Standards & Testing Agency

# Appendix C: 2019 Validity framework

## Key stage 1 English reading

**January 2020**

# Contents

## Summary

The validity frameworks are appendices to the test handbook and provide validity evidence gathered throughout every stage of the development of the national curriculum tests. It has been produced to help those with an interest in assessment to understand the validity argument that supports the tests.

## Who is this publication for?

This publication is for test developers and others with an interest in assessment.

# Claim 1: Test is representative of the subject/national curriculum

## 1.1 Are the assessable areas of the curriculum clearly defined as a content domain?

The following list explains how the content domain was developed to ensure it was clearly defined.

a. STA developed the content domain for the key stage 1 (KS1) English reading national curriculum test (NCT), based on the national curriculum programme of study (2014) for English at KS1.
b. The content domain is defined in the KS1 English reading test framework (Section 4, page 8).
c. The content domain sets out the elements of the programme of study that are assessed in the English reading test. STA test development researchers (TDRs) used wording as close to the curriculum as possible as the wording was easily translatable to a set of skills assessable in a pencil and paper test. The wording of the curriculum is such that the content domain's focus is on the comprehension skills of retrieval and inference.
d. The content domain was developed by STA's expert TDRs in consultation with the Department for Education (DfE) curriculum division. STA appointed two independent curriculum advisors to support the development of the English reading NCTs.
e. STA asked a panel of education specialists to review a draft of the content domain before it was finalised. The range of stakeholders that was involved in producing the content domain gives assurance that it is appropriate.
f. STA published the draft framework in March 2014 and the final version in June 2015. No concerns have been raised with STA about the content domain.

The evidence above confirms that the assessable areas of the curriculum are clearly defined in the content domain.

### 1.2 Are there areas that cannot be assessed in a paper and pencil test? Are there any parts of these non-assessable areas that could be assessed in a paper-based test but are better suited to different forms of assessment?

The non-assessable elements of the national curriculum are defined in Table 1. The rationale for why any element of the national curriculum is not deemed assessable in a paper-based test is also provided:

| Element of national curriculum | Rationale for not including in content domain | How this element could be assessed |
|---|---|---|
| Develop pleasure in reading, motivation to read, vocabulary and understanding by listening to, discussing and expressing views about a wide range of contemporary and classic poetry, stories and non-fiction at a level beyond that at which they can read independently | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |
| Develop pleasure in reading, motivation to read, vocabulary and understanding by becoming increasingly familiar with and retelling a wider range of stories, fairy stories and traditional tales | Requires class discussion so not suitable for a pencil and paper test | Through class discussion with the teacher |
| Develop pleasure in reading, motivation to read, vocabulary and understanding by being introduced to non-fiction<br>Develop pleasure in reading, motivation to read, vocabulary and understanding by books that are structured in different ways | These elements are too subjective to assess in a pencil and paper test | Understanding element can be assessed by teachers |

| | | |
|---|---|---|
| Develop pleasure in reading, motivation to read, vocabulary and understanding by recognising simple recurring literary language in stories and poetry | This element is too subjective to assess in a pencil and paper test | Understanding element can be assessed by teachers |
| Develop pleasure in reading, motivation to read, vocabulary and understanding by discussing their favourite words and phrases | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |
| Develop pleasure in reading, motivation to read, vocabulary and understanding by continuing to build up a repertoire of poems learnt by heart, appreciating these and reciting some, with appropriate intonation to make the meaning clear | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |
| Understand both the books that they can already read accurately and fluently and those that they listen to by drawing on what they already know or on background information and vocabulary provided by the teacher | This element requires input from the teacher | Can be assessed through a variety of means by class teachers |
| Understand both the books that they can already read accurately and fluently and those that they listen to by checking that the text makes sense to them as they read and correcting inaccurate reading | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |

| | | |
|---|---|---|
| Understand both the books that they can already read accurately and fluently and those that they listen to by answering and asking questions | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |
| Participate in discussion about books, poems and other works that are read to them and those that they can read for themselves, taking turns and listening to what others say | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |
| Explain and discuss their understanding of books, poems and other material, both those that they listen to and those that they read for themselves | This element requires class discussion so is not suitable for a pencil and paper test | Through class discussion with the teacher |

Table 1: Non-assessable elements of the national curriculum

No concerns have been raised with STA regarding the inclusion of the elements described in the non-assessable content section of the test framework.

The evidence above confirms that these areas are better suited to different forms of assessment.

### 1.3 Are the areas of the curriculum that are deemed to be assessable in a paper and pencil test an accurate reflection of the whole curriculum?

STA excluded some elements of the national curriculum from the content domain for the KS1 reading test. This is not a significant exclusion in terms of the skills of reading comprehension and so the content domain remains an accurate reflection of the national curriculum.

### 1.4 Do the rating scales within the cognitive domain provide an accurate reflection of the intended scope of teaching and learning outlined within the national curriculum?

The following list explains how the cognitive domain was developed to ensure it was an accurate reflection of the intended scope of teaching and learning outlined within the national curriculum.

a. The cognitive domain for the KS1 English reading test is defined in the KS1 reading test framework (Section 5, page 9).
b. Before developing the cognitive domain, STA reviewed the domains for similar sorts of tests. The cognitive domain for KS1 English reading was based on the work described in the paper, 'A Framework for Predicting Item Difficulty in Reading Tests' by Tom Lumley, Alla Routitsky, Juliette Mendelovits and Dara Ramalingam from the Australian Council for Educational Research (ACER). This research was based on the Programme for International Student Assessment (PISA) scale and was presented at the American Educational Research Association (AERA) conference in 2012.
c. STA synthesised and amended these existing models to take account of the specific demands of the subject and the cognitive skills of primary-aged children. The model that resulted allows TDRs to rate items across five different areas of cognitive demand.
d. TDRs mapped previous test material that corresponded to the new content domain to test the first drafts of the cognitive domain. They then refined the cognitive scales according to the results. TDRs decided to define the two ends of each scale and not the middle to avoid being overly prescriptive in their approach.
e. For Strand A, accessibility of the target information, TDRs produced a scale ranging from the minimum amount of information a pupil could be expected to read in order to answer a question, for example one or two pieces for a simple retrieval or inference question, to the top end of the scale where pupils are expected to retrieve information from across the text in order to make global inferences or assertions. TDRs envisaged that the lower end of the scale would be more applicable to paper 1 items whereas paper 2 would have wider-ranging question opportunities and would thus occupy a wider portion of the scale. Paper 1 is better placed to test A1 and A2 because a small amount of text appears on each page and it is expected that most pupils will answer questions in the order in which they progress through the booklet.
f. For Strand B, complexity of the target information, the lowest point of the scale is for tasks where the question wording indicates the area of the text where the information needed to respond can be found. For example, questions at the lower end of this scale might include a locator or a quotation to direct pupils to the relevant bit of text. However, questions rated at the higher end of this scale would not include such a locator, meaning that pupils would have to find independently the relevant part of the text to answer the question.
g. The scale for Strand C, task-specific complexity, rates simple retrieval questions at the lower end of the scale and more complex inferences at the higher end of the scale. The complexity of the text is a factor here, as more complex texts might lend themselves to more complex inferences and more straightforward non-fiction texts, for example, might lend themselves to more straightforward retrieval questions. Generally, paper 1 tests contain more questions rated at the C1 and C2 level, while paper 2 tests contain more questions at the higher end of the range, due to the more demanding nature of the texts in paper 2.

