



Standards
& Testing
Agency

Appendix E: 2019 Validity framework

Key stage 2 Mathematics

January 2020

Contents

Summary	3
Claim 1: Test is representative of the subject/national curriculum	4
Claim 2: Test results provide a fair and accurate measure of pupil performance	15
Claim 3: Pupil performance is comparable within and across schools	28
Claim 4: Differences in test difficulty from year to year are taken account of, allowing for accurate comparison of performance year on year	34
Claim 5: The meaning of test scores is clear to stakeholders	38
Table 1: Non-assessable elements of the national curriculum	5
Table 2: Number of marks available for selection in each area of the content domain	9
Table 3: Number of marks available for selection in each area of the cognitive domain	10
Table 4: Number of marks, desired range and previous range for each content domain area	13
Table 5: Desired range, previous range and number of marks in 2019 for depth of understanding	19
Table 6: Desired range, previous range and number of marks in 2019 for calculation complexity	19
Table 7: Desired range, previous range and number of marks in 2019 for spatial reasoning and data handling	19
Table 8: Desired range, previous range and number of marks in 2019 for response strategy	20

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 2 (KS2) mathematics national curriculum test based on the [national curriculum programme of study \(2014\) for mathematics at KS2](#).
- b. The content domain is defined in the [KS2 mathematics test framework](#) (Section 4, pages 8–28).
- c. The content domain sets out the elements from 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, ‘3N1’ equates to year 3, strand – number and place value and substrand 1.
- d. The wording used in the national curriculum references is the exact wording from the 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 advisers to support the development of the mathematics national curriculum tests (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.
- h. Ofqual’s [content validation study](#) for the 2016 tests stated that: ‘Given STA’s interpretation of the national curriculum framework document, the *Test Framework* documents appear to translate national curriculum teaching requirements into plausible blueprints for testing.’ Specifically, in relation to the content domain, the report stated: ‘The degree of consistency with which our independent experts rated items from the 2016 tests supports the conclusion that the way in which STA has specified both the content domain and the cognitive domain is plausible.’ This appears to support the claim that the content domain was appropriate and clear.

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
3C1, 5C1, 4C6, 5C5c, 5C6a, 6C6 and 6F11 – mental arithmetic skills	<p>Mental mathematics skills cannot be directly assessed in a paper-based test since only the pupil’s responses can be marked. For questions where only the answer is recorded, it is not possible to know the method that the pupil used or how quickly he or she completed the question.</p> <p>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.</p>	Teacher assessment
3G3b – make 3-D shapes using modelling materials	Requires practical equipment to assess validly	Teacher assessment
5M8 – estimate capacity (e.g. using water)	Requires practical equipment to assess validly	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 a small number of elements of the national curriculum from the content domain for the KS2 mathematics test. This is not a significant exclusion.

Ofqual's [content validation study](#) for the 2016 tests stated: 'Given STA's interpretation of the national curriculum framework document, the *Test Framework* documents appear to translate national curriculum teaching requirements into plausible blueprints for testing.'

The evidence above confirms that 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 KS2 mathematics test is defined in the [KS2 mathematics test framework](#) (Section 5, pages 29–32).
- 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 inter-rater consistency between TDRs and the comments gathered from the teacher panels.

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.

This ensured that the cognitive domains published in the test frameworks were valid and usable.

- f. Questions within the test are rated across four strands to inform a judgement of their overall cognitive demand.

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 that facts and procedures will need 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.

In questions with lower complexity, it is likely that a one-step process such as recalling a mathematical fact or the application of a simple procedure will be sufficient to solve the problem.

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 have multiple numeric steps or computations and the links between them within the problem will be complex.

Strand 3: Spatial reasoning and data handling

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

There is a low level of demand when all of the resources or information required to answer the question are presented within the problem (e.g. finding the perimeter of a shape by adding the length of the sides).

At intermediate levels of demand, spatial reasoning will be needed to manipulate the information presented in the question to solve the problem (e.g. reflecting a polygon in a mirror line). 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 spatial reasoning to interpret, infer or generate new information from that given before the problem can be completed (such as identifying 3-D characteristics from 2-D representations or making inferences from the given information).

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 answering a problem is clear and following simple steps will lead to completion of the problem. Some reasoned construction may be associated with organising appropriate working.

At an upper intermediate level of demand, there may be the need to construct a straightforward response. Some reasoned construction may be associated with organising more complex working.

