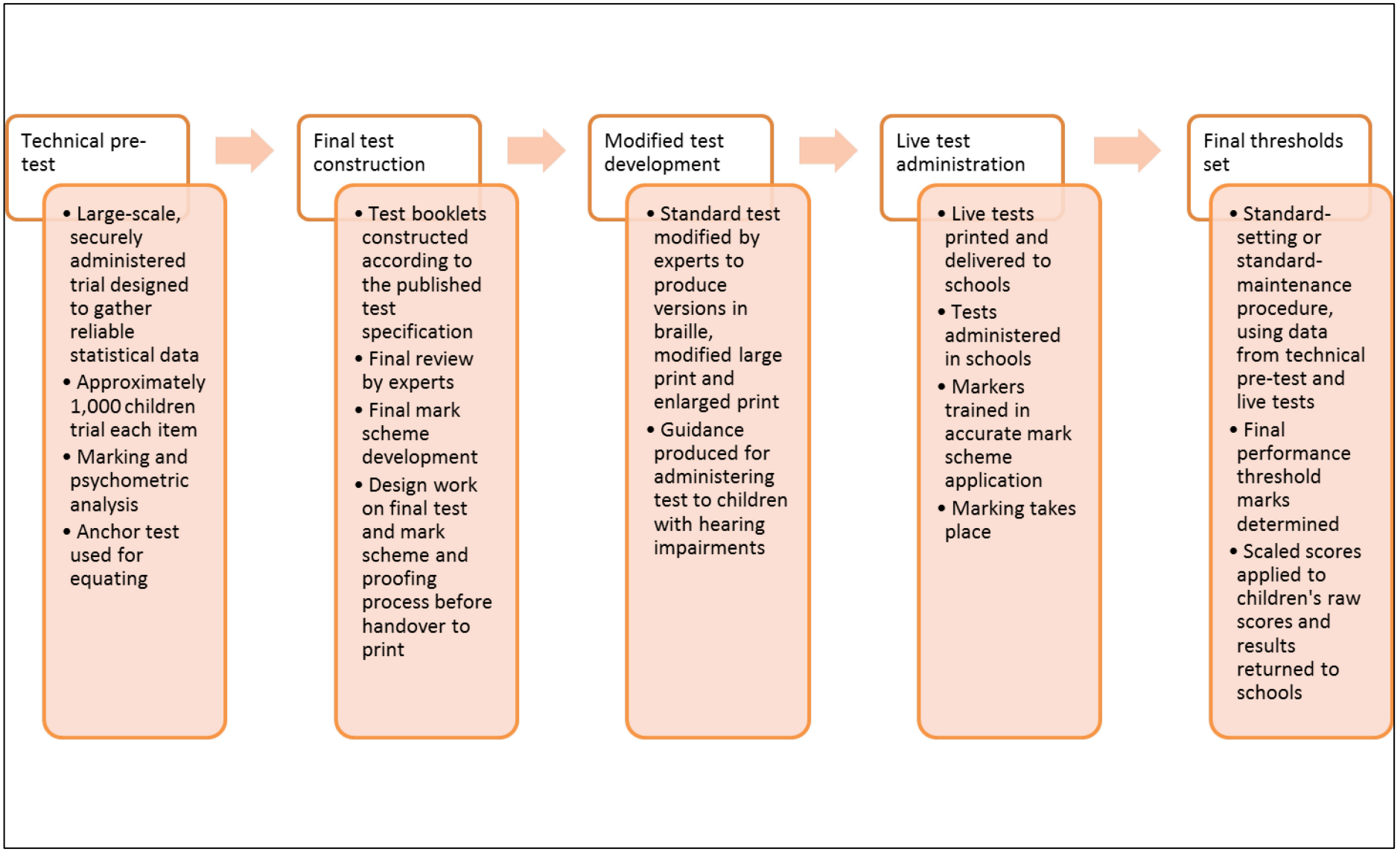

Figure 2: Overview of test development stages (continued)

3.3 STA’s item bank

Test materials and data are stored securely in a searchable database, known as an item bank, which can only be accessed by designated STA staff using DfE computers. The security of the item bank is one aspect of overall test confidentiality, which is essential for the integrity and validity of the tests.

STA developed its item bank database in August 2012. It was designed to improve the efficiency and effectiveness of the test development process, and stores all information about items in development.

An item-banking test development model means that items, questions or texts are trialled and held in a single repository, available for selection to meet the test specification. The use of an item-banking test development model contributes significantly to the validity and reliability of the tests:

- It enables STA to meet the test specification criteria precisely, because test papers are constructed using information from trialling.
- Items that do not perform well in trialling are not included in the live tests.
- Items can be selected in any combination to create the optimum test that meets the test specification.

In addition, over time, the item bank will hold a volume of items that are appropriate and ready for inclusion in a live test. Provided the item bank has sufficient materials, some stages of the process may not be needed each year, resulting in cost savings.

4 Item origination

This chapter outlines the process by which items (and texts for English reading) are initially developed. This first stage of development is usually outsourced to external suppliers, managed by STA staff.

4.1 Planning for item writing

STA conducts regular reviews of all materials in the item bank to ensure sufficient items are available to cover the full programme of study for each subject, taking into account all aspects of the test specification, such as coverage of the cognitive domain. STA decides whether to work with external suppliers or to proceed with internal item writing to provide the items required.

4.2 External item writing

STA holds a framework of item-writing agencies, who are external suppliers with experience of developing assessment material and educational resources. There are currently six item-writing agencies on the item-writing framework, which is renewed every four years.

Item-writing agencies from the framework may bid for work when invited by STA. The agencies are directed to publicly available documents such as test frameworks, past tests and sample materials to exemplify item format and design requirements. Information is also provided relating to the conventions of question wording. Further guidance about the scope for innovation and development of new or item formats may be provided to prospective item-writing agencies. The provision of this guidance ensures continuity of approach between years of testing and different item-writing suppliers.

Bids are scored according to cost and the technical requirements below:

1. expertise in item development
2. project management
3. example materials.

The contract is awarded to the bidder who achieves the highest overall score. After the award of the contract, the lead test developer for the project and a project manager work with the item-writing agency to oversee the materials being developed.

Through a series of supplier meetings, project plans and initial ideas for item writing are discussed, requirements are clarified and refined, and guidance is provided on draft items and mark schemes. The lead test development researcher for the project provides the main review of items, taking into account comments from curriculum advisors. The

project manager ensures the overall delivery of contractual requirements. It is through this dialogue with suppliers that the majority of guidance about item writing is provided as curricular or assessment issues arise.

There are two principal points of review during the item writing contract, during which all items under development will be scrutinised by the subject's test development team in consultation with curriculum advisors where appropriate. The first of these reviews takes place before the small-scale trial conducted in school by the item writing agency, and the second immediately after the small-scale trial (see section 4.5: Small-scale trialling). Following these reviews, specific revisions are requested to ensure the items meet the requirements of the test framework and national curriculum and are suitable for inclusion in national curriculum assessments.

For English reading, an additional text selection meeting is held during the early stages of the contract, in which test development researchers, curriculum advisers and teachers review proposed texts and agree which are suitable to be used. Proposed texts are also submitted to an external cultural review to ensure the subject matter is appropriate for all pupils. This provides evidence that texts included in the test are fair and accessible.

During the item-writing contract, design templates are provided to the supplier and a dialogue is maintained between the design team at STA and the item-writing agency.

Details of item writing agencies are provided in the accompanying technical appendix.

4.3 Internal item writing

Item writing is sometimes conducted within STA. This internal item writing may be targeted to address gaps in the balance of items in the item bank, the development of specific item types for research purposes, or the creation of items to address new requirements. For more straightforward areas of the curriculum, such as the arithmetic paper in mathematics, item writing is routinely undertaken internally.

Items are drafted by subject-specialist test development researchers. Depending on the nature of the item-writing project, they may do this with reference to specific published sources, international tests or research evidence. Sometimes a familiar format is reproduced, with changes made to the target word or context. Indicative mark schemes are developed alongside the items.

Once drafted, the items and mark schemes are subject to several stages of internal review. Some items will be considered suitable without changes; for others, amendments may be proposed. Some items are removed at this stage, which may necessitate further item writing. Following amendments to items, a draft is circulated to curriculum advisers, who provide comments via a written report, and subsequent amendments are made before the items proceed to a small-scale trial in schools, except for very straightforward items, such as spelling or simple arithmetic.

4.4 Initial mark schemes

Mark schemes are developed, reviewed and revised alongside the initial drafts of items. At this early stage, these mark schemes outline the general correct answer(s). Example pupil responses are gathered from subsequent large-scale trials to exemplify acceptable points or add acceptable responses to the mark schemes. Early development of the mark schemes is an important part of the process as it ensures effort invested in developing test items is not wasted because the item is subsequently deemed unmarkable.

4.5 Small-scale trialling

Where appropriate, items are trialled with a small number of pupils to provide an indication of the clarity of question wording and text accessibility. This provides the first small-scale evidence of validity for each item. Test development research staff conduct the trials of items originated in-house; item-writing agencies carry out the trials for the materials sourced externally.

Approximately 60 pupils of the target age group from at least two schools trial each test item. Consideration is given to variables such as school type, geographical location and attainment when selecting schools for trialling, so that – as far as is possible within a small sample – a range of characteristics is represented.

During the trials, pupils complete booklets of items. Some pupils are interviewed to gather their qualitative feedback. Teachers also comment on the materials.

After the trial, pupil responses are analysed for a variety of features such as the proportion of correct responses, omission rates, common responses and errors, and apparent misunderstandings of the item or text content. This small-scale qualitative information is communicated alongside the pupil and teacher feedback in an internal report, which gives further recommendations for amendments to items and the indicative mark schemes.

After these recommendations have been discussed with STA test developers and agreed amendments have been made, the final version of the item is submitted for inclusion in the item bank along with item classification tables, detailing item characteristics. The item then becomes available for potential selection for a larger-scale trial.

4.6 Item classification

At the end of the item origination process, items are securely stored within the STA's item bank database. Here they are assigned a unique identifier code and additional item metadata is stored, which is used to classify the items and monitor the characteristics of selected items against the test specification.

Item metadata may change as the item moves through the development process and any changes are recorded in the item bank. Further details of the classifications used can be found in the test frameworks[10].

[10] Available at: www.gov.uk/government/collections/national-curriculum-assessments-test-frameworks

5 Item design

Item design refers to an item's appearance and the conventions used to maintain design consistency within and between items.

5.1 Interfaces and processes

Design begins during item origination (see chapter 4: Item origination) and is an important consideration throughout the test development process because of its contribution to pupil experience, item accessibility and item validity, including the avoidance of construct irrelevant variance (see section 6.3.3: Construct irrelevant variance). Item-writing agencies produce items, including graphical elements, according to STA style and design guides. Items written in-house are typeset and designed by STA's test design team. Design guidance is given to item-writing agencies to ensure aspects of design, such as font use and layout, are applied to items as early as possible in the process and before being presented to pupils. In this way, item design is consistent through trialling and does not influence the answers pupils may give, contributing to the reliability of trialling data.