h. Strand D, response strategy, relates to the strategy pupils use to answer the question. The lower end of the scale encompasses answers of a few words or where pupils have to tick a box or draw lines to match text together. The higher end of the scale is for extended answers where there are multiple answer lines for the pupil response. The variety of response types allows pupils to demonstrate their understanding in different ways. As paper 1 does not include any 2-mark questions (which sometimes require longer answers), it is generally the case that there are more opportunities for assessing D4 in paper 2 using open 2-mark questions.
i. Strand E, technical knowledge required, relates to the amount of technical language used in the question or in the part of the text where the answer can be found. The expectation is that the majority of questions will fall in the lower half of the range for this strand. However, there are non-fiction texts, for example, where technical terms might be introduced in a supportive context so that pupils can glean their meaning. Questions where pupils are expected to understand such vocabulary may be rated higher on the scale.
j. Panels of teachers reviewed the test frameworks to validate the cognitive domains. STA asked the teachers to comment on the extent to which the cognitive domain set out the appropriate thinking skills for the subject and age group. In addition, pairs of TDRs independently classified items against the cognitive domain and compared their classifications.
k. TDRs made further refinements to the cognitive domains based on both the inter-rater consistency between TDRs and the comments gathered from the teacher panels. This ensured that the cognitive domains published in the test frameworks were valid and usable.

The evidence above confirms that the rating scales within the cognitive domain provide an accurate reflection of the intended scope of teaching and learning outlined within the national curriculum.

### 1.5 How well do the items that are available for selection in the test cover the content domain and cognitive domain as set out in the test framework?

109 items were available for the 2019 KS1 English reading test construction.

The texts available for selection for the 2019 live test are included in Table 2. RPA refers to Paper 1 – the reading prompt and answer booklet, RAB refers to Paper 2 – the reading answer booklet (which has a separate prompt).

| **Set of prompts** | **Texts and types** | **Word count<br>RPA: 400–700 words<br>RAB: 800–1100 words** |
|---|---|---|
| RPA24 | Text A (non-fiction) | 271 words |
| | Text B (fiction) | 373 words |
| | **Total** | **644 words** |
| RABX | Text C (fiction) | 355 words |
| | Text D (non-fiction) | 639 words |
| | **Total** | **994 words** |
| RPA23 | *My Big Brother JJ* (fiction) | 364 words |
| | *What is a Cowboy?* (non-fiction) | 301 words |
| | **Total** | **665 words** |
| RABW | *Liam the Park Keeper* (non-fiction) | 484 words |
| | *Dora the Storer* (fiction) | 501 words |
| | **Total** | **985 words** |

Table 2: Texts available for selection

These items covered the content and cognitive domains as shown in Table 3.

| | Requirement | RPA24 (paper 1) | RABX (paper 2) | RPA23 (paper 1) | RABW (paper 2) |
|---|---|---|---|---|---|
| No. marks available | 40 | 26 | 28 | 34 | 25 |
| 1-mark items | 34–36 | 26 | 24 | 34 | 21 |
| 2-mark items | 1–3 | 0 | 2 | 0 | 2 |
| Enemy marks | – | 7 | 11 | 17 | 11 |
| | | | | | |
| 1a | 1–8 | 2 | 3 | 3 | 4 |
| 1b | 16–32 | 16 | 17 | 24 | 13 |
| 1c | 0–3 | 0 | 2 | 0 | 0 |
| 1d | 4–14 | 8 | 6 | 7 | 8 |
| 1e | 0–2 | 0 | 0 | 0 | 0 |
| | | | | | |
| Selected | 12–20 | 9 | 14 | 10 | 13 |
| Short | 12–24 | 14 | 11 | 22 | 8 |
| Extended | 2–6 | 3 | 3 | 2 | 4 |
| | | | | | |
| D1 | 12–24 | 11 | 13 | 9 | 14 |
| D2 | 12–24 | 8 | 6 | 16 | 3 |
| D3 | 12–24 | 6 | 8 | 8 | 4 |
| D4 | 2–6 | 1 | 1 | 1 | 4 |

Table 3: Content and cognitive domain ratios

As elements of the content domain are sampled over time, any that were not available for selection were not required to be included in a final test but may be included in future tests, according to the ranges published in the test framework. For example, for content domain 1e (predict what might happen on the basis of what has been read so far), no items were available for selection. This was not unexpected and did not cause concern as the test framework does not require a test to contain a 1e item – the recommended range for 1e items is 0–2 marks.

The evidence above confirms that an appropriate range of items was available for selection to cover the content and cognitive domain.

### 1.6 Have test items been rigorously reviewed and validated by a range of appropriate stakeholders? To what extent has feedback led to refinements of test items?

STA designed the test development process to ensure a range of stakeholders reviews and validates items throughout development. These stages are:

a. Item writing: STA item writers, TDRs and external curriculum advisors review items. The reviewers suggest improvements to items and STA makes the improvements before the next stage.
b. Expert review 1 and 2: a wide range of stakeholders reviews the items to confirm they are appropriate. This stakeholder group includes teachers, subject experts, special educational needs and disability (SEND) experts, inclusion experts and local authority staff. TDRs collate the feedback and decide on the amendments to the items in a resolution meeting with STA staff and curriculum advisors.
c. Item finalisation after trialling: TDRs and psychometricians review items after each trial using the evidence of how the item performed. TDRs can recommend changes to items based on this evidence. Items that are changed may be considered ready to be included in a technical pre-test (TPT) or a live test, depending on their stage of development. If the change is more significant, TDRs may decide that they need to review the item further.

The technical appendix of the test handbook contains information about the item-writing agencies and expert review panels.

STA holds a final expert review (expert review 3) after constructing the live test. At this meeting, STA asks stakeholders to review the completed test. If the panel identifies a problem with any items, STA may replace these items. The technical appendix of the test handbook contains information about expert review 3.

STA keeps the evidence relating to the review and validation of individual items in its item bank.

The evidence above confirms that test items have been rigorously reviewed and validated by a range of appropriate stakeholders and that this feedback has led to refinements of test items.