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

Ofqual carried out an independent review of the cognitive domains as part of its investigation into the 2016 tests. Its [content validation study](#) stated: 'The degree of consistency with which our independent experts rated items from the 2016 tests supports the conclusion that the way in which STA has specified both the content domain and the cognitive domain is plausible.' This appears to support the claim that the cognitive domain was appropriate and clear.

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?

794 items were available for the 2019 KS2 mathematics test construction, comprising:

- 379 arithmetic items (419 marks)
- 415 reasoning items (474 marks)

These items covered the content and cognitive domains as shown in tables 2 and 3. A dash indicates that an element of the content domain was not available for selection, as items with this attribute cannot be developed.

Content area	Arithmetic (marks)	Reasoning (marks)	Total (marks)
Number	15	69	84
Calculation	242	101	343
Fractions	140	75	215
Ratio	22	12	34
Algebra	–	22	22
Measures	–	71	71
Geometry	–	38	38
Position	–	33	33
Statistics	–	53	53
Total	419	474	893

Table 2: Number of marks available for selection in each area of the content domain

Depth of understanding	Arithmetic (marks)	Reasoning (marks)	Total
1	319	58	377
2	100	276	376
3	–	112	112
4	–	28	28
Calculation complexity	Arithmetic (marks)	Reasoning (marks)	Total
1	9	106	115
2	294	272	566
3	116	82	198
4	–	14	14
Spatial reasoning and data handling	Arithmetic (marks)	Reasoning (marks)	Total
1	419	308	727
2	–	101	101
3	–	40	40
4	–	25	25
Response strategy	Arithmetic (marks)	Reasoning (marks)	Total
1	339	283	622
2	–	104	104
3	80	69	149
4	–	18	18

Table 3: Number of marks available for selection in each area of the cognitive domain

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.

Please note that the items selected for the 2019 KS2 mathematics test were selected from three different TPTs: 2016, 2017 and 2018. The TPT 2016 items went through a two-trial model and were fully reviewed at expert reviews 1 and 2. The 2017 and 2018 TPT items went through the one-trial model and were fully reviewed at least once.

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)?

STA holds an item finalisation meeting involving TDRs and psychometricians after each trial. The purpose of this meeting is to review all the evidence on each item to make decisions about the next stage of development. 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.

The 2016 items selected for the 2019 KS2 mathematics test went through an IVT. After an IVT, the following outcomes are available for each item:

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

All items selected for the KS2 mathematics 2019 test went through a TPT. After a TPT, the following outcomes are available for each item:

- a. Item is available for inclusion in a live test: the evidence shows it is performing as intended.
- b. Item requires minor amends: it will need to be retrialled before inclusion in a live test.
- c. Item is archived: 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. STA keeps the evidence relating to the review and validation of individual items in its item bank.

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 KS2 mathematics test meets the requirements of the test framework as follows:

Content domain	Target	Previous range*	2019
Whole test	110	110	110
Number, ratio and algebra	83–93	85–88	84
Measurement, geometry and statistics	17–27	22–25	26
Arithmetic	40	40	40
Number, ratio and algebra	40	40	40
Measurement, geometry and statistics	–	–	–
Reasoning	70	70	70
Number, ratio and algebra	44–52	45–48	44
Measurement, geometry and statistics	17–27	22–25	26

Table 4: Number of marks, desired range and previous range for each content domain area

* 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 expert review 3 on 3 October 2018. Their comments are summarised below:

- a. There was a positive reception of the proposed test.
- b. Paper 1 was considered to be challenging but this was welcomed; the panel felt the start of the paper was very positive with the answer first, which was a consideration STA made at test construction.

- c. Papers 2 and 3 cover the breadth of the curriculum with different areas being assessed ('loved the range'). Although demanding, there was nothing adverse ('a good level of challenge'; 'not a problem').
- d. They welcomed the three 'explain' questions and thought that most of the questions assessed a good level of reasoning (relational understanding), not just regurgitated facts. Many members of the panel thought the marking was very fair.
- e. It was felt by the panel that paper 2 should be paper 3 (and vice versa) because of the perceived difficulty of paper 2 following paper 1. They felt this would have a negative effect on pupils and cause 'backlash'. STA made this change prior to project board 3.

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 retrialled 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. 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 reviews 1*, 2 and 3); 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.

* The TPT 2016 items selected for the 2019 KS2 mathematics test were reviewed by at least three independent expert review panels because they were developed under a two-trial model; the 2017 and 2018 TPT items were reviewed by expert panels twice as they were developed under a one-trial model.