Following each expert review (see chapter 6: Question and item review), STA may amend items' wording, layout or graphical elements. The design team will carry out amendments to the items, with additional images commissioned from external illustrators as necessary. Because items may change significantly during development, flexibility and common standards are important. STA's designers work on materials using specific software to produce editable documents capable of being amended and changed at each test development stage.

Once amended, items are placed in booklets for use during in-school trialling and live testing. The booklet format is developed in conjunction with the print and logistics, marking and trialling teams to ensure materials are compatible with live test and trialling requirements and specifications. From the first stage of trialling (item validation trial), materials are produced as print-production ready and compatible with our live marking suppliers' scanning specifications, so they do not require amendment for use in live tests.

Mark schemes are developed as Word documents throughout the development process as this allows easy alteration by test development researchers. Prior to being used in live test administration and entering the public domain, mark schemes are typeset.

5.2 Key considerations

Item design can affect validity; items that are poorly laid out are difficult to comprehend and access. All STA test materials are developed to conform to the relevant design principles set out in the document, Fair access by design[11], namely:

- ensure the purpose of each task is clear, with due consideration given to readability and legibility
- ensure tasks address assessment criteria explicitly without unnecessary prescription
- avoid a requirement for pupils to demonstrate skills that are not essential to the subject being tested.

In addition, materials produced must be suitable for large-scale print reproduction, scanning, on-screen marking and final distribution online. STA uses colour swatches (using the Pantone Management System), fonts (using the Adobe Open Type font library) and design software (using Adobe Creative Cloud and Adobe PDF) to produce materials that meet the final print specifications.

Stylistic conventions (such as the use of bold, spacing of elements of an item, appearance of tables) are detailed in STA design guidelines so that they are consistent within and between the items of a particular subject, and consistent between subjects as far as is practicable.

Consistent use of writing style and design guidelines are also important for the item bank development process that STA uses (see chapter 3: The test development process). Consistent style and design allows items held in the item bank from different item-writing agencies or developed in different years to be brought together more easily.

5.3 Style and design guides

Style and design guidelines provide a basis for item development so that items from different item-writing agencies are consistent. The guidelines are active documents, updated to reflect changes in the use of language and design resulting from feedback from the trialling and development process.

5.4 Design of graphical elements

Items include a variety of graphical elements, including photographs, illustrations (including diagrams), tables and graphs. Photographs and illustrations can be produced

[11] Available at: www.nocn.org.uk/assets/0000/0683/Fair-Access-by-Design.pdf

by an item-writing agency or commissioned from an external design supplier. Photographs of pupils are only taken and used in an item when parental permission has been sought and obtained. Copyright of photographs and illustrations is always transferred to STA if they are commissioned from an external agent. Records of these copyrights are maintained and stored in case of future queries.

Tables and graphs are produced as an integral part of an item alongside the text and are subject to their own design constraints. The appearance of tables, graphs and diagrams is standardised within subjects and efforts have also been made to standardise them between subjects. Where differences exist, they are a consequence of subject-specific requirements.

5.5 Copyright

The use of third-party copyright materials is limited as much as possible. While the STA is able to use third-party copyright materials relatively freely for the purposes of assessment in accordance with Section 32(3) of the Copyright, Designs and Patents Act[12], final test materials, once used in live testing, are available for free to schools via the GOV.UK website under Open Government Licence (OGL)[13].

Under the terms of the OGL, materials are available for open use, which may infringe the copyright of third-party copyright holders. Therefore, where possible, images and illustrations are directly commissioned for the purpose of testing with the copyright transferred to STA.

Materials that cannot be directly commissioned have their third-party source recorded by the relevant subject team and acknowledged on the final test materials. This information is included in an annual 'Third-party copyright report' that must accompany the materials when released under OGL.

[12] Date of issue: 1988. Available at: www.legislation.gov.uk/ukpga/1988/48/contents

[13] Available at: www.nationalarchives.gov.uk/doc/open-government-licence/version/1/open-government-licence.htm

6 Question and item review

Item review is the process of checking content, wording and layout of questions and items to ensure appropriateness, validity, clarity and accuracy.

6.1 Stages of review

Items and their mark schemes are reviewed throughout the test development process by a wide team of internal and external parties to ensure they assess the national curriculum appropriately. Items are first reviewed through informal trialling during initial development, as described in chapter 4: Item origination. Following item origination, they enter the expert review process.

The expert review stage of the process involves independent quality assurance of test materials and is critical to the validity evidence that supports the tests. It takes place twice during the test development cycle and once at the end of the process when the final test is constructed (see section 6.1.6: External review at the end of the development process):

- expert review 1 – before the item validation trial
- expert review 2 – before the technical pre-test
- expert review 3 – after the live test is constructed.

The expert panels review and comment on the suitability of test items, identify possible issues and suggest improvements. The dates and composition of the panels who took part in the expert review of materials at each stage of development are detailed in the accompanying technical appendix.

Security is of critical importance during the expert review process and all those involved in the process are required to sign confidentiality agreements.

6.1.1 Panel definition and set up

Three separate expert panels review the items at both expert review 1 and expert review 2. These panels are recruited through recruitment emails to schools, local authorities, subject associations and universities.

The panels are updated with new reviewers regularly. Each panel is provided with guidance about the purpose of their review and the intended outcomes of the meeting.

The expert review panels are chaired by a test development researcher.

6.1.2 Teacher panel

Teacher panels are designed to gather feedback on how materials reflect current classroom practice. The panel comprises practising teachers, headteachers, or teachers who have recently left the profession. They have experience of teaching the target age group for the test, meaning that STA is gathering relevant and valid advice. The panel will consist of individuals of differing experience levels and from differing school types and geographical locations so that a range of views is gathered, representing – so far as is possible within a small group – the range of teachers and school types that will encounter the live tests. The panel is asked to comment on:

- how the items reflect current classroom practice
- whether the materials are of appropriate difficulty for the age group and ability range
- whether the materials are a suitable assessment of the programme of study
- the format, design and layout of items.

The teacher panel receives the materials for review in secure conditions on the morning of the meeting. The panel is given time to read and work through the items and mark up their copies. This process allows members to review the materials as they would on test day. Members are asked to comment on the suitability of draft materials and coverage of the curriculum and suggest any improvements that would make it a fairer assessment. After reading the materials, the panel works as a group to comment on each individual item.

6.1.3 Test review group

The test review group (TRG) comprises subject specialists, local authority advisers and multi-academy trust advisers, markers and practising teachers and headteachers who have previously served on teacher panels. Approximately one-third of TRG panel members are changed each time, so that each person stays on the panel for approximately three years.

The group comments on:

- how the items reflect classroom practice
- the technical accuracy of the content
- whether the materials are of appropriate difficulty for the age group and ability range
- whether the materials are a suitable assessment of the programme of study and the programme of study references are accurate
- the difficulty of test items in comparison with previous years.

Both test review group and inclusion panel (see section 6.1.4: Inclusion panel) members receive test materials, mark schemes and a report template at least one week before the

meeting takes place. They review each item, identify any issues and consider improvements in advance of the meeting, recording their comments on the report template. At the meeting they provide feedback on each item, listen to each other's perspective and attempt to reach a consensus on each item.

6.1.4 Inclusion panel

The inclusion panel is composed of experts in teaching pupils with special educational needs (SEN) and English as an additional language (EAL). Typically, the panel includes:

- an expert in teaching the visually impaired
- an expert in teaching the hearing impaired
- an educational psychologist
- those with expertise in teaching pupils with autistic spectrum disorders
- those with expertise in teaching pupils with dyslexia
- those with expertise in teaching pupils with EAL.

The panel reviews the accessibility of the standard test for pupils with this range of needs.

6.1.5 Decision making after external review

Resolution meetings are held shortly after expert review meetings 1 and 2. Resolution meetings are chaired by the senior test developer or test developer leading that particular test and are attended by other test development researchers or research assistants for the subject; the deputy director, head of test development research or head of assessment research and psychometrics; a psychometrician or senior psychometrician; curriculum advisers and a project co-ordinator. An assessment researcher may also attend.

The meetings consider all of the evidence from the expert review panels and resolve all the issues raised through the expert review stage, deciding what changes will be made to materials as a result. Evidence from item validation trialling may also be considered if it is available. Different reviewers' suggested changes may contradict each other. The lead test development researcher is responsible for deciding which recommendations to implement, taking all the evidence into account. If no solution can be found to an issue identified with an item, the item is likely to be removed from the test development process.

6.1.6 External review at the end of the development process

A single expert panel reviews the items at the end of the test development process (expert review 3). It comprises members of the teacher panel, test review group and inclusion panel. Expert review 3 is to review the suitability of the constructed test, with specific consideration for overall difficulty and breadth of coverage of the content domain.

At expert review 3, the panel receives the test booklets, mark schemes and guidance documents on the morning of the meeting. The panel is given time to read and work through the test booklets before providing whole-group feedback on the suitability of the test. Test materials are not sent out in advance of expert review 3. This is to ensure the live test remains secure and to allow panellists to review the materials as they would on test day. The dates and composition of the panels who reviewed materials at expert review 3 are detailed in the accompanying technical appendix.

6.2 Curriculum advisors

Curriculum advisors are recruited to support the quality assurance of items throughout the test development process. They work alongside STA's test development researchers to ensure the tests provide an accurate, valid and appropriate assessment of the curriculum. At least two curriculum advisors work on each subject and key stage to ensure a balance of feedback.

Curriculum advisors need to have:

- substantial and recent expertise in their chosen subject at the relevant key stage
- an understanding of the national curriculum, its structure and ongoing strategy of improvement
- an understanding of adjacent key stages within their chosen subject area, such as an understanding of key stage 1 and key stage 3 for a role that focuses specifically on key stage 2
- an understanding of the national curriculum with respect to teacher assessment and the exemplification of national standards that support reliability, validity and consistency
- an understanding of the issues being considered in any recent and forthcoming national curriculum and assessment reviews and how they may impact on the assessments
- the ability to communicate effectively with stakeholders, providing constructive feedback at meetings.

6.2.1 Procurement process

Curriculum advisors are appointed through the National College for Teaching and Leadership's Operational Associates framework. Anyone registered on the framework is eligible to apply for a curriculum advisor post. Applicants are required to respond to a series of questions to assess their skills and expertise. Responses are scored by three members of the relevant subject team within the test development research unit. Contracts are awarded to those with the highest overall scores.

6.2.2 Input to test development

Curriculum advisors typically review materials at the following stages of test development:

- initial drafts of questions and mark schemes from item-writing agencies
- expert review 1, including participation in resolution meeting 1 and project board 1 (see chapter 10: Governance)
- expert review 2, including participation in resolution meeting 2 and project board 2 (see chapter 10: Governance)
- final test review meeting, project board 3 and mark scheme finalisation meeting
- marker training materials development meetings
- standard setting.

The test development stages were summarised in chapter 3: The test development process.

6.3 Main areas for item review

The factors taken into account by STA and its reviewers are explained below.

