Standards
& Testing
Agency

Appendix B: 2019 Validity framework

Key stage 1 Mathematics

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) mathematics national curriculum test (NCT), based on the national curriculum in England: mathematics programme of study key stage 1 and 2.
b. The content domain is defined in the KS1 mathematics test framework (Section 4, pages 8–15).
c. The content domain sets out the elements of the programme of study that are assessed in the mathematics test. Elements from the curriculum are ordered to show progression across the years. A referencing system (national curriculum references) is used in the content domain to indicate the year, the strand and the substrand, for example, '1N1' equates to year 1, strand – number and place value and substrand 1.
d. The wording used in the national curriculum references uses the exact wording from the national curriculum.
e. STA's expert test development researchers (TDRs) developed the content domain in consultation with the Department for Education (DfE) curriculum division. STA appointed two independent curriculum advisors to support the development of the mathematics NCTs.
f. STA asked a panel of education specialists to review a draft of the content domain before it was finalised. The range of stakeholders who were involved in producing the content domain gives assurance that it is appropriate.
g. 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
1C8 – with the support of the teacher	The 'with the support of the teacher' element applies only to classroom assessment.	Teacher assessment
2C1 – recall and use addition and subtraction facts to 20 fluently 2C2a – add and subtract numbers mentally 2C4 – solve problems with addition and subtraction: applying their increasing knowledge of mental methods 2C8 – solve problems involving multiplication and division, using mental methods	Mental mathematics skills cannot be directly assessed in a paper-based test since it is only possible to mark what the pupil records. For questions where only the answer is recorded, it is not possible to know the method that pupils used or how quickly they completed the question. Pupils who are fluent with numbers will be able to use their mental arithmetic skills to find efficient strategies for completing calculations under test conditions. Therefore, good mental arithmetic skills will enable pupils to recall and apply number knowledge rapidly and accurately.	Teacher assessment
2C2b – using concrete objects	The 'using concrete objects' element applies only to classroom assessment.	Teacher assessment
2S2a and 2S2b – asking questions	The 'ask questions' element is more suited to classroom assessment.	Teacher assessment

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 mathematics test. These were not significant exclusions 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 mathematics test is defined in the KS1 mathematics test framework (Section 5, pages 16–19).
b. Before developing the cognitive domain, STA reviewed the domains for similar sorts of tests. The cognitive domain for mathematics was based on the research by Hughes et al (1998)[1], Webb (1997)[2] and Smith and Stein (1998)[3].
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 different areas of cognitive demand.
d. 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.
e. TDRs made refinements to the cognitive domains based on both the consistency between TDR judgements and the comments gathered from the teacher panels. This ensured the cognitive domains published in the test frameworks were valid and usable.

[1] Hughes, S., Pollit, A. and Ahmed, A. (1998). ‘The development of a tool for gauging demands of GCSE and A-Level exam questions’. Paper presented at the BERA conference, The Queens University Belfast.

[2] Webb, L.N. (1997). ‘Criteria for alignment of expectations and assessments in mathematics and science education’. Research Monograph, No. 8., Council of Chief School Officers.

[3] Smith, M.S. and Stein, M.K. (1998). ‘Selecting and creating mathematical tasks: from research to practice’. Mathematics Teaching in Middle School, 3, pp344–350.

f. Questions within the test are rated across four classifications to inform a judgement of their cognitive demand according to the strands below:

Strand 1: Depth of understanding

This strand is used to assess the demand associated with recalling facts and using procedures to solve problems.

Questions requiring less depth of understanding require simple procedural knowledge, such as the quick and accurate recall of mathematical facts or the application of a single procedure to solve a problem.

At intermediate levels of demand, a question may require the interpretation of a problem or application of facts and procedures. However, the component parts of these questions are simple and the links between the parts and processes are clear.

At a high level of demand, a greater depth of understanding is expected. Questions may require facts and procedures to be used flexibly and creatively to find a solution to the problem.

Strand 2: Computational complexity

This strand is used to assess the computational demand of problems.

In questions with low complexity, there will be no numeric operation.

At an intermediate level of complexity, more than one numeric step or computation will be needed to solve the problem.

At a high level of complexity, questions will involve more than two processes or numeric operations.

Strand 3: Spatial reasoning and data interpretation

This strand is used to assess the demand associated with the representation of geometrical problems involving 2-D and 3-D shapes, position and movement. This strand is also used to assess the demand associated with interpreting data.

There is a low level of demand when all the resources or information required to answer the question are presented within the problem (e.g. counting the number of sides of a given 2-D shape).

At intermediate levels of demand, spatial reasoning will be needed to manipulate the information presented in the question to solve the problem (e.g. find a line of symmetry on a simple shape or interpret a 2-D representation of a 3-D shape). Pupils may need to select the appropriate information in order to complete the problem (e.g. from a table, chart or graph).