### 1.7 Have test items and item responses from trialling been suitably interrogated to ensure only the desired construct is being assessed (and that construct-irrelevant variance is minimised)?

Following each trial, an item finalisation meeting takes place involving TDRs and psychometricians. The purpose of the meeting is to review all available evidence and make decisions on the most appropriate next stage for each item. For each item, the following evidence is reviewed:

a. classical analysis and item response theory (IRT) analysis of the performance of items including difficulty and discrimination.
b. differential item functioning (DIF) analysis, by gender for the item validation trial (IVT) and by gender and English as an additional language (EAL) for the TPT.
c. analysis of coding outcomes and coder feedback.
d. reviews of children's responses to items to see how children are interacting with questions.

After the IVT, the following outcomes are available for each item:

a. Proceed to expert review 2 stage unamended since there is sufficient evidence that the question is performing as intended.
b. Proceed to expert review 2 stage with amendments since, although there is some evidence that the item is not performing as intended, the issue has been identified and corrected.
c. Revert to expert review 1 stage with amendments since the issues identified are considered major and the item will need to be included in an additional IVT.
d. Archive the item as major issues have been identified that cannot be corrected.

After the TPT, the following outcomes are available for each item:

a. Item is available for inclusion in a live test since the evidence shows it is performing as intended.
b. Item requires minor amendments and will need to be re-trialled before inclusion in a live test.
c. Item is archived since a major issue, that cannot be corrected, has been identified.

Any item that is determined to be available for inclusion in a live test has therefore demonstrated that it assesses the appropriate construct. Evidence related to individual items is stored within the item bank and is not repeated here, though is available should specific issues be identified.

The evidence above confirms that test items and item response from trialling have been suitably interrogated to ensure only the desired construct is being assessed and that construct-irrelevant variance is minimised.

### 1.8 Does the final test adequately sample the content of the assessable curriculum (whilst meeting the requirements within the test framework)? Is a range of questions included that are appropriate to the curriculum and classroom practice?

The 2019 KS1 English reading test meets the requirements of the test framework as shown in Tables 4–8.

| Text | Type | Word count* |
|---|---|---|
| *My Big Brother JJ* | fiction | 364 |
| *What is a Cowboy?* | Non-fiction | 284 |
| Total paper 1 | | 648 |
| *Liam the Park Keeper* | Non-fiction | 484 |
| *Dora the Storer* | fiction | 501 |
| Total paper 2 | | 985 |

Table 4: Text types and word counts

*Live target: 400–700 words for paper 1 and 800–1100 words for paper 2.

| | Target | Previous range | 2019 |
|---|---|---|---|
| Number of marks | 40 | 40 | 40 |
| Number of items | – | 36–38 | 38 |
| Number of 1-mark items | 34–36 | 32–36 | 36 |
| Number of 2-mark items | 2–6 | 2–4 | 2 |
| Number of marks in booklet 1 | 20 | 20 | 20 |
| Number of items in booklet 1 | – | 20 | 20 |
| Number of marks in booklet 2 | 20 | 20 | 20 |
| Number of items in booklet 2 | – | 16–18 | 18 |

Table 5: Mark allocations

| Content reference | Target | Previous range | 2019 |
|---|---|---|---|
| 1a | 1–8 | 4–5 | 6 |
| 1b | 16–32 | 23–29 | 24 |
| 1c | 0–3 | 2 | 0 |
| 1d | 4–14 | 8–10 | 10 |
| 1e | 0–2 | 0–1 | 0 |

Table 6: Content domain allocations

| Response strategy | Target | Previous range | 2019 |
|---|---|---|---|
| D1 | 12–24 | 14–20 | 18 |
| D2 | 12–24[1] | 9–17 | 13 |
| D3 | See D2 | 5–8 | 5 |
| D4 | 2–6 | 2–4 | 4 |

Table 7: Response strategy allocations

| Item type | Target | Previous range | 2019 |
|---|---|---|---|
| Selected response | 12–20 | 13–21 | 16 |
| Short response | 12–20 | 17–21 | 19 |
| Extended response | 2–6 | 2–6 | 5 |

Table 8: Item type allocations

[1] The target range for D2 and D3 is combined.

Teachers, subject experts, markers, inclusion experts and independent curriculum advisors reviewed the test at expert review 3 on 9 October 2018. Their comments are summarised below:

a. Attitudes towards the tests were largely positive.
b. Paper 2 was particularly popular. Reviewers thought Liam the Park Keeper was engaging and accessible and Dora the Storer had a positive theme and was a solid choice for the end of paper 2.
c. Reviewers appreciated the balance of text types and themes in paper 1. They thought that My Big Brother JJ was an appropriately straightforward opening text that was prevented from becoming predictable by the mishap with the paint and Mum's unexpected reaction. They commented favourably on the layout and contents of What is a Cowboy? and perceived this text as equally accessible to all pupils.
d. Reviewers expressed some concern over gender stereotyping in My Big Brother JJ: JJ added sports-themed elements to the mural while Jasmine added flowers. They felt this was somewhat compounded by the inclusion of only male cowboys in What is a Cowboy? and Dora pushing a pram in Dora the Storer.
e. Reviewers felt the tests seemed comparable in difficulty to those of previous years and commented that they would be pleased to see these particular texts at test time.

The TDR presented this evidence at STA's project board 3 and the deputy director for assessment development signed off the test.

The evidence above confirms that the final test adequately samples the content of the assessable curriculum, whilst meeting the requirements within the test framework, and that a range of questions is included that are appropriate to the curriculum and classroom practice.

# Claim 2: Test results provide a fair and accurate measure of pupil performance

### 2.1 How has item-level data been used in test construction to ensure only items that are functioning well are included in the test?

The following list indicates how STA collects and uses item level data.

a. STA trials all test materials in a TPT during which approximately 1000 pupils from a stratified sample of schools see each item. This trial provides STA with enough item-level data to be confident it knows how an item will perform in a live test.
b. STA reviews qualitative and quantitative data from the TPT and reports on each item's reliability and validity as an appropriate assessment for its attributed programme of study.
c. TDRs remove any items from the pool of available items that do not function well or that had poor feedback from teachers or pupils. These items may be amended and re-trialled in a future trial.
d. STA holds a test construction meeting to select the items for the live test booklets. The meeting's participants consider: the item's facility (i.e. its level of difficulty); the ability of the item to differentiate between differing ability groups; the accessibility of the item; the item type; presentational aspects; question contexts; coverage in terms of assessing the content and cognitive domains – for each year and over time; and conflicts between what is assessed within test booklets and across the test as a whole.
e. At this stage, TDRs and psychometricians may swap items in or out of the test to improve its overall quality and suitability.
f. TDRs and psychometricians use a computer algorithm and item-level data to construct a test that maximises information around the expected standard, as well as across the ability range, while minimising the standard error of measurement (SEM) across the ability range. The TDRs and psychometricians consider the construction information alongside the test specification constraints and their own expertise to make a final decision on test construction.