6.3.1 Programme of study

Items are reviewed to ensure the programme of study references are accurate. These categorisations help monitor the overall balance of curriculum coverage in the tests.

6.3.2 Test specification

Items and tests are considered against the criteria laid out in the test specification, such as:

- item type
- cognitive domain
- context of questions.

6.3.3 Construct irrelevant variance

Construct irrelevant variance (CIV) occurs when pupils interpret an item in an unintended way. As a result, their performance on the item measures something other than that intended by the test. This means that the item can become unexpectedly easy or difficult and does not contribute to the aim of the test.

Reviewing an item for CIV involves looking for unintended ways in which a pupil might respond to the question and deciding whether these are valid misinterpretations of the item.

6.3.4 Distractors in selected response items

Selected response items (for example, multiple-choice items) require careful review of distractors (incorrect answer options). There are many aspects to be considered for distractors, which include:

- comparative length and complexity of language of distractors and correct answer options
- whether distractors are plausible in the context of the question, for example, that they work semantically and syntactically as answers to the question
- whether any of the distractors could be regarded as alternative correct responses.

6.3.5 Accessibility of questions

All reviewers consider the needs of pupils with SEN and EAL; some reviewers are specifically appointed for their expertise in one of these areas. STA applies the principles of universal design, where the standard versions of the tests are designed to be accessible to as many pupils as possible.

Modified versions of the tests are developed for pupils with certain types of SEN (see chapter 10: Governance). However, there are a significant number of pupils with SEN who will use the standard version of the tests.

6.3.6 Other considerations

In addition to the factors explained in detail above, reviewers take account of the following when considering the suitability of an item:

- accuracy
- clarity of question text
- formatting of item text
- clarity of instructions
- appropriateness of contexts
- clarity and accessibility of diagrams, photos and illustrations
- layout of questions
- number of marks
- appropriateness of difficulty level
- comparability with questions from previous year(s)
- compliance with house style.

7 Trialling

STA conducts trials with samples of pupils to assess whether items and their associated coding frames (see 7.5.1: Development of coding frames) are valid assessments of pupils' knowledge and skills. STA outsources the administration of test trials, including coding, to approved suppliers, but maintains overall management of the trials to ensure they are conducted in a robust, reliable and confidential manner. Information from trialling is used to make decisions about whether an item enters the next stage of the test development process.

This section provides an overview of the item validation trial (IVT) and technical pre-test (TPT). Informal trialling also occurs during item origination, as described in chapter 4: Item origination.

7.1 Item validation trial

An item validation trial (IVT) is the first large-scale trial in the test development process. The purpose of the IVT is to determine each item's suitability and general difficulty for the target age group and ability range. Qualitative and quantitative data are collected to inform the development of each item and its mark scheme.

For IVT, items are grouped into trial booklets that reflect coverage in the live test. Each IVT item is administered to approximately 300 pupils. The sample of schools that participate in an IVT is as representative as possible of the characteristics of the national cohort, in terms of school attainment and region. The dates and numbers involved in the IVTs are provided in the accompanying technical appendix.

Pupils taking the IVT in 2014 and 2015 had not studied the new national curriculum in full; this was taken into account when considering the outcomes of the trials.

There are three possible outcomes for each item following the IVT:

1. The item is progressed to the next stage of the development process with no changes.

2. The item is progressed to the next stage of the development process with amendments.

3. The item is archived because the evidence collected indicates that the item is not valid, reliable, or of an appropriate level of difficulty and cannot be improved.

7.1.1 IVT trial design

Each item usually appears once within the suite of IVT trial booklets. The trial booklets are constructed according to the test specification, as detailed in chapter 2: The tests.

An example of a trial design is presented in Figure 3. In the example, a collection of nine items is used to make three trial booklets, with each item appearing once in the entirety of the trial.

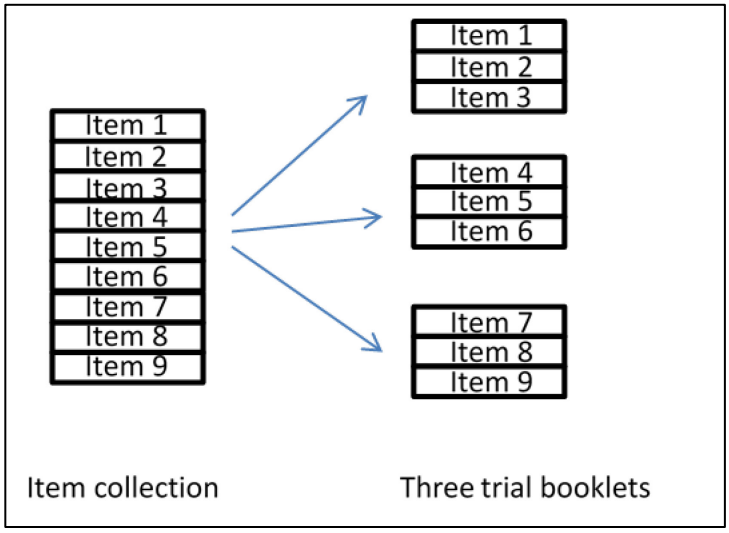

Figure 3 Item validation trial design

7.2 Technical pre-test

A technical pre-test (TPT) is the second, large-scale trial in the test development process. The purpose of the TPT is to gather detailed statistics to help support the final test construction process. Qualitative and quantitative data are collected about each item to determine its suitability for a live test, and the outcomes from the TPT are used to inform the final mark scheme construction. Furthermore, statistics are used to link the relative difficulty between tests from one year to the next as a comparability measure. This is achieved by including common 'anchor' items or booklets (see section 7.3.1: Anchor test) which are used in consecutive TPTs.

For TPT, items are grouped into trial booklets that reflect coverage in the live test. Each TPT item is administered to approximately 1,000 pupils. The sample of schools that participate in the TPT is as representative as possible of the national cohort of pupils, in terms of school attainment and region. The dates and numbers involved in the TPTs are provided in the accompanying technical appendix.

Pupils taking the TPT in 2015 would not have studied the new national curriculum in full; this was taken into account when considering the outcomes of the trial.

There are three possible outcomes for the items following a TPT:

1. The item is approved for use in a live test.

2. The question is not approved for use in a live test and is amended and retrialled.

3. The item is not approved for use in a live test and is archived because the evidence collected indicates that the item is not valid, reliable, or of an appropriate level of difficulty and cannot be improved.

All items that are included in all live tests are subject to at least one TPT. The TPT also included anchor items or booklets (see section 7.3.1: Anchor test) to link standards across years, with the exact approach varying between subjects.

7.2.1 TPT trial design

The technical pre-test trial design is more complex than that used for IVT; the same item appears in more than one booklet within the suite of trial booklets. This helps to quantify the effect of any differences in overall ability between the groups of pupils completing the different booklets.

Designing the trial so that items appear in more than one trial booklet also minimises the possibility that the performance of the item has been influenced by other items in the same booklet (item 'context') or by where the item appeared in the booklet (item 'position').

For the mathematics tests and the English grammar, punctuation and spelling tests, every item appears twice. For key stage 1 English reading, most items appear twice and for key stage 2 English reading, approximately one-third of the items appear twice. This is because the trial design for English reading, which is based around texts, means there are more limitations to the overall number of items that can be trialled.

Figure 4 shows an simplified example of a TPT design. This design facilitates the aim of each item appearing twice at different positions in the trial. The trial design dictates that any booklet within the trial is attempted by 500 pupils for each item to be trialled by 1,000 pupils in total.

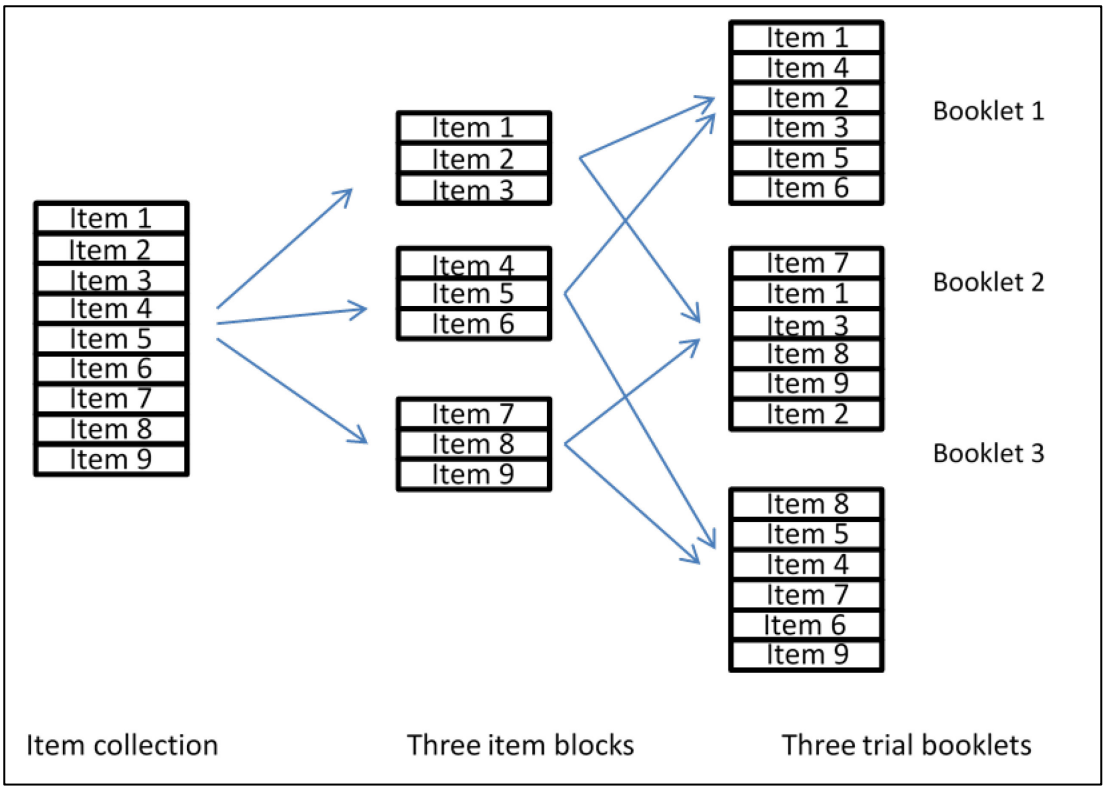

Figure 4. A sample TPT design

In Figure 4, a collection of nine items is divided into three item blocks and then the three blocks are used to make three trial booklets by combining any two blocks. Items appear twice in the trial booklets at different positions and in the context of other items. The TPT can vary in size depending on the number of items that need to be trialled.

7.2.2 Aims of the TPT for the 2016 test

As 2016 was the first year of assessing the 2014 national curriculum, the aims of the TPTs were broader than will be the case in future years. In addition to constructing a live test, anchor tests were created for key stage 1 and updated for key stage 2 to reflect the 2014 national curriculum.

7.3 Trial booklet construction

Trial booklets for IVT and TPT are constructed according to the chosen trial design and test specification for each subject.