** The TPT 2016 items selected for the 2019 KS2 mathematics test were trialled at least twice with an IVT and a TPT because they were developed under a two-trial model; the 2017 and 2018 TPT items were developed under a one-trial model and were trialled at least once at TPT.

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.

Qualitative data is collected during the TPT:

- 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. TDRs summarise the information and presents it at an item finalisation meeting.

The senior test development researcher (STDR), TDRs, the senior psychometrician and the deputy director for assessment development attended item finalisation for the 2019 KS2 mathematics test. The attendees considered the information the TDR presented and decided whether items were suitable for live test construction.

TDRs and psychometricians select items for live test construction based on the outcomes of item finalisation. They use qualitative data to confirm that the items selected were suitable. TDRs and psychometricians consider 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.

Based on this evidence, TDRs amend or archive items that are not deemed acceptable.

A combination of stakeholders reviewed the proposed live 2019 KS2 mathematics test at expert review 3. This group included teachers and inclusion, curriculum, assessment and mathematics experts. At this meeting, panellists could challenge items and the TDR used the item data to either defend that challenge or support it. If the panel deemed an item unacceptable, the TDR could swap it with a suitable item from the TPT. In the proposed 2019 KS2 mathematics test, the panel identified one item that was an enemy with another item in the paper. The TDR proposed an alternative item, which the expert review panel agreed was suitable.

TDRs collate the data from expert review 3 and present 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 KS2 mathematics.

STA held a one-day mark scheme finalisation meeting for the 2019 KS2 mathematics test. At this meeting, an expert group of senior markers reviewed the live test, mark scheme and responses from the trialling, and suggested improvements to the mark scheme to ensure that markers can apply it accurately and reliably. These amendments do not affect the marks awarded for each question.

After this meeting, STA and the external marking agency used the amended mark scheme and the trialling responses to develop marker training materials. The purpose of these materials is to ensure that markers can consistently and reliably apply the mark scheme.

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 and STA TDRs wrote the items that make up the 2019 KS2 mathematics test.
- b. STA gives item writers a clear brief to use the relevant parts of the national curriculum document for KS2 mathematics when writing their items. This ensures that the items are age appropriate as they are based on a curriculum that experts have deemed suitable.
- c. During the item-writing stage, agencies conduct very small-scale trials with approximately 18–45 pupils who are in Year 6 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 STA 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 advisers, 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 KS2.
- 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 KS2.

- f. The 2019 KS2 mathematics 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 tables below show that the 2019 KS2 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	33–65	36–38	36
2 and 3	32–70	65–68 (2) 39–42 (3) 23–26	69 (2) 44 (3) 25
4	0–20	5–9	5

Table 5: Desired range, previous range and number of marks in 2019 for depth of understanding

Calculation complexity	Target	Previous range	2019
1	1–15	6–10	11
2 and 3	77–105	97–101 (2) 59–77 (3) 24–38	95 (2) 69 (3) 26
4	2–20	3–5	4

Table 6: Desired range, previous range and number of marks in 2019 for calculation complexity

Spatial reasoning and data handling	Target	Previous range	2019
1	65–95	82–95	89
2 and 3	15–35	13–27 (2) 9–15 (3) 4–12	12 (2) 8 (3) 4
4	0–10	0–2	9

Table 7: Desired range, previous range and number of marks in 2019 for spatial reasoning and data handling

Response strategy	Target	Previous range	2019
1	40–60	57–60	57
2 and 3	30–70	48–52 (2) 15–21 (3) 31–33	50 (2) 22 (3) 28
4	1–10	1–2	3

Table 8: Desired range, previous range and number of marks in 2019 for response strategy

- g. 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. TDRs and psychometricians 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.
- h. 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 the higher end of the ability range too.

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 qualitative and quantitative evidence has been used 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 KS2 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 KS2 mathematics test consisted of representation from the autistic spectrum disorder (ASD) specialism, cultural reviews, SEND, behavioural issues and EAL. Within this review process, panellists highlight any potential bias and suggest ways to remove it. TDRs consider all the available evidence and present 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. TDRs use 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 DIF analysis 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 KS2 mathematics test were flagged as having moderate or large DIF, TDRs and psychometricians 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 a minimal change in the item's difficulty.
- g. For 93% of the items in the 2019 KS2 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. In producing the 2019 KS2 mathematics test, six items required modification that test developers felt changed the construct of the question. These were in questions 21 and 22 (parts a and b) in paper 2 and questions 9, 10 and 22 in paper 3. There were no items that could not be modified for use. However, STA replaced the line graph for question 22 in paper 2 with a table, following advice from the modifier.
- h. For 93% of the items in the 2019 KS2 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 KS2 mathematics MLP test, six items required modification that test developers felt changed the construct of the question: questions 21 and 22 (parts a and b) in paper 2 and questions 9, 10 (part b) and 22 in paper 3. There were no items that could not be modified for use. However, STA replaced the line graph for question 22 in paper 2 with a table, following advice from the modifier.