The evidence above confirms that item-level data has been used in test construction to ensure only items that are functioning well are included in the test.

### 2.2 How has qualitative data been used in test construction to ensure only items that are effectively measuring the desired construct are included in the test?

STA collects qualitative data from a range of stakeholders throughout the test development cycle and uses it to develop items that are fit for purpose. STA consults stakeholders through the following methods:

a. three independent expert review panels: teacher panel (at expert reviews 1, 2 and 3); inclusion panel (at expert review 1); and test review group panel (at expert reviews 1, 2 and 3).
b. teacher and administrator questionnaires.
c. responses captured by codes at trialling.
d. reviews of pupil responses.
e. observations of trialling.
f. pupil focus groups during trial administrations at item-writing stage conducted by the item-writing agency and at IVT and TPT conducted by administrators and/or teachers.
g. coding and marker meetings including their reports.
h. curriculum expert reports.

TDRs and psychometricians analyse qualitative data at each stage of the process in preparation for trials and live tests alongside the quantitative data gathered. TDRs revisit the data throughout the development process to ensure they are making reliable judgements about the item and the construct it is measuring. STA considers the results of the analysis at key governance meetings: item finalisation, resolution and project board.

Following the TPT, a range of qualitative data has been collected and analysed, including:

a. pre-trial qualitative data from previous expert reviews and trials.
b. coded item responses from trialling.
c. script archive trawl based on codes captured at trialling.
d. teacher and administrator questionnaires, which include evidence given by focus groups of pupils.
e. coders' reports from trialling.
f. curriculum advisor report from resolution.
g. modified agency report comments.

TDRs and psychometricians analyse this data alongside quantitative data before item finalisation. The TDR summarises the information and presents it at an item finalisation meeting.

The senior test development researcher (STDR), the TDR, the senior psychometrician, the project manager and the head of assessment development research attended item finalisation for the 2019 KS1 English reading test. The attendees considered the

information the TDR presented and decided whether items were suitable for live test construction.

The TDR and psychometrician selected items for live test construction based on the outcomes of item finalisation. They used qualitative data to confirm that the items selected were suitable. The TDR and psychometrician considered the following:

a. each item's suitability in meeting the curriculum reference it is intended to assess.
b. stakeholders' views on the demand and relevance of the item.
c. any perceived construct-irrelevant variance (CIV).
d. curriculum suitability.
e. enemy checks – items that cannot appear in the test together.
f. context.
g. positioning and ordering of items.
h. unintentional sources of easiness and/or difficulty.

A combination of stakeholders reviewed the proposed live 2019 KS1 English reading test at expert review 3. This group included teachers, inclusion, curriculum, assessment and English experts. At this meeting, panellists can challenge items and the TDR may use the item data to either defend that challenge or support it. If the panel deems an item unacceptable, the TDR may swap it with a suitable item from the TPT. The panel did not identify any items in the 2019 KS1 English reading test that needed to be swapped.

The TDR collated the data from expert review 3 and presented it alongside the quantitative data for the live test at project board 3. The purpose of this meeting is to scrutinise and critically challenge the data to ensure the test meets the expectations published in the test framework for KS1 English reading.

STA held a one-day mark scheme finalisation meeting for the 2019 KS1 English reading test. At this meeting, an expert group of senior markers reviewed the live test and responses from trialling and suggested improvements to the mark scheme. These amendments do not affect the marks awarded for each question.

In addition, STA held a one-day mark scheme user acceptance testing (UAT) meeting at which six panellists, who were KS1 teachers, trialled the proposed mark scheme on pupil responses to ensure the mark scheme was fit for purpose and could be applied accurately.

The attendees tested the equivalent of ten pupil scripts from the TPT script archive. They were not able to see how the responses had been coded at TPT. The pupil responses included a variety of item types and response types (e.g. answers that had been crossed out and replaced).

The evidence above confirms that qualitative data has been used in test construction to ensure only items that are effectively measuring the desired construct are included in the test.

### 2.3 Is an appropriate range of items that are age appropriate and cover the full ability range included in the final test?

The following list demonstrates how STA ensured an appropriate range of items were included in the final test.

a. External item-writing agencies wrote the items that make up the 2019 KS1 English reading test.
b. STA gives item writers a clear brief to use the relevant parts of the national curriculum document for KS1 when writing their items. This ensures that the items are age appropriate as they are based on a curriculum that a range of experts has deemed suitable. The item-writing contract also states texts must be appropriate for pupils in Year 2 and TDRs judge the acceptability of the item-writing agency's work against this criterion, amongst others.
c. During the item-writing stage, agencies conduct very small-scale trials with approximately 20 to 30 pupils who are in Year 2 or, if overseas, with pupils of an equivalent age. This helps to gauge whether children can interpret items correctly. This also provides the item-writing agency with insights into the most age-appropriate language to use in the items.
d. The TDR reviews the items after the small-scale trials have completed to ensure that they meet the requirements of the national curriculum. A range of experts, including independent curriculum advisors, reviews the items at this stage as part of expert review 1. STA gives the panel members a terms of reference document that asks them to consider whether the items are appropriate for children at the end of KS1.
e. STA also invites test administrators and teachers to give feedback on the test items in a questionnaire. The questionnaire has a specific area for feedback on whether the items are appropriate for children at the end of KS1.
f. The 2019 KS1 English reading test covers the full range of abilities. The test is made up of a range of different cognitive domains, as specified in the test framework. The 2019 KS1 English reading test meets the desired coverage of all strands of the cognitive domain, as set out in the test specification.

Table 9 shows the number of marks available in the 2019 KS1 English reading test for each strand of the cognitive domain.

| | Cognitive domain strand | 2019 |
|---|---|---|
| Accessibility of target information | A1 | 6 |
| Accessibility of target information | A2 | 24 |
| Accessibility of target information | A3 | 8 |
| Accessibility of target information | A4 | 2 |
| Complexity of target information | B1 | 19 |
| Complexity of target information | B2 | 15 |
| Complexity of target information | B3 | 6 |
| Complexity of target information | B4 | 0 |
| Task-specific complexity | C1 | 17 |
| Task-specific complexity | C2 | 12 |
| Task-specific complexity | C3 | 11 |
| Task-specific complexity | C4 | 0 |
| Technical knowledge | E1 | 34 |
| Technical knowledge | E2 | 6 |
| Technical knowledge | E3 | 0 |
| Technical knowledge | E4 | 0 |

Table 9: Cognitive domain

Table 10 shows for each substrand of cognitive domain Strand D (response strategy) the desired range of marks and the number of marks available in the 2019 KS1 English reading test.

| Strand D response strategy | Target | 2019 |
|---|---|---|
| D1 | 12–24 | 18 |
| D2 | 12-24[2] | 13 |
| D3 | See D2 | 5 |
| D4 | 2–6 | 4 |

Table 10: Strand D response strategy

For reading, items are ordered to reflect the chronology of the text.