At the highest level of demand, there may be the need to use complex manipulation or interpretation of the information as part of the problem.

Strand 4: Response strategy

This strand describes the demand associated with constructing a response to a question.

At a low level of demand, the strategy for solving a problem is given as part of the presentation of the problem.

At a lower intermediate level of demand, the strategy for solving a problem is clear. Very little construction is required to complete the task.

At an upper intermediate level of demand, there may be simple procedures to follow that will lead to completion of the problem.

At a high level of demand, the question will require that a simple strategy is developed (and perhaps monitored) to complete the task. The answer may need to be constructed, organised and reasoned.

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?

319 marks were available for the 2019 KS1 mathematics test construction, comprising:

- 153 arithmetic items arithmetic (153 marks)
- 157 reasoning items (166 marks)

These items covered the content and cognitive domains as shown in Tables 2–6. A dash indicates that an element of the content domain was not available for selection, as items with this attribution cannot be developed.

Content area	Arithmetic (marks)	Reasoning (marks)	Total (marks)
Number and place value	18	30	48
Calculation	119	60	179
Fractions	16	17	33
Measurement	–	24	24
Geometry – properties of shapes	–	19	19
Geometry – position and direction	–	5	5
Statistics	–	11	11
Total	153	166	319

Table 2: Content domain

Depth of understanding	Arithmetic and reasoning (marks)
1	43
2	165
3	87
4	24

Table 3: Cognitive domain – depth of understanding

Calculation complexity	Arithmetic and reasoning (marks)
1	68
2	135
3	94
4	22

Table 4: Cognitive domain – calculation complexity

Spatial reasoning and data handling	Arithmetic and reasoning (marks)
1	259
2	19
3	30
4	22

Table 5: Cognitive domain – spatial reasoning and data handling

Response strategy	Arithmetic and reasoning(marks)
1	30
2	230
3	37
4	22

Table 6: Cognitive domain – response strategy

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 make 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 (SEND) and disability 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: STA test development researchers 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)? Is a range of questions included that are appropriate to the curriculum and classroom practice?

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 has been identified that cannot be corrected.

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, although it is available should specific issues be identified.

Please note that the items selected for the 2019 KS1 mathematics test were selected from three different TPTs: 2016, 2017 and 2018.

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 mathematics test meets the requirements of the test framework as shown in Table 7.

Content domain	Target	Previous range*	2019
Whole test	60	60	60
Number, calculations and fractions	48–51	49–51	49
Measurement, geometry and statistics	9–12	9–11	11
Arithmetic	25		
Number, calculations and fractions	25	25	25
Measurement, geometry and statistics	–	–	–
Reasoning	35		
Number, calculations and fractions	23–26	24–26	24
Measurement, geometry and statistics	9–12	9–11	11

Table 7: Test coverage of content domain

* Previous range is taken from the marks selected from the previous three years of tests: 2016, 2017 and 2018.

Teachers, subject experts, markers, inclusion experts and independent curriculum advisors reviewed the test at the expert 3 meeting on 11 October 2018. Their comments are summarised below:

a. The test provides sufficient challenge and covers the ability range with a good mix of questions in both papers.
b. Paper 1 appeared more difficult than previous years, but not unreasonably so. They were happy to see questions where the answer box appears first.
c. Paper 2 has an excellent use of visuals that adequately support the accessibility of questions.

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 in 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 from the pool of available items any 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 have 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 the data 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 quantitative and qualitative 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 and the deputy director for assessment development attended item finalisation for the

2019 KS1 mathematics 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.
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 or difficulty.

A combination of stakeholders reviewed the proposed live 2019 KS1 mathematics test at expert review 3. This group included teachers, inclusion, curriculum, assessment and mathematics 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 problems with items in the 2019 KS1 mathematics test.

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 mathematics.

STA held a one-day mark scheme finalisation meeting for the 2019 KS1 mathematics test. At this meeting, an expert group of senior markers reviewed the live test, mark scheme, 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 11 panellists, who were current 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 mark scheme on 132 pupil responses from the TPT script archive. They were not able to see how the response 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 UAT attendees' comments are summarised below:

a. The panel was content with the overall test.
b. The general marking principles (GMPs) were accessible and manageable.
c. They liked the pupil responses used, which were accessible and good examples were given.

d. The mark scheme was accessible, with a few minor amends suggested to guide teachers (e.g. adding references to GMPs in additional guidance).
e. Paper 1: arithmetic paper was straightforward to mark with additional information added to the header of the mark scheme to direct markers to specific GMPs.