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 KS2 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 advisers 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 (e.g. 'explain') 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 KS2 mathematics test. TDRs use the trialling coding frames to inform the content of the mark schemes and select 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. TDRs and the external marking agency use the script archive and the mark schemes to develop marker training materials. Marker training materials comprise pupil scripts with pre-agreed marks in the form of:
- training scripts – used to train markers on the application of the mark scheme
 - practice scripts – used in training sessions for markers to practise marking specific points
 - standardisation scripts – used to check that markers are applying the mark schemes correctly before they can start marking
 - validity scripts – used during the marking window to ensure that markers continue to apply the mark scheme correctly
- d. STA holds a mark scheme finalisation meeting, composed of TDRs, the psychometrician, independent curriculum advisors, senior trialling coders, senior markers and representatives from the external marking agency. 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.
- e. After this meeting, STA and the external marking agency use the amended mark scheme and the trialling responses to develop marker training materials. The purpose of these materials is to ensure that the mark scheme is robust and that markers can apply it reliably and consistently.
- f. External markers, who receive training to ensure that the agreed mark schemes are applied consistently and fairly, mark KS2 tests. The external marking agency creates the training materials in consultation with the STA TDR, the STA psychometrician and the independent curriculum advisors.
- g. The training suite includes question-specific training on how to apply the mark scheme and exemplar responses drawn from the trialling process. Markers must successfully complete a standardisation exercise on each question to check that they are able to apply the mark scheme correctly and consistently before they are able to mark. During the marking period, the external marking agency checks the accuracy of marking using responses for which the senior marking team, STA TDR and psychometrician have agreed marks. Markers must demonstrate their accuracy by marking these scripts to the agreed standard to be able to continue marking on that question.

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 STA 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 that 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 as well as 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 that 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 of 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 that coding frames reflect pupil responses.
- e. A second trial – the TPT – is administered, 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 that 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 which 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 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 KS2 mathematics

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 KS2 mathematics 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 review the coding frames before they are used. This group includes the STA curriculum advisors, the psychometrician 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 trialling agency provides an archive of scanned pupil scripts and STA's 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 KS2 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. 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 STA TDRs, STA

psychometricians, independent curriculum advisors, the senior marking team and representatives from the external marking agency. 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. Following the mark scheme finalisation meeting, STA provided the external marking agency with the TPT script archive. The marking agency used this to develop a suite of training materials to support the external marking of the 2019 KS2 mathematics test. Working with TDRs, STA psychometricians and the external curriculum advisors, the external marking agency developed the materials over a series of meetings. The marking agency designed the training materials to further exemplify the marking principles stated in the mark scheme and to support markers in applying these principles.
- i. To ensure quality of marking, the marking agency selects standardisation scripts. Following their training, markers must score these scripts accurately before they are released to mark. To ensure ongoing accuracy of marking, the external marking agency selects validity items. The senior marking team selects and marks these scripts, which TDRs review. The marking agency then uses these scripts during the live marking period to check that markers are applying the mark scheme accurately and consistently. As part of the development of the marker training materials, a panel of markers conducts a user acceptance test (UAT) of how accurately markers can apply the mark scheme, exemplified by the training materials, to the marking of the practice and standardisation scripts. For the 2019 KS2 mathematics test, this UAT highlighted that three of the scripts needed to be replaced.

The evidence above confirms that 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?

To ensure that markers apply the mark schemes as intended, STA follows these processes:

- the development of mark schemes, as previously outlined
 - the training process, as previously outlined
 - the quality assurance process during marking
 - the quality assurance process following marking
- a. At the training meeting, the markers see examples for each item (the numbers vary according to item difficulty) and receive training on how the mark scheme should be applied.
 - b. Each marker then completes a number of practice scripts for each item. Supervisory markers provide feedback on their performance to ensure that they understand how to apply the mark scheme.