Most of the test information is focused around the expected standard, although items are selected to ensure there is information at both the lower end and at the higher end of the ability range.

The evidence above confirms that an appropriate range of items that are age appropriate and cover the full ability range is included in the final test.

### 2.4 What evidence has been used (qualitative and quantitative) to ensure the test does not disproportionately advantage or disadvantage any subgroups?

The following list demonstrates how STA ensured the test does not disproportionately advantage or disadvantage any subgroups.

a. TDRs have interpreted a wide range of evidence to ensure the 2019 KS1 English reading test does not disproportionately advantage or disadvantage the following subgroups: non-EAL and EAL; girls and boys; no SEN and SEN; pupils with visual impairments (modified paper); and braillists (modified paper).
b. Expert panels of teachers, educational experts and inclusion specialists reviewed the items and considered whether they were suitable for inclusion in a trial. The inclusion specialists for the 2019 KS1 English reading test consisted of representation from the visual impairment, dyslexia, SEND and hearing impairment specialisms. Within this review process, panellists highlight any

[2] The target range for D2 and D3 is combined.

potential bias and suggest ways to remove it. The TDR considers all the available evidence and presents it in a resolution meeting to decide which recommendations to implement.

c. Data relating to the performance of EAL/non-EAL and girls/boys are identified in classical analysis after the TPT. The TDR uses this quantitative information (facility and per cent omitted) in conjunction with the qualitative evidence from the teacher questionnaires and administrator reports to flag any items that appear to be disproportionately advantaging or disadvantaging a group. As an acknowledgement that pupils in these groups have a wide range of ability, TDRs treat this information with some caution during the decision-making process for each item.
d. STA also carries out a statistical analysis – differential item functioning (DIF) – after the trial. The purpose of this is to identify differences in item performance based on membership in EAL/non-EAL and girls/boys groups. Moderate and large levels of DIF are flagged. As DIF only indicates differential item performance between groups that have the same overall performance, the test development team considers qualitative evidence from the teacher questionnaires and previous expert review panels to help determine whether the item was biased or unfair.
e. None of the items available for inclusion in the 2019 KS1 English reading test were flagged as having moderate or large DIF.
f. Alongside the development of the standard test, STA works closely with a modified test agency to produce papers that are suitable for pupils who require a modified paper. TDRs and modifiers carefully consider any modification to minimise the possibility of disadvantaging or advantaging certain groups of pupils who use modified papers. STA and the modifier make these modifications and ensure minimal change in the item's difficulty.
g. For the majority of the items in the 2019 KS1 English reading braille test, the modifier used standard modification to minimally change the format of items or did not modify items at all. Sometimes, the modifiers are unable to modify an item in a way that maintains its original construct. None of the items in the 2019 KS1 English reading braille test required modifications that changed the construct of the question and STA did not have to replace any of the items.
h. For the majority of the items in the 2019 KS1 English reading modified large print (MLP) test, the modifier used standard modification to minimally change the format of items or did not modify items at all. Sometimes, the modifiers are unable to modify an item in a way that maintains its original construct. None of the items in the 2019 KS1 English reading test required modifications that changed the construct of the question and STA did not have to replace any of the items.

The evidence above confirms that an appropriate range of qualitative and quantitative evidence is used to ensure that the test does not disproportionately advantage or disadvantage any subgroups.

### 2.5 Have pupil responses been interrogated to ensure pupils are engaging with the questions as intended?

The following list demonstrates how STA interrogates pupil responses.

a. STA collects pupil responses for the KS1 English reading test in the IVT and TPT.
b. STA codes responses for each item to collect information on the range of creditworthy and non-creditworthy responses pupils might give. TDRs develop coding frames. Independent curriculum advisors and senior coders review the coding frames. TDRs refine the coding frames both before and during trialling based on this feedback.
c. When coding is complete, the trialling agency provides STA with a PDF script archive of the scanned pupil scripts and a report from the lead coders.
d. STA psychometricians provide classical and distractor analysis to TDRs at IVT and TPT (plus IRT analysis at TPT).
e. TDRs analyse the data, review the report and scrutinise pupil scripts. TDRs may target specific items that are behaving unexpectedly and use the pupil scripts to provide insight into whether pupils are engaging with the questions as intended. TDRs can request script IDs to help them target specific responses from children based on the codes awarded.
f. At TPT, they also randomly select scripts across the ability range and aim to look through the majority of the 1000 responses – particularly for the extended response items. TDRs present the information they have collected from script reviews with other evidence at the item finalisation meeting. TDRs use this evidence to make recommendations for each item.

The evidence above confirms that pupil responses have been interrogated to ensure pupils are engaging with the questions as intended.

### 2.6 Is the rationale for what is creditworthy robust and valid? Can this rationale be applied unambiguously?

The following list demonstrates how STA determines what is creditworthy.

a. TDRs include indicative mark allocations in the coding frames they have developed for the IVT and TPT. TDRs discuss creditworthy and non-creditworthy responses with stakeholders at the expert review panels. Senior coders review the coding frames during the coding period. If necessary, TDRs may add codes or examples to the coding frames to reflect pupil responses.
b. TDRs draft mark schemes for each question after constructing the KS1 English reading test. TDRs use the trialling coding frames to inform the content of the mark schemes and selects pupil responses from the trial to use as examples in the mark scheme. These responses are clear examples of each mark point. TDRs may also include responses that are not creditworthy.

c. STA holds a mark scheme finalisation meeting, composed of TDRs, psychometricians, independent curriculum advisers and senior trialling coders. The participants review the live test and responses from trialling and suggest improvements to the mark scheme so that markers can apply it reliably and consistently.
d. KS1 tests are marked internally in schools. As part of the expert review 3 meeting, a panel of teachers and subject experts conduct UAT of the mark schemes. TDRs collate pupil scripts for each question from the trialling process and allocates marks according to the proposed mark scheme. The panel members mark the pupil scripts and their marking is compared with that done by TDRs to see whether the mark scheme can be applied consistently and unambiguously.

The evidence above confirms that the rationale for what is creditworthy is robust and valid and can be applied unambiguously.

### 2.7 Are mark schemes trialled to ensure that all responses showing an appropriate level of understanding are credited and that no responses demonstrating misconceptions or too low a level of understanding are credited?