For TPTs, statistical data gathered in the IVT is also reviewed to ensure the trial booklets are of an appropriate difficulty level.

Within each test booklet, items are generally ranked in difficulty order, with the exception of English reading where items appear according to the chronology of texts. However, at this stage it is not uncommon to have some easy items throughout the booklets. This can help ensure that low-attaining pupils are not discouraged too early on, but are motivated to continue even when the overall difficulty of the items may be increasing.

Booklets are designed and scrutinised to ensure:

- there is no overlap of items that may prompt or support the answer to other items
- each contains an appropriate range of different item types (as set out in the test specification)
- contexts that involve names of pupils include a mix of pupils of different gender and ethnicities.

7.3.1 Anchor test

Anchoring is the process used to maintain standards year-on-year by comparing the performance of each trial sample of pupils on a standard ('anchor') test or set of anchor items that are administered as part of every TPT. The performance on the anchor items is used to ensure equivalent relative difficulty between successive live tests. The anchor items are selected according to the test specification for each subject; they are representative of a live test.

The anchor items used in the TPT for the 2016 tests will be used in subsequent TPTs. Some of these were developed alongside the live test materials and went through the same development stages (see chapter 3: The test development process); in key stage 2, some anchor items from previous years that met the requirements of the new national curriculum were re-used.

Anchor tests are kept secure to ensure they will not have been seen by pupils involved in trialling in advance of the trial. English reading test TPTs for key stages 1 and 2 use an external anchor booklet. External anchoring of TPTs involves having an additional test (or test booklet) which is attempted by a specific sample of pupils during the TPT window.

An alternative to external anchoring is to include anchor items within the trial booklets (internal anchoring). This model was used for the TPTs in mathematics and English grammar, punctuation and spelling.

Anchor items need to be indistinguishable from the items being trialled, so that pupils approach all items in the same way. The anchor items are marked in the same way as trial items in the TPT.

Anchor items need to be reliable and the coding frames (described in section 7.5: Trial coding) must be unambiguous so they can be coded consistently from year-to-year. By keeping the administration and coding of the anchor test consistent, the performance of items should not vary considerably. If STA detects variance, test developers and psychometricians will investigate possible causes. Reasons could include:

- differences in the sample
- the position or context of an item in the test paper
- changes in the curriculum being taught

- unforeseen differences in administration or marking.

When differences are detected, the anchor item concerned is not used to link standards and may be replaced for future years.

7.4 Trialling agencies

Trialling agencies, procured from the DfE trialling agency framework, administer the trials. To be included on the framework, agencies must:

- meet security standards
- demonstrate sufficient resources and experience.

Details of the trialling agencies can be found in the accompanying technical appendix.

7.4.1 Recruiting trial schools

From 2016, participation in trials is a statutory requirement for schools, meaning the samples for the trials for the 2017 tests were designed to be representative of the population of assessing schools in England, by school attainment and region. However, during the development of the 2016 tests, schools could choose whether to participate. STA therefore provided the trialling agency with a specification of the samples required to be representative of key stage 1 or key stage 2 assessing schools in England.

Samples were stratified by school attainment and region. The trialling agency then contacted local authorities and schools to recruit a suitable sample. Once the samples of schools had been recruited, chi-square tests for each stratifier were conducted to test the representativeness of the samples and indicated that the achieved samples met STA's specification.

7.4.2 Trial administration

Trial administrators are recruited and trained by the trialling agency. The trialling agency also prints and collates the trial booklets and securely distributes these to the trial administrators shortly before the trial starts. Trial administrators conduct the trials in schools; administration is intended to mirror live test administration.

Questionnaires are used to collect qualitative information from teachers and administrators about aspects of the test booklets and their performance. Any incidents that occur during the trialling are also noted on the administrator questionnaire, so that these can be taken into account when reviewing data from the trial.

The trialling agency produces a final report for the trial. This is composed of:

- an administration report, detailing the achieved trial samples, the representativeness of the samples as per STA's specification, and summary of the trial
- a data report collating findings of the teacher and administrator questionnaire data.

7.5 Trial coding

During trialling, STA uses coding instead of marking for both the IVT and the TPT as it provides in-depth quantitative analysis on how items have performed. Pupils' scripts are not marked: they are 'coded'. This means that those reviewing pupil responses (coders) give a code to a specific response, so that test developers can identify the number of pupils in the trial giving that response.

Coding is therefore used to capture more detailed information on aspects of question performance than is possible using a standard mark scheme approach. At the analysis stage, these codes are converted into marks so the items can be scored. Coding can be done on paper or on screen.

A wide range of coding may be used, but some examples of coding and their purposes are outlined below:

- common, but incorrect, responses
- responses types that are borderline creditworthy or non-creditworthy
- creditworthy responses beyond the key stage / programme of study
- misconceptions STA wish to identify
- analysis of the performance of incorrect answer options in multiple-choice items.

7.5.1 Development of coding frames

During item origination, items are developed with initial mark schemes. STA then converts these to coding frames, which are refined during the early stages of the process (see section 7.5.3: Pre-coding).

7.5.2 Recruitment of trial coders

Trial coders are expected to have relevant teaching experience and / or experience of marking key stage 2. It is also desirable that the majority of trial coders have some experience in previous rounds of trial coding.

Each coding team is headed by a lead coder, supported by a deputy lead coder. The lead / deputy lead ('senior coders') are responsible for training coders and quality assurance of coding. They also deal with coding queries, passing these to test development researchers for a decision, where necessary.

7.5.3 Pre-coding

Before coding for each trial begins, pre-coding takes place, in which senior coders review the coding frames against a small sample of actual pupil responses from the trial. The purpose of pre-coding is:

- to test the draft coding frame on actual scripts and to find a range of responses that could act as exemplars in the coding frames
- to identify unexpected responses and to ensure they are catered for in the coding frame or coder training materials
- to make amendments to the coding frames ready for coder training
- to select scripts to develop any coder training materials
- to select practice scripts
- for TPT only, to select standardisation scripts and seeding or benchmarking scripts (see section 7.5.5: Quality assurance during coding).

For both IVT and TPT, a minimum of 100 scripts from the trial are provided for review at pre-coding. On-screen coding may allow far more than this to be available.

At the end of pre-coding:

- coding frames, training materials and practice materials have been agreed and finalised
- for TPT only, standardisation scripts and seeding or benchmarking scripts have been agreed and finalised
- where applicable, themed response tables have been checked for agreement with amended coding frames and finalised for use
- coding decisions are finalised and passed to the trialling agencies.

7.5.4 Training coders

The senior coders, trialling agency and STA work collaboratively to deliver coder training events, which contribute to coding reliability by making sure that all coders receive the same training.

On the training day, the trial coders familiarise themselves with the materials, with direction and support from their lead coder. They work through each item in the coding frame methodically, using the training exemplars selected at pre-coding. If any issues are identified, which were missed at pre-coding, the coding frames are amended to ensure consistent and reliable coding. Coders then apply the coding frames to the practice materials selected during pre-coding and, for TPT only, complete the standardisation scripts to ensure they are able to apply the coding frame accurately and consistently prior to the start of the actual coding.

7.5.5 Quality assurance during coding

Quality assurance processes are dependent on the mode of coding.

7.5.5.1 Paper-based coding (IVT)

At IVT, senior coders check one in five booklets of each coder's allocation. If coder error is found, the frequency of quality assurance by senior coders is increased and senior coders also provide the coder with additional guidance on accurate application of the coding frame for the affected question(s). In some instances, double coding is conducted as an additional check. Where necessary, the senior coders carry out targeted recoding to address problems that are identified. Test development researchers oversee the quality assurance process.

7.5.5.2 TPT coding

At TPT, the quality assurance process begins with coders completing standardisation items, for which the codes have been agreed at pre-coding. The lead coders check each coder's responses to ensure the coding frame has been applied accurately and to clarify any points of contention.

After the successful completion of standardisation, trial coding commences. As a quality assurance check, senior coders check 20% of each coder's allocation. If coder error is found, further training is given and the frequency of quality assurance increases.

For on-screen coding, seeding items are included in each coder's allocation. Like the standardisation items, these have been pre-coded and coders must allocate the same codes that were decided at pre-coding.

For paper-based coding, benchmarking scripts are included in each coder's allocation at a pre-determined point. Like the standardisation items, these have been pre-coded and coders must allocate the same codes that were decided at pre-coding. Senior coders feed back regularly to their teams to discuss issues as they arise. Coders are encouraged to raise any problematic responses to the lead and / or deputy lead coder for advice and discussion.

Coders could be stopped from coding all or part of a paper based on evidence from standardisation and ongoing quality assurance. This decision would be taken following discussion between the lead coder, test development researcher and trialling agency. Targeted re-coding is then undertaken as required.

7.5.6 Qualitative evidence collection

Coders are asked for feedback during and after coding to contribute to the qualitative evidence associated with each item. In particular, coders:

- highlight exemplars that raised particular issues for the coding frame so that they could be included in live training, or used to inform changes for the next step in the process
- feed back on the coding frame, training materials and items themselves
- feed back on processes to enable continuous improvement.

Some of the factors that STA considers when reviewing coder feedback include:

- Are there responses to a question that could be considered correct, but are not credited by the coding frame?
- Are there responses to a question that are difficult to code because they are not sufficiently addressed by the coding frame?
- Is there a mismatch between the item and coding frame?
- Is there sufficient differentiation of mark points in open response multi-mark items?
- Is the coding frame rewarding answers where it is not clear that the pupil fully understands the concept or skill being assessed?

This information will feed into decisions on whether the item proceeds to the next stage of test development unchanged or whether changes are required to improve the item.

8 Analysis of trialling data

Statistical analysis of the items is carried out following each trial to inform the test development process. The focus of this kind of analysis is to evaluate the performance of the items rather than the pupils.

From this analysis, STA gathers crucial, objective information about the difficulty of each item and how well it measures the test construct, as well as highlighting any potential flaws such as bias towards different groups of pupils or issues with the mark scheme. The types of analysis fall broadly under two areas of test theory: classical test theory (CTT) and item response theory (IRT).

8.1 Classical analysis

Most CTT-based analysis is aimed at identifying how much information the observed score tells us about an individual's true score (see section 15: Glossary). Classical analysis encompasses a range of statistics at item, booklet and test level.

Booklet or test level information includes score distributions, measures of internal consistency reliability and comparisons of mean scores by gender. Item level statistics include the mean score expressed as a percentage of marks, percentage of pupils achieving each possible score, percentage of pupils who omitted the item, percentage of pupils who omitted the item and all remaining items in the booklet. STA also examines the correlation between the item score and the total score, which has been corrected to exclude the item. This gives an indication of how well the item differentiates between pupils of differing ability.

8.2 Item response analysis

IRT refers to a family of statistical models in which the probability of success on an item is a function of various item parameters, depending on the model. IRT allows the comparison of items that have not appeared in the same test together and pupils who have taken different tests, by putting them all onto a single common scale, linked by common items or pupils. Due to its complexity, IRT is correspondingly more difficult to carry out, requiring specialist software.