- c. Before they are allowed to mark, markers must complete a set of five standardisation scripts for each item. This checks that their marking agrees with the agreed marks for that item. Supervisory markers provide feedback on their performance to address any issues and, if necessary, markers may complete a second set of five standardisation scripts.
- d. The external marking agency undertakes quality assurance during live marking by placed validity items. These items have a predetermined code agreed by TDRs and lead/deputy coders, and they appear randomly in each batch of marking. The marking agency will suspend a marker from that item if there is a difference between the agreed mark and the mark awarded by the marker. The marker cannot continue marking that item until they have received further training on the mark scheme. If a marker fails for a third time, the external marking agency stops them from marking that item.
- e. During live marking, markers may flag any responses that do not fit the mark scheme for guidance from their supervisory marker. A marker may also check their marking with their supervisory marker. If necessary, supervisory markers may escalate queries to lead markers or TDRs to be resolved.
- f. Supervisory markers quality-assure the marking and provide feedback to markers if they make errors in applying the mark scheme. The supervisory marker may review complete batches at any time, if necessary, and may ask markers to review their submitted batches to update marks after receiving additional training. The supervisory marker may follow the agreed procedures for stopping a marker at any point if they have concerns about accuracy.
- g. After live marking is complete, the external marking agency provides STA with a report on marking. This report contains some qualitative data on the quality of marking for each item.
- h. If schools dispute the application of the mark scheme, they can send pupil scripts for review with a senior marker.

The evidence above provides a summary of how STA ensures markers are applying 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. STA test development identifies bias 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 socio-economic status.
- b. STA collects quantitative data 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. STA considers the fairness, accessibility and bias of each test question in three rounds of expert reviews.⁴ Teacher panels, test review groups (TRGs) and inclusion panels (comprising SEND, EAL, culture/religion and educational psychology experts) scrutinise questions, their answering strategies and their contexts. TDRs further examine any questions that raise concerns about bias or unfairness to either minimise the bias or remove the question from the test if no revision is possible.
- d. STA produces modified tests, including braille and large print versions, for pupils who are unable to access the standard NCTs. The content of the modified test is kept as close to the original as possible. This is to rule out test-critical changes and to prevent any further bias from being introduced through modification. STA consults modification experts throughout the test development process to ensure that no bias is introduced.
- e. Further information about diversity and inclusion in the NCTs can be found in the [KS2 mathematics test framework](#) (Section 7, page 41).

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

⁴ For the 2019 KS2 mathematics test, the TPT 2016 selected items have been through at least three expert reviews and the 2017 and 2018 TPT items have been reviewed at least twice.

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. STA asks external stakeholders to review test items throughout the test development process, usually through expert reviews. Anyone who is involved in expert review panels is required to sign confidentiality forms, and the requirements for maintaining security are clearly and repeatedly stated at the start and throughout the meetings. At teacher panels, attendees receive a pack of items in the meeting to comment on. They return this pack to STA at the end of the day. TRGs review the items in advance of the meeting. STA sends items to TRG members via STA's approved parcel delivery service and provides them with clear instructions on storing and transporting materials. TRG members return their materials to STA at the end of the meeting.
- c. When items are trialled as part of the IVT or TPT, the trialling agency must adhere to the security arrangements that are specified 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 are required to 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 collecting all materials after the visit before using the approved courier to return them to the trialling agency.
- d. STA outsources all print, collation and distribution services for NCTs 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 (DSAM) assessment. This ensures all aspects of IT/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 for the delivery of materials to schools.
- e. STA provides schools with guidance about handling NCA test materials securely, administering the tests, using access arrangements appropriately and returning KS2 test scripts for marking. Local authorities (LAs) have a statutory duty to make monitoring visits to at least 10% of their schools that are participating in the phonics screening check and KS2 tests. These visits are unannounced and may take place before, during or after the test or check periods. The monitoring visits check that schools are storing materials securely, administering tests correctly, and packaging and returning materials as required. At the end of the administration, headteachers must complete a statutory headteacher's declaration form (HDF) and submit it to STA. The HDF confirms that either the tests or checks

have been administered according to the published guidance or that any issues have been reported to STA.

- f. Each year approximately 4600 markers are involved in the marking of KS2 tests. The marking agency restricts the number of markers who have access to live materials prior to test week in order to maintain the confidentiality of the test material, while still allowing for an adequate test of the developed marker training materials and also ensuring a high-quality marking service is achieved. Approximately 20 senior supervisory markers have access to live test content from the November before the test in May in order to develop marker training material. Around six supervisory and non-supervisory markers take part in UAT of the developed training materials during January/February. The external marking agency gives around 500 supervisory markers access to materials in March/April before the tests are administered in May. This is to enable supervisory markers to receive training in their use, ahead of their training of marker teams. The remaining 4100 markers receive their training following the administration of the tests.
- g. Markers must sign a contract that stipulates the requirement to maintain the confidentiality of materials before they have sight of any material. Confidentiality of material is emphasised to all markers at the start of every meeting/training session. The external marking agency will not offer a supervisory role to markers if their own child is to sit the KS2 tests, although they may receive a standard marking contract. This ensures that they do not see materials until after the tests have been administered.
- h. The external marking agency holds marker training events at venues that meet agreed venue specifications, which ensures that they comply with strict security procedures.