The following list demonstrates how STA trialled the mark schemes.

a. STA develops mark schemes alongside their associated items.
b. Item-writing agencies and TDRs draft mark schemes during the initial item-writing stage. TDRs and external curriculum reviewers review these mark schemes.
c. TDRs refine the mark schemes through two rounds of large-scale trialling. Approximately 300 pupils see each item in the IVT. TDRs draft coding frames so they can group pupil responses into types rather than marking them correct or incorrect. Coding allows TDRs to understand how pupils are responding to questions and whether their answers are correct or incorrect. TDRs and psychometricians consider the qualitative data gathered from coding along with quantitative data to make recommendations for changes to the mark schemes. This ensures the mark scheme includes an appropriate range of acceptable responses and examples of uncreditworthy responses.
d. The trialling agency provides STA with a digital script archive of all the pupil answer booklets. TDRs are able to review pupil scripts to view example pupil responses. Reviewing the script archive in this way enables TDRs to ensure coding frames reflect pupil responses.
e. A second trial is administered – the TPT – during which approximately 1000 pupils see each item. TDRs amend coding frames using the information gathered during the IVT. After TPT administration is complete and before marking commences, a group of lead coders review a subset of TPT scripts to ensure the coding frames reflect the range of pupil responses. TDRs and lead coders agree amendments to the coding frames before coding begins.
f. When coding is complete, lead coders write a report for STA that contains their

reflections on the coding process, highlights any specific coding issues and makes recommendations on whether each item could be included in a live test. This report forms part of the qualitative evidence that is reviewed by TDRs.

g. After TPT coding is complete, TDRs consider the lead coder reports and other statistical and qualitative information to make recommendations on which items are performing as required. At this stage, TDRs review pupil scripts and consider the data gathered from coding to ensure that all responses that demonstrate the required understanding are credited and that responses that do not demonstrate the required understanding are not credited.
h. When the TDR and psychometrician have constructed the live test, TDRs use the coding information and pupil responses from the TPT to draft mark schemes. The wording of the mark scheme is finalised. In a small number of cases, STA may need to partially or wholly re-mark a question in the live test to account for changes to the mark scheme after finalisation. For the 2019 KS1 English reading test, no questions had marking changes and so the analysis was not re-run.

The evidence above confirms that mark schemes are trialled to ensure that all responses showing an appropriate level of understanding are credited and that no responses demonstrating misconceptions or too low a level of understanding are credited.

### 2.8 Do the mark schemes provide appropriate detail and information for markers to be able to mark reliably?

The following list demonstrates how STA ensured the mark scheme is appropriate.

a. TDRs developed the mark schemes for the 2019 KS1 English reading test using coding frames that were used in the trialling process. STA uses coding frames to capture the range of responses that pupils give, both creditworthy and non-creditworthy. This allows TDRs to understand how effective an item is and to identify any issues that could affect the accuracy of marking.
b. TDRs draft initial coding frames, which are refined during expert review and trialling. A range of stakeholders reviews the coding frames before they are used. This group includes STA curriculum advisors, psychometricians and some senior coders.
c. TDRs may make further amendments to the coding frames during coding to reflect the range of pupil responses seen. They may also include additional codes to capture previously unexpected responses. TDRs may amend the wording of codes to better reflect how pupils are responding or to support coders in coding accurately.
d. Following the IVT, TDRs update coding frames to include exemplar pupil responses and to reflect the qualitative data that the senior coders provide. Their feedback focuses on whether the coding frames proved fit for purpose, identifying any issues coders faced in applying the coding frames and making suggestions for amendments.

e. Following each trial, the trailing agency provides an archive of scanned pupil scripts and psychometricians provide analysis of the scoring of each item. After IVT, TDRs receive classical and distractor analysis. After TPT, TDRs receive classical, distractor and IRT analysis. TDRs analyse this data and review pupil responses in the script archive in preparation for an item finalisation meeting, where they make recommendations about each item and comment on the effectiveness of the coding frames.
f. After the 2019 KS1 English reading test was constructed, TDRs used the coding information and pupil responses from the TPT to draft mark schemes. To maintain the validity of the data collected from the TPT, STA makes only minor amendments between the TPT coding frame and the live mark scheme. TDRs may refine the wording of the mark scheme or the order of the marking points for clarity and may include exemplar pupil responses from the script archive.
g. STA holds a mark scheme finalisation meeting, composed of TDRs, psychometricians, independent curriculum advisers and senior coders from the trials. The focus of the meeting is to agree that the mark scheme is a valid measure of the test construct and that markers can apply it consistently and fairly.
h. KS1 tests are marked internally in schools. As part of the expert review 3 meeting, a UAT is conducted on the mark scheme by a panel of current KS1 teachers who apply the mark scheme to a range of scripts selected from the TPT archive by TDRs. The outcomes of this test may result in further amendments for clarification and the addition of further exemplification to the mark scheme to ensure it is accessible and can be applied consistently in schools.

The evidence above provides a summary of how mark schemes are developed to provide appropriate detail and information for markers to mark reliably.

### 2.9 Are markers applying the mark scheme as intended?

The KS1 English reading test is marked internally in schools and the results are not reported, therefore STA does not have evidence that the markers apply the mark schemes as intended. However, STA designed the test development process to result in marking that is as consistent as possible. This is done through the thorough development of mark schemes with expert feedback at various stages, the input of lead coders who provide feedback on the process of using the coding frames and UAT to provide evidence that KS1 teachers can apply the mark scheme as intended.

# Claim 3: Pupil performance is comparable within and across schools

## 3.1 Is potential bias to particular subgroups managed and addressed when constructing tests?

The following list demonstrates how STA considers potential bias.

a. In test development, bias is identified as any construct-irrelevant element that results in consistently different scores for specific groups of test takers. The development of the NCTs explicitly takes into account such elements and how they can affect performance across particular subgroups, based on gender, SEND, disability, whether English is spoken as a first or additional language and socioeconomic status.
b. Quantitative data is collected for each question to ensure bias is minimised. DIF is calculated for each question to show whether any bias is present for or against pupils of particular genders or who are or are not native English speakers. The DIF values are then used to guide test construction in order to minimise bias.
c. The fairness, accessibility and bias of each test question are also assessed in three rounds of expert reviews. Texts, items, contexts and illustrations are scrutinised in teacher panels, test review groups (TRGs: comprising senior academic and educational experts) and inclusion panels (visual/audio impairment, SEND, EAL, culture/religion and educational psychology experts). Questions that raise concerns about bias or unfairness are identified and are further examined in-house to either minimise the identified bias or remove the question from the test if no revision is possible.
d. For those pupils who are unable to access the NCTs as they are, alternative test versions are made available, for example braille versions and large print versions. While it is essential that tests are made available in modified formats, the content of the modified tests is kept as close to the original as possible to rule out test-critical changes or any further bias introduced through modification. To ensure this is the case, modification experts are consulted throughout the test development process.
e. Further information about diversity and inclusion in the NCTs can be found in KS1 reading test framework (Section 7, page 18).