Item response theory assumes there is an underlying latent construct being measured, that is continuous in nature. An item characteristic curve plots the probability of correct response as a function of the latent construct. Item characteristic curves are summed together to plot a test characteristic curve for each test (see accompanying technical appendix).

Another important concept in IRT is that of item information. This provides a display of item contribution to the latent construct and depends to large extent on the differentiation

power of the item along the latent construct continuum. Item information function plots can also be summed together to plot a test information function for each test (see accompanying technical appendix).

Item response theory has a number of assumptions that must be statistically tested to ensure its use is appropriate:

- the items in the test fit the IRT model
- the underlying latent construct is unidimensional
- the items in the test are independent of each other.

8.3 Distractor analysis

Pupil responses are coded for both trials rather than marked: numeric codes are assigned to represent how the pupil responded. Distractor analysis is used to analyse the response codes to provide more information about how the pupils answer the items. For example, in multiple-choice items, distractor analysis shows how often incorrect options were selected.

8.4 Differential item functioning

Differential item functioning (DIF) is a statistical analysis carried out to flag differences in item performance based on group membership. This analysis is useful when looking at all of the evidence for test construction, however, DIF can only indicate that there is differential item performance between groups (boy / girl, EAL / non-EAL) that have the same overall performance, it cannot determine the cause of the differential performance. Further qualitative exploration of an item leading to a reasoned, substantive explanation for the DIF is required before an item could be considered ‘biased’. It is important to acknowledge that group differential performance does not on its own indicate bias or lack of fairness. In all cases, there were no substantive reasons identified on items that flagged as having DIF (see accompanying technical appendix).

8.5 Analysing qualitative data

Qualitative evidence is gathered throughout the trial administration and trial coding stages. Feedback is evaluated alongside the data, recorded in the item bank and used to inform item and coding frame refinement.

8.5.1 Summary of types of evidence gathered

Throughout each trial, administrators have the opportunity to observe pupils taking the tests. They are asked to report any findings to STA via a questionnaire completed at the end of the administration window.

Teachers whose classes participate in the trial also have the opportunity to review the materials and observe their pupils sitting the tests. They also complete a questionnaire.

The questionnaires provide an opportunity for those involved in the trialling to give their opinions on the suitability of the questions in terms of accessibility, ease of reading, clarity and layout, manageability and timing.

Pupils' comments also feed into the trial administrator and teacher questionnaires.

8.5.2 Revising items and coding frames

Qualitative data can support information provided by quantitative data and bring to light issues with accessibility or reasons why pupils may have misinterpreted a question.

However, because it is subjective and because of the relatively small numbers responding to the questionnaires, some caution is required when using qualitative data to make judgements on whether changes should be made to questions and coding frames.

9 Test construction

Test construction is the process by which questions are put together to form test booklets that meet the test specification set out in the test framework (see chapter 2: The tests).

Test development researchers and psychometricians are responsible for constructing the tests based on the evidence from trialling. After the live tests have been constructed, they are reviewed at expert review 3 and by curriculum advisors to ensure the tests are fit for purpose and meet the test specification.

9.1 Item finalisation

All questions considered for a live test have been through a technical pre-test, which has provided evidence that they are functioning well. Items going forward to the live test should not be amended after the technical pre-test as this could affect pupil performance on the item and render the data held unreliable.

An item finalisation meeting is held following each trial. At this meeting, test developers and psychometricians confirm which items from the trial are suitable for inclusion in the next round of trialling or in a live test. Items that are deemed suitable must:

- have appropriate facilities (difficulty levels)
- differentiate between groups of pupils of differing ability
- be accessible in terms of language, layout and illustrations.

Both quantitative and qualitative data from the trial is reviewed to give a report on each item's reliability and validity as an appropriate assessment for its attributed programme of study references.

Any items that do not function well or that had poor feedback from teachers or pupils are either removed from the pool of available items for selection or amended to be re-trialled for future use.

Occasionally, it is not possible for the items in the live test to be identical to the trialled TPT versions. Small changes may be required at the test construction phase, for example, to keep the gender and ethnicity of pupils named in the papers balanced or to make sure each test paper carries the correct total number of marks.

While there is a small chance this could affect the performance of a question in the live test compared to its performance in the technical pre-test, it is unlikely to be significant and is monitored through the live analysis. Such changes are only made when strictly necessary.

9.2 Constructing the test

STA's psychometricians use question and item metadata to construct a test that meets the test specification and which optimises the measurement precision within the appropriate range on the ability scale. A test construction meeting is then held to select the items for IVT, TPT or live test booklets. At this meeting, test developers use the psychometricians' initial selection as a starting point. The meeting's participants consider the proposed booklets, taking into account item type, presentational aspects, question contexts and coverage, and whether there are any conflicts between what is assessed within test booklets and across the test as a whole. At this stage, items may be swapped in or out of the test to improve its overall quality and suitability.

Once the questions have been chosen for each test, they are put in an appropriate order. For mathematics and English grammar, punctuation and spelling, more accessible questions are put at the start of the test and more demanding questions towards the end, so the test paper increases in difficulty. For English reading, although an attempt is made to put easier questions earlier in the test, the questions are required to follow the chronology of the texts and so some easier questions will appear later in the booklet. The tests usually start with one or two easy questions to allow pupils to familiarise themselves with the context and conditions.

The overall test is reviewed in the combination of booklets that will form the complete test. This is to check for overlaps (for example, in the English grammar, punctuation and spelling tests, it would not be appropriate for one of the spelling words to also appear in the question booklet).

9.3 Reviewing and confirming the test

The tests are subject to a further review process after test construction. They are reviewed first by the test development team and the psychometrician to ensure none of the test constraints have been missed during the test construction meeting.

External stakeholders then review proposed live tests at expert review 3. Details of this panel are provided in chapter 6: Question and item review and the details of those who participated in expert review 3 are provided in the accompanying technical appendix.

Live tests are then presented to project board 3 (see chapter 10: Governance). Once the project board is satisfied that the tests have met the test specification and are of sufficient quality, the tests are signed off to go into production.

9.4 Quality assurance and proof reading

The tests go through a rigorous quality assurance process before the materials are signed off to print. There are three major handover stages. At each stage, the tests and

mark schemes are proofread by a range of people to ensure all the material is accurate and error-free. Between each stage, any errors found are corrected and rechecked.

The tests and mark schemes are proofread by people with particular areas of expertise, including professional editorial proofreaders and subject experts. They make sure that:

- there are no typographical errors
- house style is applied consistently
- grammar is correct and reflects the grammatical rules and conventions tested in the English grammar, punctuation and spelling tests
- ISBN numbers, product codes and barcodes are correct
- requirements are met for the scanning of papers, data capture and marking
- there is consistency across subjects and key stages
- there are no content overlaps
- the tests give pupils all the information they need to answer the questions
- test content is factually correct and would stand up to scrutiny
- mark schemes and tests reflect each other
- the mark schemes are usable, correct and adhere to established conventions
- the number of marks is correct
- items are labelled correctly.

All of these checks occur at least once throughout the handover stages. Test development researchers also proofread the materials at every stage and have overall responsibility for the quality of the tests. They are responsible for collating the comments from each proofing round and ensuring any amendments are made.

10 Governance

Overall governance of each test in development is managed through fortnightly checkpoint meetings involving the whole project team for that subject, plus relevant members of STA's senior management team. Outcomes of these meetings are reported into monthly delivery and technical sub-programme boards as appropriate, chaired by a Deputy Director and attended by members of STA senior management team. Escalation is through normal STA routes, including the STA Risk and Security Committee, to the Executive Management Board.

In addition, project boards take place at three points in the test development process: before item validation trial, before technical pre-test and before the materials go live. The purpose of the project board is: to review documentation on the quality of the items and tests; to ensure there is sufficient material available to carry on to the next stage of development, and for the tests going live; to ensure the correct process has been followed, the tests meet the specification and the materials are fit for purpose.

The project board will consider the evidence carefully before approving the test materials for trial or live testing.

Statutory national testing is regulated by Ofqual, whose representatives may observe sub-programme or project boards.

10.1 Project board 1

Project board 1 takes place shortly after the resolution meeting for expert review 1 (see chapter 2: The tests, for a summary of the test development process). The purpose of the meeting is to formally approve materials proceeding to the item validation trial.

At project board 1, the following evidence is considered:

- trial booklets
- a summary of the process to this point
- an overview of evidence from expert review 1 and any feedback from the curriculum advisors on the materials
- coverage of items in the trial against the test framework
- coverage of items within the item bank to ensure areas not covered in the trial are provided for in the item bank.

10.2 Project board 2

Project board 2 takes place shortly after the resolution meeting for expert review 2. The purpose of the meeting is to formally approve and sign off materials for trial in the technical pre-test.

At project board 2, the documentation comprises of:

- trial booklets
- an outline of the technical pre-test model being used
- a summary of the process to this point
- an overview of evidence from expert review 2 and the curriculum advisors
- coverage of materials in the trial against the test framework
- coverage of items within the item bank to ensure areas not covered in the trial are provided for in the item bank; in total, the materials in the trial and the item bank must be able to produce a live test that meets the specification.

10.3 Project board 3

Project board 3 takes place shortly after expert review 3. The purpose of project board 3 is to formally approve the materials going forward as the live test. Any recommendations for changes to the constructed test following expert review 3 must be made and agreed at project board 3.

Project board 3 reviews the validity evidence for the test, including whether the correct process has been followed and that the proposed test meets all aspects of the specification. If these conditions are met, the materials can be approved as fit for purpose.

The dates of the project boards are given in the accompanying technical appendix.

10.4 Standards confirmation and maintenance

The final major governance meeting of the process is the standards confirmation or standards maintance meeting, where the scaled score conversion tables are signed off for the tests (see chapter 13: Standard setting). The dates of these meetings are in the accompanying technical appendix.

11 Test administration

The test administration team in STA is responsible for developing test guidance and managing the processes related to test administration.

The test operations division of STA oversees the print, logistics, return of paper-based national curriculum tests including marking of key stage 2 tests and the provision of the downloadable key stage 1 English grammar, punctuation and spelling test through National Curriculum Assessment (NCA) tools.

11.1 Test orders

Maintained schools, academies and free schools are not required to place a test order for standard versions of the tests. Quantities of standard key stage 1 English reading and mathematics test materials are sent to schools based on their autumn census data.

Quantities of standard key stage 2 test materials are sent to schools based on their census and pupil registration data. Schools can order modified versions of the tests on NCA tools, if required, by the end of November. Modified versions of the key stage 1 tests are available in modified large print and braille. Modified versions of the key stage 2 tests are available in enlarged print, modified large print and braille.

Independent schools choosing to participate in the key stage 1 and / or key stage 2 tests must place test orders on NCA tools for both standard and modified versions, and issue privacy notices to parents, by the end of November.

Details of the amount of test materials sent to schools are provided in the accompanying technical appendix.