The data collected from expert review 3 is then presented alongside the quantitative data for the live test at project board 3. At this board meeting, the data is scrutinised and critically challenged to ensure the test meets the expectations as stated in the test framework for KS1 mathematics.

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 mathematics test.
b. STA gives item writers a clear brief to use the relevant parts of the national curriculum document for KS1 mathematics 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.
c. During the item-writing stage, agencies conduct very small-scale trials with approximately 30 pupils 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 been 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 mathematics test covers a full range of abilities. The test is made up of a range of different cognitive domains, as specified in the test framework.
g. Tables 8–11 show that the 2019 KS1 mathematics test meets the desired coverage of all strands of the cognitive domain, as set out in the test specification.

Depth of understanding	Target	Previous range	2019
1	0–20	7–9	5
2 and 3	30–60	47	47
4	0–10	4–6	8

Table 8: Depth of understanding

Calculation complexity	Target	Previous range	2019
1	10–20	12–15	11
2 and 3	30–50	39–42	45
4	0–10	6–7	4

Table 9: Calculation complexity

Spatial reasoning and data handling	Target	Previous range	2019
1	45–55	46–50	49
2 and 3	0–15	9–11	10
4	0–5	1–3	1

Table 10: Spatial reasoning and data handling

Response strategy	Target	Previous range	2019
1	0–10	4–7	4
2 and 3	40–60	48–49	50
4	0–10	7	6

Table 11: Response strategy

h. TDRs place items in the test booklet in order of difficulty as part of the test construction process. The easiest items are at the beginning of the test and the most difficult ones are at the end. The TDR and psychometrician make decisions on the difficulty of each item using information from both classical analysis and IRT. The data on individual items helps to make up a picture of the overall test characteristics.
i. Most of the test information on ability 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 mathematics 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 panel for the 2019 KS1 mathematics test consisted of representation from hearing and visual impairment experts, a SEND representative, an EAL representative, dyslexia and dyscalculia representatives and an educational psychologist. Within this review process, panellists highlight any 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), along 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. STA acknowledges that pupils in these groups have a wide range of ability so treats 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 is biased or unfair.

e. Although none of the items available for inclusion in the 2019 KS1 mathematics test were flagged as having moderate or large DIF, the TDR and psychometrician considered the balance of items with negligible DIF at test construction alongside all other test constraints.
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 81% of the items in the 2019 KS1 mathematics braille test, the modifier used standard modification to minimally change the format of items or did not modify items at all. Sometimes an item cannot be modified in a way that maintains the construct of the original question. In producing the 2019 KS1 mathematics test, 11 items required modification that test developers felt changed the construct of the question. These items were questions 3, 6, 9, 14, 15, 16, 17, 18, 20, 21 and 28, all in Paper 2. One item, question 17 in Paper 2, could not be modified for use and so STA replaced it with a modifiable item with similar characteristics.
h. For 86% of the items in the 2019 KS1 mathematics modified large print (MLP) test, the modifier used standard modification to minimally change the format of items or did not modify items at all. In producing the 2019 KS1 mathematics MLP test, eight items required modification that test developers felt changed the construct of the question. These items were questions 3, 6, 9, 14, 15, 17, 18 and 20, all in Paper 2. One item, question 17, could not be modified for use and so STA replaced it with a modifiable item with similar characteristics.

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 mathematics 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, TDRs 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 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. It if is 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 mathematics 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 reviews 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 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 all responses that demonstrate the required understanding are credited and responses that do not demonstrate the required understanding are not credited.
h. When TDRs and psychometricians have constructed the live test, TDRs use the coding information and pupil responses from 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 mathematics test, two questions needed to be re-marked and the analysis 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 mathematics 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 the 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 mathematics 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. The TDR

may refine the wording of the mark scheme or the order of the marking points for clarity and they 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 the TDR. 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 mathematics 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 test 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 the test framework for KS1 mathematics.

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 staff within STA 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 on them for maintaining security are clearly and repeatedly stated at the start and throughout the meetings. Teacher panels will be 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 they 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 IVT or TPT, the trialling agency must adhere to the security arrangements within 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 school. They are responsible for ensuring all materials are collected after the 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 supplier's compliance with its 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 are 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 test and ensure they are able to demonstrate their attainment. Access arrangements are included to increase access without providing an unfair advantage to the 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 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; readers; prompters; rest breaks; written or oral translations; and apparatus in mathematics tests.
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 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 Test Development Subprogramme 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 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[4] 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.

[4] 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 2014 national 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 mathematics is 85–115, and all scaled scores were represented in 2019. 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 the 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 process was approved by the STA 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 the 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 as 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 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 year’s 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 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 about 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 within 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: 978-1-78957-494-4 Product code: STA/20/8511/e

Follow us on Twitter:
@educationgovuk

Like us on Facebook:
facebook.com/educationgovuk