The evidence above confirms that potential bias to particular subgroups is managed and addressed when constructing tests.

### 3.2 Are systems in place to ensure the security of test materials during development, delivery and marking?

The following list demonstrates how STA ensured security.

a. All STA staff who handle test materials have undertaken security of information training and have signed confidentiality agreements.
b. Throughout the test development process, external stakeholders are asked to review test items. This is predominantly as part of expert reviews. All those involved in expert review panels are required to sign confidentiality forms, and the requirements for maintaining security are clearly and repeatedly stated in all meetings. Teacher panels are provided with a pack of items in the meeting to comment on, which are signed back in to STA at the end of the day. TRGs review the items in advance of the meeting. Items are sent to TRG members via STA's approved parcel delivery service and panellists are provided with clear instructions on storing and transporting materials. Materials are collected back in via a sign-in process after the TRG meeting.
c. When items are trialled as part of an IVT or TPT, the trialling agency must adhere to the security arrangements laid out in the trialling framework. This includes administrators undertaking training at least every two years, with a heavy emphasis on security. Administrators and teachers present during trialling sign confidentiality agreements. Administrators receive the items for trialling visits (via an approved courier service) and take the items to the schools. They are responsible for ensuring all materials are collected after each visit before returning them to the trialling agency via the approved courier.
d. All print, collation and distribution services for NCTs are outsourced to commercial suppliers; strict security requirements are part of the service specifications and contracts. STA assesses the suppliers' compliance with our security requirements by requiring suppliers to complete a Departmental Security Assurance Model assessment, which ensures all aspects of information technology/physical security and data handling is fit for purpose and identifies any residual risk. These arrangements are reviewed during formal STA supplier site visits. All suppliers operate a secure track and trace service for the transfer of proof/final live materials between suppliers and STA and the delivery of materials to schools.

The evidence above confirms that systems are in place to ensure the security of test materials during development, delivery and marking.

### 3.3 Is guidance on administration available, understood and implemented consistently across schools?

STA publishes guidance on gov.uk 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.

### 3.4 Are the available access arrangements appropriate?

The following list provides details on access arrangements.

a. Access arrangements are adjustments that can be made to support pupils who have issues accessing the tests and ensure they are able to demonstrate their attainment. Access arrangements increase accessibility without providing an unfair advantage to a pupil. The support given must not change the test questions and the answers must be the pupil's own.
b. Access arrangements address accessibility issues rather than specific disabilities or SEND. They are based primarily on normal classroom practice and the available access arrangements are, in most cases, similar to those for other tests such as GCSEs and A levels.
c. STA publishes guidance on gov.uk about the range of access arrangements available to enable pupils with specific needs to take part in the KS1 tests. 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; and who have processing difficulties.
d. The range of access arrangements available includes: early opening to modify test materials (for example, photocopying on to coloured paper); additional time; scribes; transcripts; word processors or other technical or electronic aids; rest breaks and written or oral translations.
e. Headteachers and teachers must consider whether any of their pupils will need access arrangements before they administer the tests.
f. Schools can contact the national curriculum assessments helpline or use NCA tools for specific advice about how to meet the needs of individual pupils.
g. Ultimately, however, a small number of pupils may not be able to access the tests, despite the provision of additional arrangements.

The evidence above provides a summary of the access arrangements available whilst maintaining the validity of the test.

### 3.5 Are the processes and procedures that measure marker reliability, consistency and accuracy fit for purpose? Is information acted on appropriately, effectively and in a timely fashion?

KS1 assessments are internally marked in schools. Owing to the stage of assessment, the mark schemes are more straightforward and reliability is easier to achieve than with complex mark schemes. Section 2.8 contains information on how STA seeks to maximise reliability and usability during the development of the mark schemes. Those marking the tests participate in local authority-provided external moderation activities.

### 3.6 Are the statistical methods used for scaling, equating, aggregating and scoring appropriate?

Methods that are used for scaling and equating NCTs are described in Section 13.5 of the test handbook.

These methods have been discussed and agreed at the Technical Board and agreed to be appropriate by the STA Technical Advisory Group (consisting of external experts in the field of test development and psychometrics).

There are no statistical methods used for scoring NCTs. The tests are scored or marked as described in Section 12 of the test handbook. The processes for training markers and quality assuring the marking ensure that the mark schemes are applied consistently across pupils and schools.

The evidence above confirms that the statistical methods used for scaling and equating are appropriate.

# Claim 4: Differences in test difficulty from year to year are taken account of, allowing for accurate comparison of performance year on year

## 4.1 How does STA ensure appropriate difficulty when constructing tests?

STA has detailed test specifications that outline the content and cognitive domain coverage of items. Trial and live tests are constructed using this coverage information to construct balanced tests. Live tests and some of the trial tests will be constructed using a computer algorithm with constraints on specific measurement aspects to provide a starting point for test construction. This is further refined using STA's subject matter and psychometric expertise.

TPTs are conducted to establish the psychometric properties of items STA is able to establish robust difficulty measures for each item (using a two-parameter IRT analysis model) and, consequently, the tests that are constructed from them have known overall test difficulty. These difficulty measures are anchored back to the 2016 test, thus allowing both new and old items to be placed on the same measurement scale and thereby ensuring a like-for-like comparison.

The evidence above shows how STA ensures appropriate difficulty when constructing the tests.

## 4.2 How accurately does TPT data predict performance on the live test?

IRT is a robust model used for predicting performance of the live test. It allows STA to use the item information from a TPT and to estimate item parameters via linked items. Furthermore, $D^2$ analysis[3] is used to compare item performance across two tests, booklets or blocks. This allows STA to look at potential changes in performance of the items between two occurrences.

As long as sufficient linkage is maintained and the model fits the data (based on meeting stringent IRT assumptions), pre-test data can give a reliable prediction of item performance on a live test.

The evidence above shows how STA uses TPT data accurately to predict performance on the live test.

[3] O'Neil, T., Arce-Ferrer, A. (2012). Empirical Investigation of Anchor Item Set Purification Processes in 3PL IRT Equating. Paper presented at NCME Vancouver, Canada.

### 4.3 When constructing the test, is the likely difficulty predicted and is the previous year's difficulty taken into account?

The first test of the new curriculum occurred in 2016. STA aims for all tests following that to have a similar level of difficulty. This is ensured by developing the tests according to a detailed test specification and by trialling items. Based on the TPT data, STA constructs tests that have similar test characteristic curves to the tests of previous years. Expected score is plotted against ability. Differences are examined at key points on the ability axis: near the top, at the expected standard and near the bottom, with two additional mid-points in between. The overall difficulty with respect to these five points is monitored during live test construction, with differences from one year to the next minimised as far as possible.