11.2 Pupil registration for key stage 2 tests

All pupils enrolled at maintained schools, including special schools, academies and free schools, who will complete the key stage 2 programme of study in the current academic year, must be registered for the tests on NCA tools. This includes pupils who are working below the overall standard of the tests and ultimately won't take them, and pupils who are working at the overall standard but cannot access the tests.

There is no pupil registration process at key stage 1.

11.3 Delivery of test materials

Test materials are delivered to school addresses taken from the Get information about schools website[14]. Schools receive all key stage 1 and key stage 2 test materials, including any modified test orders, in April.

The headteacher, or a delegated senior member of staff, must check the contents of their delivery against the delivery note to ensure the correct number and type of test materials have been received. Key stage 1 test papers must not be opened until the school is administering the test for the first time. Key stage 2 test papers must not be opened until the test is about to be administered on the day specified in the test timetable.

Headteachers must ensure the security of the tests is maintained so that no pupil has an unfair advantage. Schools must follow the guidance on keeping test materials secure and treat them as confidential from the point they are received in school until the end of the relevant test administration period.

11.4 Test administration

Schools must administer the key stage 1 tests in English reading and mathematics during May. The tests do not have set days for their administration, and they may be administered to groups of pupils on different days. Pupils must only be allowed to take each test once.

All key stage 2 tests must be administered on the days specified in the statutory timetable in May. Headteachers are responsible for deciding the start time of the test, but all pupils should take each test at the same time. Tests must not be taken before the day specified in the statutory timetable.

If it is not possible for all pupils to take a key stage 2 test at the same time on the day specified in the statutory timetable, schools must notify STA of a start-time variation. A start-time variation allows an individual pupil, or part of the cohort, to take the test on the same day but at a different time from the rest of the cohort. Schools must complete the notification on NCA tools before the test begins, but do not need a response from STA to proceed at the nominated time.

If it is not possible for all pupils to take a key stage 2 test on the day specified in the statutory timetable, schools must apply for a timetable variation on NCA tools. If approved by STA, a timetable variation allows an individual pupil, or part or whole of the

[14] Available at: https://get-information-schools.service.gov.uk/

cohort, to take the test up to five school days after the original day of the test. Schools must wait for approval from STA before beginning the test on a new day.

Key stage 1 tests are marked internally within schools. Teachers use the results of the English reading and mathematics tests to help make a secure judgement for their final teacher assessment (TA) at the end of key stage 1. The tests make up one piece of evidence for overall TA. There is no requirement for schools to administer the optional English grammar, punctuation and spelling test or use the result to inform TA.

Schools must send all key stage 2 test scripts for external marking. Test administrators should return test scripts to the headteacher immediately after each test. Headteachers are responsible for making sure the school's completed test scripts are immediately collated, packed and sealed correctly. All test papers must be collected, ensuring every pupil is accounted for. Key stage 2 test results are returned to schools at the beginning of July.

Full administration details are published in the assessment and reporting arrangements (see section 11.6: Communications and guidance).

11.5 Monitoring visits

Monitoring visits, on behalf of the local authority (LA) or STA, are made, unannounced, to a sample of schools administering the tests. They will check whether the school is following the published test administration guidance on:

- keeping the test materials secure
- administering the key stage 2 tests
- packaging and returning key stage 2 test scripts.

The monitoring visits are part of STA's assurance that the tests are being administered consistently in all schools and support the validity evidence in relation to standardisation.

If a school receives a monitoring visit, they must allow visitors to:

- see all key stage 1 and key stage 2 test materials, and any relevant delivery notes
- observe any key stage 2 tests being administered
- see evidence to show that pupils using access arrangements, for example, prompters, scribes or readers, are doing so in accordance with normal classroom practice
- see copies of correspondence and other documents sent to, and received from, the LA or STA about the administration of the tests.

STA will carry out a full investigation if a monitoring visitor reports:

- administrative irregularities

- potential maladministration.

These investigations are used to make decisions on the accuracy or correctness of pupils' results.

11.6 Communications and guidance

STA publishes guidance throughout the test cycle to support schools with test orders, pupil registration, keeping test materials secure, test administration and packing test scripts. This guidance is developed to ensure consistency of administration across schools and therefore supports the validity evidence in relation to standardisation.

This guidance[15] includes:

- Key stage 1: Assessment and Reporting Arrangements (ARA)[16]
- Key stage 2: Assessment and Reporting Arrangements (ARA)[17]
- Key stage 1: Test Administration Guidance[18]
- Key stage 1: Modified Test Administration Guidance[19]
- Key stage 2: Test Administration Guidance[20]
- Key stage 2: Modified Test Administration Guidance[21]
- Keeping test materials secure[22]
- Attendance register and test script dispatch instructions[23] (Key stage 2 only)
- Monitoring visits guidance[24]
- Assessment updates[25] (emailed weekly to schools and stakeholders during term time)

[15] The guidance documents linked in this section are the most recent versions of the documents. Similar documents were provided in 2016 and 2017, though these versions are no longer available online.

[16] Available at: www.gov.uk/government/publications/2018-key-stage-1-assessment-and-reporting-arrangements-ara

[17] Available at: www.gov.uk/government/publications/2018-key-stage-2-assessment-and-reporting-arrangements-ara

[18] Available at: www.gov.uk/government/publications/key-stage-1-tests-test-administration-guidance-tag

[19] Available at: www.gov.uk/government/publications/key-stage-1-tests-modified-test-administration-guidance-mtag

[20] Available at: www.gov.uk/government/publications/key-stage-2-tests-test-administration-guidance-tag

[21] Available at: www.gov.uk/government/publications/key-stage-2-tests-modified-test-administration-guidance-mtag

[22] Available at: www.gov.uk/government/publications/key-stage-2-tests-and-phonics-screening-check-keep-materials-secure/guidance-on-the-security-of-key-stage-2-tests-and-phonics-screening-check-materials

[23] Available at: www.gov.uk/government/publications/key-stage-2-attendance-register-and-test-script-dispatch

[24] Available at: www.gov.uk/guidance/key-stage-2-tests-and-phonics-screening-check-monitoring-visits

[25] Available at: www.gov.uk/government/collections/sta-assessment-updates

- Videos and webinars[26]

11.7 Access arrangements

Some pupils with specific needs may need additional arrangements to be put in place so that they can take part in the key stage 1 and key stage 2 tests. Access arrangements are adjustments that can be made to support pupils and ensure they are able to demonstrate their attainment. This supports STA in determining that the tests are fair for all groups of pupils. Headteachers and teachers must consider whether any of their pupils will need access arrangements before they administer the tests.

Access arrangements should be based primarily on normal classroom practice and they must never provide an unfair advantage to a pupil. The support given must not change the test questions and the answers must be the pupils' own. Failure to apply for, or administer, access arrangements appropriately can result in a maladministration investigation at the school and pupils' results can be annulled.

Access arrangements can be used to support pupils:

- who have difficulty reading
- who have difficulty writing
- with a hearing impairment
- with a visual impairment
- who use sign language
- who have difficulty concentrating
- who have processing difficulties.

When planning for the tests, schools are advised to think of any needs their pupils have and whether they receive additional support as part of normal classroom practice. However, some pupils may not be able to access the tests, despite the provision of additional arrangements.

During a monitoring visit for the key stage 2 tests, local authorities may ask to see evidence that any additional support provided in the tests is part of normal classroom practice. Evidence will vary according to the type of arrangement and the key stage it is required for. Evidence may include notes recorded in teaching plans, individual pupil support plans or a pupil's classwork to demonstrate the type of support provided in the classroom.

[26] Available at: https://registration.livegroup.co.uk/sta

Schools are not required to make either applications or notifications for access arrangements for the key stage 1 tests. They may be used at the discretion of the headteacher. However, schools may be asked to provide evidence that any support given is part of normal classroom practice as part of a TA moderation visit.

Some access arrangements for the key stage 2 tests must be applied for in advance:

- early opening
- additional time
- compensatory marks.

There are some access arrangements that do not require an application, but schools are required to notify STA about their use:

- use of scribes
- use of transcripts
- use of word processors or other technical or electronic aids.

Schools are required to submit notifications for these arrangements after all the tests have been administered.

Other access arrangements may be put in place without prior approval or the need to notify STA. However, the use of these arrangements must reflect normal classroom practice.

These arrangements include:

- readers
- prompters
- rest breaks
- written or oral translations
- apparatus in mathematics tests
- modified test papers.

Full details of each of the access arrangements permitted in the tests is available from these links:

- Key stage 1: how to use access arrangements[27]
- Key stage 2: how to use access arrangements[28]

[27] Available at: www.gov.uk/guidance/key-stage-1-tests-how-to-use-access-arrangements

[28] Available at: www.gov.uk/guidance/key-stage-2-tests-how-to-use-access-arrangements

11.8 Modified test papers

STA develops modified versions of the tests, which are primarily designed for pupils with significant visual impairments, although they may be suitable for pupils with other needs, such as dyslexia.

At key stage 1 the test papers are available in:

- Modified Large Print (MLP)
- braille.

The size and type of font used in the standard version of the key stage 1 tests have been designed to be more accessible to pupils with visual impairments, so enlarged print (EP) versions are not produced.

At key stage 2 the test papers are available in:

- EP
- MLP
- braille.

EP versions are produced in a larger format: booklet and all text, pictures, and non-scaled diagrams are larger than the standard versions.

MLP versions are also in the larger format, but more white space is present. Some diagrams are substituted for a high contrast design or require the use of physical models.

Braille versions of the test, available in Unified English Braille (UEB), are suitable for pupils with extremely limited or no vision. Diagrams are produced in tactile formats or as physical models. The MLP and braille versions of the test are developed by a specialist modified test agency on behalf of STA.

The modified test agency also provides advice to schools about which modified materials may be suitable for pupils. Details of the number of modified test papers sent to schools are provided in the accompanying technical appendix.

12 Marking of the key stage 2 tests

Each year, approximately 580,000 pupils complete the key stage 2 national curriculum tests. The tests for mathematics, English reading and English grammar, punctuation and spelling are marked by a team of over 4,000 markers. STA outsources the marking of key stage 2 tests to a specialist supplier, but maintains overall responsibility for the marking process to ensure it is conducted in a robust, reliable manner.

12.1 Mode of marking

The majority of key stage 2 tests are marked on screen, with approximately one per cent marked on paper, as they cannot be scanned. Items are classified as either expert, standard or clerical, based upon the complexity of the item response required. Each item type is marked by markers with the appropriate level of expertise.

Markers are trained to mark one item type. Expert and standard markers mark on screen at home. Clerical markers mark at a marking centre.

12.2 Management of marking

The marking team work in a hierarchical structure, with separate teams for each test subject. Each subject team, for example English reading, is headed by a senior marking team who are responsible for developing the training and standardisation materials and for the delivery of marker training to ensure markers can apply the mark scheme correctly and mark accurately. They are also responsible for ongoing quality assurance of live marking.

A quality assessor oversees the quality of the development of assessment materials and delivery of training for each subject.