As another measure of difficulty comparability, the scaled score range is also estimated and is checked to ensure that it covers the expected and appropriate range compared with previous years. The scaled score range for KS1 English reading is 85–115, and there were only three scaled scores not represented in 2019: 109, 112 and 114. Scale score representation is monitored year on year and in 2019 was similar to previous years.

The evidence above confirms that the likely difficulty is predicted when constructing the test and that the previous year's difficulty is taken into account.

### 4.4 When constructing the test, how is the likely standard predicted? Is the approach fit for purpose?

Using the IRT data from TPT, STA is able to estimate the expected score for every item at the expected standard (an ability value obtained from the 2016 standard-setting exercise). This estimation is possible because the IRT item parameter estimates have been obtained using a model that also includes previous years' TPT and live items, allowing STA to place the parameters on the same scale as the 2016 live test. So, during test construction, the sum of the expected item scores at that specific ability point is an estimate of where, in terms of raw score, the standard (i.e. a scaled score of 100) will be.

Once a final test is established, additional analysis is carried out to scale the parameters to the 2016 scale in order to produce a scaled score conversion table, which estimates the standard for the test.

The approach is fit for purpose and was approved by the Technical Advisory Group in 2017 and confirms that STA's approach to predicting the likely standard is fit for purpose.

### 4.5 What techniques are used to set an appropriate standard for the current year's test? How does STA maintain the accuracy and stability of equating functions from year to year?

The expected standard was set in 2016 using the Bookmark method, with panels of teachers, as outlined in Section 13 of the test handbook.

The standard set in 2016 has been maintained in subsequent years using IRT methodology, as outlined in Section 13.5 of the test handbook. This means the raw score equating to a scaled score of 100 (the expected standard) in each year requires the same level of ability, although the raw score itself may vary according to the difficulty of the test. If the overall difficulty of the test decreases, then the raw score required to meet the standard will increase; if the overall difficulty increases, then the raw score needed to meet the standard will decrease. Similarly, each raw score point is associated with a point on the ability range, which is converted to a scaled score point from 85 to 115.

In order to relate the new tests in each year to the standard determined in 2016, a two-parameter graded response IRT model with concurrent calibration is used. The IRT model includes data from the 2016 live administration and data from TPTs, including anchor items repeated each year and items selected for the live test. The parameters from the IRT model are scaled using the Stocking-Lord scaling methodology to place them on the same scale used in 2016 to determine the standard and scaled scores. These scaled parameters are used in a summed score likelihood IRT model to produce a summed score conversion table, which is then used to produce the raw to scaled score conversions. This methodology was reviewed by and agreed with the STA Technical Advisory Group in 2017.

In order to ensure that the methodology used is appropriate, assumption checking for the model is undertaken. Evidence for the following key assumptions is reviewed annually to ensure the model continues to be appropriate. Evidence from assumption checking analysis is presented at standards maintenance meetings to inform the sign-off of the raw score to scaled score conversion tables. The assumptions are as follows:

a. Item fit: that the items fit the model. An item fit test is used however, owing to the very large numbers of pupils included in the model, results are often significant. Item characteristic curves, modelled against actual data, are inspected visually to identify a lack of fit.
b. Local independence: that all items perform independently of one another and probability of scoring on an item is not impacted by the presence of any other item in the test. This assumption is tested using the Q3 procedure, where the difference between expected and actual item scores is correlated for each pair of items. Items with a correlation of higher than 0.2 (absolute value) are examined for a lack of independence.
c. Unidimensionality: that all items relate to a single construct. Unidimensionality is examined using both exploratory and confirmatory factor analysis, with results compared against key metrics.

d. Anchor stability: that anchor items perform in similar ways in different administrations, given any differences in the performance of the cohort overall. Anchor items are examined for changes in facility and discrimination. The $D^2$ statistic is used to identify any items that differ in terms of their IRT parameters, by looking at differences in expected score at different points in the ability range. Additionally, detailed logs are maintained recording any changes to anchor items. Following a review of this evidence, any anchor items thought to be performing differently are unlinked in the subsequent IRT analysis.

The evidence above confirms that STA uses appropriate techniques to set the standard for the current years test and maintain the accuracy and stability of equating functions from year to year.

# Claim 5: The meaning of test scores is clear to stakeholders

### 5.1 Is appropriate guidance available to ensure the range of stakeholders – including government departments, local government, professional bodies, teachers and parents – understand the reported scores?

Before the introduction of the new NCTs (and scaled scores) in 2016, STA had a communication plan to inform stakeholders of the changes that were taking place. This included speaking engagements with a range of stakeholders at various events and regular communications with schools and local authorities through assessment update emails.

STA provides details on scaled scores on gov.uk for KS1 and KS2. This information is available to anyone but is primarily aimed at headteachers, teachers, governors and local authorities. STA also produces an end-of-term leaflet for KS1 and KS2 for teachers to use with parents.

The evidence above confirms that appropriate guidance is available to ensure the range of stakeholders understand the reported scores.

### 5.2 Are queries to the helpdesk regarding test scores monitored to ensure stakeholders understand the test scores?

Since the introduction of scaled scores in 2016, the number of queries relating to test results has steadily declined. This provides reassurance that stakeholders' understanding is improving year on year.

- 2015–2016: 642 enquiries categorised as 'scaled scores' or 'calculating overall score' (out of 1881 enquiries about results)
- 2016–2017: 299 enquiries categorised as 'scaled scores' or 'calculating overall score' (out of 1312 enquiries about results)
- 2017–2018: 251 enquiries categorised as 'scaled scores' or 'calculating overall score' (out of 1179 enquiries about results)
- 2018–2019: 117 enquiries categorised as 'scaled scores' or 'calculating overall score' (out of 1114 enquiries about results)

The evidence above confirms that queries to the helpdesk regarding test scores are monitored to ensure stakeholders understand the test scores.

### 5.3 Is media coverage monitored to ensure scores are reported as intended? How is unintended reporting addressed?

Media coverage is monitored by STA on a weekly basis, and coverage of NCTs and scores are captured as part of this. Social media is monitored by STA during test week, in part to identify any potential cases of maladministration.

In 2019 the return of results media coverage had no notable cases of misrepresentation of results.

The evidence above confirms that media coverage is monitored to ensure scores are reported as intended.

Standards & Testing Agency



About this publication:

enquiries www.education.gov.uk/contactus

download www.gov.uk/government/publications

Reference: 978-1-78957-495-1 Product code: STA/20/8512/e

Follow us on Twitter: @educationgovuk

Like us on Facebook: facebook.com/educationgovuk