12.2.1 Expert and standard marking hierarchy

Senior marking team: Overall responsibility for the subject. Responsible for developing marker training and quality assurance materials and for conducting quality assurance throughout the marking cycle.
Supervisors: Support delivery of marker training. Supervises a team of markers throughout the marking cycle. Quality assure marking and provide ongoing feedback on accuracy of marking.
Markers (expert and standard): Complete marking for allocated items.

Table 1: Expert and standard marking hierarchy

12.2.2 Clerical marking hierarchy

Senior marking team (as above): Overall responsibility for the subject. Responsible for developing marker training and quality assurance materials.
Marking manager: Responsible for the set-up, training, overall running and successful completion of marking for all items classified as clerical.
Clerical marking supervisors: Responsible for training, supervising and quality assurance for a group of markers.
Clerical marker: Completes marking for allocated items.

Table 2: Clerical marking hierarchy

All clerical marking is carried out at a marking centre, overseen by a marking manager. Clerical markers mark across all subjects. Each subject's senior marking team work with the marking manager to develop training and quality assurance materials for clerical marking.

12.3 Marker recruitment

Markers are recruited based on their experience and, if they have marked key stage 2 tests previously, the quality of their marking. Expert and standard markers are all required to have qualified teacher status and teaching experience in the relevant subject and/or key stage. Clerical markers, who mark items requiring no professional judgement, are graduates or graduands, but are not required to have teaching experience in the relevant subject.

12.4 Development of marker training materials

12.4.1 Expert and standard items

Marker training materials are developed to ensure markers have a clear understanding of the key stage 2 mark schemes and can consistently apply the correct marks to pupil responses for each item. The training materials are developed using exemplar pupil responses from the technical pre-test of the items.

The training materials are developed by the senior marking team and curriculum experts in collaboration with STA test development researchers and psychometricians and the STA marking team. Through a series of meetings, the training materials are refined to

ensure they accurately illustrate the key marking points from the mark scheme. They include correct and incorrect pupil responses to ensure markers fully understand which responses are creditworthy, which are not, and why. During the meetings, selected pupil responses are reviewed and assessed for suitability and to ensure the response has been marked correctly based on the mark scheme.

Further detail on how these different responses are introduced in the marker training is provided below.

STA's role during the development of marker training materials is to ensure the training materials are in line with the intention of the key stage 2 tests and mark schemes and to approve the materials.

Training responses

Training responses are used to train markers on the key marking points from the mark scheme. Pupil responses are selected to illustrate the variety of ways a pupil could answer a question and how the response should be marked.

Practice responses

Practice responses are used by markers to practise marking following their training so they can be given further feedback on the accuracy of their marking before attempting standardisation. Practice responses include examples of the range of answers given by pupils during trialling.

Standardisation responses

Standardisation responses are selected to make sure markers are tested on their marking of a range of responses in order to verify they:

- have understood the mark scheme and correctly apply the marking principles
- can use their professional judgement within the context of the mark scheme
- can accurately mark to the standard required.

For each item, a standardisation set of between five and eight pupil responses is created. To pass standardisation for an item, markers have to match the correct mark for the responses within the standardisation set for that item. For a very small number of items requiring a high degree of professional judgement, a small tolerance is applied across the standardisation set.

All supervisors and markers have to complete item level standardisation successfully before they are permitted to start 'live' marking.

Supervisors have to pass standardisation for all items their teams mark.

User Acceptance Testing (UAT) of the training materials

The training materials are tested before they are finalised. Qualitative and quantitative feedback from the UAT process is used to refine the training materials further before they are used to train all markers.

12.4.2 Clerical items

Training materials for clerical items are selected by the senior marking team. Materials are reviewed and approved by STA test development researchers to ensure they accurately illustrate the key marking points from the mark scheme.

12.5 Marker training

Supervisors are trained to mark expert and standard items. General marking supervisors are trained to mark clerical items. Markers are trained to mark either expert, standard or clerical items depending on their role. All training is delivered face-to-face and led by a supervisor.

12.6 Practice and standardisation

After training, supervisors and expert and standard markers mark practice items. This is their first opportunity to mark items independently and receive feedback from their supervisor. This is followed by completing standardisation responses for each item.

Clerical markers are not required to complete practice or standardisation as clerical items have one possible correct response and no expert knowledge or judgement is required to mark them.

12.7 Quality assurance of live marking – all item types

Ongoing marking quality for all markers is assessed through the inclusion of validity responses in a marker's allocation. The validity responses selected exemplify the marking principles included in the mark schemes. They are given a pre-determined mark in collaboration between the senior marking team, STA test development researchers and the STA marking team. The markers cannot identify which responses are validity items.

If a marker does not perform to the pre-determined standard, they are stopped from marking. Any completed marking considered at risk of being incorrect would then be re-marked.

12.8 Marking of modified and unscannable test scripts

Modified tests include large print, modified large print and braille test scripts. Standard tests can be unscannable for a variety of reasons, such as photocopied test papers. Modified and unscannable test scripts are marked at a central marking panel consisting of supervisors who have been trained and have passed standardisation for all test items.

Marking quality assurance for these scripts is also the responsibility of the senior marking team. A selection of the scripts marked by each marker is reviewed to ensure the marking is within the tolerance set by STA. Immediate feedback is given where necessary and any marking considered to be incorrect is re-marked. Any marker who fails to meet the agreed tolerance would be stopped from marking and their marking re-marked by another marker.

A marker who can read braille and has been trained, and passed standardisation for all items, marks the braille tests. Where this is not possible, a transcript is made of the braille test script and the script is marked by a supervisor. Quality assurance checks of both the scripts marked in braille and of the transcriptions made from the braille scripts are carried out to ensure marking and transcripts are free from errors.

12.9 Marking reviews

Schools can apply for a review of marking if they believe there is evidence that a mark scheme has not been applied correctly or a clerical error has occurred. For the 2016 and 2017 tests, a review application was deemed to be successful if it resulted in a change of three or more marks or in a change to the achievement of the expected standard. The criteria for a successful review is reviewed annually.

Test scripts that are marked on screen are generally reviewed on screen. Test scripts that are marked on paper are reviewed on paper.

The review marker reviews all the pupil's test paper(s) for the subject for which the school has requested a review. All marks awarded are reviewed to check if the original application of the mark scheme was accurate. As a result of the review, marks are amended if the original marking was not in line with the published mark scheme. Review marking is managed through the supervisory hierarchy in place for marking the key stage 2 tests.

12.9.1 Development of review training materials

Review marker training materials and re-standardisation materials are developed by the senior marking team. The materials do not introduce new marking principles, but are used to remind review markers of the mark scheme and their previous training. This ensures they can consistently and accurately review the marks awarded. The materials

focus on themes identified from review applications made by schools and are based on exemplar pupil responses from the tests.

All materials are signed off by STA test development researchers.

12.9.2 Review marker recruitment

Experienced supervisory markers who have achieved the highest quality marking standards throughout the marking period are recruited to complete review marking.

12.9.3 Review marking process

All review markers are trained using the original training materials for all item types and review training materials and are required to pass re-standardisation for all items before starting review marking.

Review markers are supervised by senior members of the marking hierarchy. Their review marking is sampled to ensure they are correctly applying the mark scheme and that marking quality is maintained. If a review marker does not apply the mark scheme as required, or does not adhere to the correct review procedures, they would be stopped from review marking. Their completed review marking would then be checked again by another review marker.

13 Standard setting

A standard setting process was conducted after the live 2016 administration of key stage 1 and key stage 2 national curriculum tests to set the initial standard on the tests. Standard setting is only conducted when a new assessment has been introduced. All future administrations focus on maintaining the standard set on these initial tests.

Data from the live administration is necessary to ensure the standard can be set in the right place. The performance descriptors, developed as part of the test framework (see chapter 2: The tests) also played an integral part in the standard setting process.

13.1 Data used for standard setting

Since test data is not routinely collected for key stage 1, a representative sample of schools were required to provide pupil responses from an early data collection exercise that took place in April 2016. The schools involved in the key stage 1 standard setting sample were only required to submit data from one subject; the testing of the other subjects was conducted during the standard key stage 1 May window.

Three stratified random samples of approximately 700 schools were selected to take part in the early administration of the key stage 1 tests in April 2016 in order for data to be available for standard setting. At least 25,000 pupils took each of the tests during this data collection exercise. The tests were marked by teachers in the schools and returned for data capture.

For key stage 2, data from the live administration was used to carry out the analysis prior to standard setting (see section 13.2: Standard setting methodology). In all, data from over half a million pupils were used to determine the standards for key stage 2.

13.2 Standard setting methodology

A Bookmark standard setting procedure was used to set the key stage 1 and key stage 2 standards. This procedure is used widely internationally to set standards and was also used to set the standard on the phonics screening check, the level 6 reading and mathematics tests and the level 3-5 English grammar, punctuation and spelling tests when they were introduced. The use of the method was discussed with STA's technical advisory group and was signed off at STA's technical sub-programme board.

Two distinct Bookmark standard setting sessions were carried out for each of the key stage 1 and key stage 2 tests. Running two sessions allows the sessions to act as validation exercises for each other as two independent panels of teachers participated. Care was taken to standardise the presentation of all information as much as possible across the two meetings. Baseline and final evaluations were gathered from the teacher panellists to provide further validity evidence regarding the standard setting outcomes.

The teachers were selected from over 2,500 who responded to the expression of interest that was sent to all schools. The objective was to have a representative group of teachers involved each day in terms of region and school type, and similar representation each day in terms of teaching experience and subject specific expertise. However, this was constrained somewhat by availability and was impacted by withdrawals after the initial invitation to attend. A maximum of one teacher per school was selected across all the key stage 1 and key stage 2 standard setting meetings to ensure participation from as many schools as possible. The number of participants attending each standard setting session is presented in the accompanying technical appendix.

Standard setting panellists were provided with an ordered item booklet, whereby each of the mark points on each item on a test was presented in order of difficulty. A two-parameter item response theory (IRT) model using a response probability of two-thirds (Cizek and Bunch, 2009[29]) was used to assemble the ordered item booklets.

The task for participants was then to use the performance descriptors to identify the position in the ordered item booklet where a minimally capable pupil just at the expected standard would have a two-thirds chance of achieving the mark. This took place over several rounds, with participants working individually, in small groups and as a whole group to converge on the recommended threshold for each test.

The dates of the standard setting meetings are provided in the accompanying technical appendix.

After standard setting, key STA staff and a curriculum advisor attended a standards confirmation meeting for each subject. Representatives from Ofqual and the teacher unions were present to observe the decision-making process.

The purpose of this meeting was to sign off the mark relating to the expected standard for each test. Within the meeting, the psychometricians who had been present at the standard setting meeting summarised key points from discussion within the meetings, along with the recommendation for the cut score for each subject. A scaled score conversion chart was also presented for each test. Following discussion of key points from the meetings, the cut score for each test was ratified.

13.3 2016 key stage 1 standard setting outcomes

In the baseline evaluation in each subject, all participants (except one in English grammar, punctuation and spelling) agreed with the statement, ‘I am comfortable with the performance descriptor.’ Participants were asked whether they understood the procedure

[29] Cizek, G. J. and Bunch, M. B. (2009). Standard setting: A guide to establishing and evaluating performance standards on tests. Sage: Thousand Oaks.

as it had been explained to them. Again, all participants (except one in English grammar, punctuation and spelling) agreed with the statement. Finally, participants were asked whether, 'the discussion of the standard setting procedure was sufficient to allow me to be confident that my colleagues and I will be able to determine a standard.' All participants in all subjects at key stage 1 agreed with that statement.

In the final evaluation in each subject, all participants across the three subjects were fairly positive about their standard setting experience, indicating they understood the task, had been given clear instructions and that the group discussions had been very useful. Across the three subjects, there was a mix of response around the participant's comfort with the agreed bookmark. There were several participants who were very comfortable with the agreed bookmark, a large number of participants were comfortable with the agreed bookmark, and there were a small number of panellists who were somewhat or very uncomfortable with the agreed bookmark.

The baseline and final evaluations provided validity evidence that the participants: understood the task; generally accepted the performance level descriptors and felt fairly comfortable with their final judgements; and the outcomes from each of the two meetings were sufficiently close to be averaged. Therefore, the recommendation from the standard setting meetings was to approve the following cut scores:

- English reading – 22 out of 40 marks
- Mathematics – 37 out of 60 marks
- English grammar, punctuation and spelling – 25 out of 40 marks.

13.4 2016 key stage 2 standard setting outcomes

In the baseline evaluation in each subject, all participants agreed with the statement, 'I am comfortable with the performance descriptor.' Participants were asked whether they understood the procedure as it had been explained to them. Again, all participants agreed with the statement. Finally, participants were asked whether, 'the discussion of the standard setting procedure was sufficient to allow me to be confident that my colleagues and I will be able to determine a standard.' All participants (except one in mathematics) agreed with that statement.

In the final evaluation in each subject most participants across the three subjects were fairly positive about their standard setting experience, indicating they understood the task, had been given clear instructions and that the group discussions had been very useful. Across the three subjects there was a mix of responses around the participant's comfort with the agreed bookmark. There were several participants who were very comfortable with the agreed bookmark, a large number of participants were comfortable with the agreed bookmark, and there were a small number of participants who were somewhat or very uncomfortable with the agreed bookmark. In key stage 2, there were a few participants who chose not to answer this question.

The baseline and final evaluations provided validity evidence that the participants: understood the task; generally accepted the performance level descriptors and felt fairly comfortable with their final judgements; and the outcomes from each of the two meetings were sufficiently close to be averaged. Therefore, the recommendation from the standard setting meetings was to approve the following cut scores:

- English reading – 21 out of 50 marks
- Mathematics – 60 out of 110 marks
- English grammar, punctuation and spelling – 43 out of 70 marks.

13.5 Standards maintenance

For the 2017 test onwards, a statistical scaling process is being used to take the standard set on each 2016 test to determine the statistically equivalent score on the new test.

In key stage 1, the data from the 2015 TPT, 2016 TPT and the live data from 2016 were matched and a three-group, two-parameter graded response model was run, enabling us to equate from the 2016 live sample collected for standard setting to the 2017 selected test items in the 2016 TPT data.

In key stage 2, standalone anchor tests were administered to a representative sample of pupils who took the live test. The anchor test data was matched to the live test data for the IRT analysis and a two-group, two-parameter graded response model was run, enabling us to equate from the 2016 live test to the 2017 live test.

The methodology for producing scaled score conversion tables in 2016 was discussed and agreed at the Technical Sub-Programme Board / Technical Advisory Group meeting in April 2016.

A linear IRT scale transformation was used such that:

$$SC_i = \frac{\sigma(SC)}{\sigma(\theta)}\theta_{Xi} + \left[100 - \frac{\sigma(SC)}{\sigma(\theta)}\theta_{cut}\right]$$

Equation 1: Linear IRT scale transformation

where, SC is the scaled score, θ_{Xi} is the IRT ability value corresponding to the raw score Xi, and θ cut is the IRT ability value representing the expected standard cut score. Scaled scores were rounded down to the integer below. For the purposes of the scaled score calculation, theta was estimated using the summed score likelihood based approach as implemented in flexMIRT, a specialist IRT software package.

The cut scores identified by this process were validated through a standards maintenance meeting, in a similar way as for the 2016 scaled scores.

14 Common assessment criteria

Statutory national assessments are regulated by Ofqual. Ofqual's statutory objectives are to promote standards and public confidence in national assessments, and its primary duty is to keep all aspects of national assessments under review. Ofqual focuses on the validity of assessment and takes a risk-based approach: observing, scrutinising and reporting on key aspects of assessment arrangements relating to validity.

This chapter sets out the evidence that STA has generated in relation to the common criteria in Ofqual's 2011 regulatory framework for national assessments[30]: validity, reliability, comparability, minimising bias and manageability. At the time of writing, Ofqual's regulatory framework for national assessments is under review.

14.1 Validity

The development of a validity argument must start with an understanding of the purpose of the assessment. The purpose of the key stage 1 and key stage 2 tests is to measure performance in relation to relevant areas of the national curriculum.

To determine whether the test is a sufficiently valid assessment of the level of attainment that pupils have achieved, there are two main questions that need to be answered:

1. Is the test an appropriate assessment of the relevant sections of the national curriculum programme of study?

2. Are the reported outcomes of the test appropriate with respect to the expected standard?

In relation to the first question, the test was developed using the process described in this handbook, which aligns with international best practice. During this process, evidence has been collected relating to the content of the test and whether it appropriately assesses the national curriculum.

The reviewers and experts involved in the development of the tests have provided qualitative validity evidence to complement quantitative trialling data; in totality, the process has provided sufficient data to enable STA to construct tests that meet the test specifications.

[30] Ofqual. (2011). Regulatory Framework for National Assessments: National Curriculum and Early Years Foundation Stage. Available at: www.gov.uk/government/publications/regulatory-framework-for-national-assessments

As a result, STA is confident that the tests represent an appropriate assessment of the relevant national curriculum programme of study.

In relation to the second question, the new content of the tests in 2016 meant that there was no link to the expected standard (which was denoted by levels) in previous tests. A robust process was followed to set the standard so that it could be maintained for future years. The analysis that we have undertaken that links the standards on the 2017 tests to standards set on the 2016 national curriculum tests provides evidence that standards have been maintained.

Ofqual's annual report[31] and accounts for 2016 to 2017, which is presented to Parliament, highlights the regulator's interest in the technical aspects of the test development process, particularly as these aspects relate to validity. STA's technical approach to standard setting was scrutinised and evaluated. Ofqual was "satisfied that the STA had adopted an appropriate and professionally recognised standard-setting technique and that it had applied this process carefully and effectively."

14.2 Reliability

There are various sources of error, which can be measured by different reliability statistics. Specific reliability studies, such as those designed to measure test re-test reliability or marking consistency were not carried out for this assessment. However, there are other measures that can be reviewed.

Reliability statistics are given in the accompanying technical appendix and demonstrate good levels of reliability.

14.3 Comparability

As the 2016 test was the first test of the new national curriculum and there were significant changes from the previous tests and a change to the expected standard, direct comparisons cannot be made with tests or performance in previous years.

The tests each year are developed to the same test framework to support comparability. As described in chapter 7: Trialling, the test development process uses an anchor test or anchor items to equate each TPT and live test, therefore ensuring comparability of performance from 2016 to 2017 and for future live tests.

[31] Available at: www.gov.uk/government/publications/ofqual-annual-report-for-the-period-1-april-2016-to-31-march-2017

14.4 Minimising bias

The questions that were selected for inclusion in the 2016 and 2017 tests were reviewed for bias throughout the test development process, including through feedback from inclusion panels and the monitoring of differential item functioning during data analysis (see accompanying technical appendix). A full suite of access arrangements and modified tests was available for pupils to ensure fair access to the tests. This provides confidence that bias was minimised in the 2016 and 2017 tests.

14.5 Manageability

The administration of the 2016 and 2017 national curriculum tests follows established arrangements for end of key stage testing in schools. Administration guidance documents to support schools were provided in accordance with the normal timetable.

The majority of administrations were undertaken appropriately, with upheld key stage 2 maladministration cases representing 0.4% of all schools administering in 2016, indicating that administration of the tests is manageable for schools.

15 Glossary

Phrase	Definition
Anchor test	a set of secure items administered annually to inform test construction and standards maintenance
Coding	used in trialling, the act of assigning numeric codes to different types of pupil responses
Coding frame	a document outlining the requirements for assigning numeric codes to pupil responses
Chi-square test	a statistic testing the relationship between the row and the column in a table
Cognitive domain	thinking and processing skills specific to a test subject
Content domain	the assessable elements of the programme of study
Cut score	see Threshold (of the expected standard)
Differential item functioning	a statistic which indicates whether different groups of pupils at the same level of attainment have the same probability of correct response on an item
Distractor	the incorrect options given for a multiple-choice item
Equating	the statistical process of establishing the relationship between scores on different forms of the same test
Inter-rater	consistency between two or more judges
Item	within assessment, a ‘question’ is referred to as an item – the lowest level to which a mark can be awarded
Invitation to quote	procurement documentation outlining requirements for tendering for work
Item bank	database holding item, test and process information
Item classification	metadata collected on item characteristics to ensure tests have appropriate coverage over time
Item writing agency	an outside organisation that supplies items to the STA
Mark scheme	a document outlining the requirements for awarding marks on an item level
Modified test agency	an outside organisation that supplies modified test versions and advice to schools regarding the use of modified materials with pupils

Phrase	Definition
Observed score	the reported score for a pupil
Reliability	indicates the dependability, consistency or freedom from random measurement error of a test score
Scaled score	the reported outcome of the tests against the scale developed for the tests in 2016
Scaled score conversion	the process of converting raw scores to scaled scores
Seeding	items embedded in marking / coding to ensure on-going marking / coding quality
Selected response item	a test question that requires pupils to select their answer from one of several possibilities given
Standardisation items/scripts	depending on the marking mode, these items/scripts are used to ensure markers / coders are marking / coding accurately – usually presented before any marking of pupil responses is allowed
Standard setting	a data-driven judgement-based process used to determine the threshold(s) on any given test
Test framework	subject-specific documents containing detail to guide the test development process
Themed response table	a table designed to illustrate the differences between the best quality responses that do not get a mark with the minimum quality responses that do get a mark, in order to increase marking reliability
Threshold (of the expected standard)	the specific point on the score scale that differentiates between those who have not reached the expected standard and those who have – also referred to as ‘cut score’
True score	a classical test theory concept interpreted as the average of observed test scores over an infinite number of administrations of the same test (per pupil)
Validity	the degree to which evidence gathered in the test development process supports the test outcome interpretation that the test measures what is intended

Standards & Testing Agency

About this publication:

enquiries www.education.gov.uk/contactus

download www.gov.uk/government/publications

Reference: STA/17/8129/e pdf version ISBN: 978-1-78315-957-4

Follow us on Twitter: @educationgovuk

Like us on Facebook: facebook.com/educationgovuk