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. STA develops this guidance to ensure consistency of administration across schools.

The LA make unannounced monitoring visits to a sample of schools administering the tests. They will check whether the school is following the published test administration guidance on:

- keeping the KS2 test materials secure
- administering the KS2 tests
- packaging and returning KS2 test scripts

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

- administrative irregularities
- potential maladministration

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

The evidence above shows how STA provides guidance on administration and monitors that it is available, understood and implemented consistently 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 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 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 KS2 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. 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?

The external marking agency carries out a range of checks to ensure that only markers who demonstrate acceptable marking accuracy and consistency mark NCTs.

- a. Following training, markers must complete a set of five standardisation scripts for each item before receiving permission to mark that item. This checks that their marking agrees with the agreed marks for that item. If their Absolute Mark Difference (AMD)⁵ for any one item is outside the agreed level of tolerance, the marker will have failed standardisation. A supervisory marker will provide feedback and the marker may complete a second set of five standardisation scripts. This step ensures that markers who cannot demonstrate accurate application of the mark scheme and the marking principles will not take part in live marking.
- b. The external marking agency carries out quality assurance during live marking through the placement of validity items. These items have a predetermined code agreed by TDRs and lead/deputy coders and appear randomly in each batch of marking. The external marking agency will suspend a marker from marking an item if there is a difference between the agreed mark and the mark awarded by the marker. The marker cannot continue marking that item until they have received further training on the mark scheme. If a marker fails for a third time, the external marking agency stops them from marking that item. If the marking agency stops a marker from marking an item, they will redistribute that marker's items for other markers to re-mark.
- c. This process ensures that all markers are applying the mark scheme accurately and consistently throughout the marking window and is the standard approach to ensuring the reliability of marking.
- d. The external marking agency and STA TDRs and psychometricians set AMD bands and validity thresholds for each item so that markers can be monitored to ensure that marking errors are minimised and within an acceptable level. In 2019 the maths AMD bands and validity thresholds were set by the Marking Programme Leader and agreed with the TDR and the psychometrician. These were based on the complexity of the items and the number of marks available.

The evidence above provides a summary of the processes and procedures that measure reliability, consistency and accuracy of marking.

⁵ The AMD is the difference between the marks awarded to an item on a standardisation set by a marker and the predetermined definitive mark assigned by the senior marking team.

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 which 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 analysis⁶ 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. STA predicted the threshold of the expected standard at project board 3 to be 55. The final threshold for the 2019 KS2 mathematics test was 58.

The evidence above confirms that TPT data accurately predicts performance on the live test.

⁶ 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 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 KS2 mathematics is 80–120, and only 119 was not 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 which 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 draft test is established, additional analysis is carried out to scale the parameters to the 2016 scale in order to produce a draft scaled score conversion (SSC) table which estimates the likely standard for the test.

For KS2, once the test has been administered and the live data is available, this analysis is run again (this time including the live data) to obtain the final scaled score outcome, which also helps STA to judge the accuracy of its previous estimation.

The approach is fit for purpose. The process mirrors that completed for the final standards maintenance exercise, which was approved by the STA Technical Advisory Group in 2017.

The evidence above 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 that 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. Similarly, if the overall difficulty increases, then the raw score needed to meet the standard will decrease. Each raw score point is associated with a point on the ability range which is converted to a scaled score point from 80 to 120.

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 as well as data from TPTs, including anchor items which are repeated each year as well as the items which are 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 was 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?

Prior to 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 it is primarily aimed at headteachers, teachers, governors and LAs. 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

© Crown copyright 2020

This publication (not including logos) is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

To view this licence:

visit www.nationalarchives.gov.uk/doc/open-government-licence/version/3

email psi@nationalarchives.gsi.gov.uk

write to Information Policy Team, The National Archives, Kew, London, TW9 4DU

About this publication:

enquiries www.education.gov.uk/contactus

download www.gov.uk/government/publications

Reference: 978-1-78957-497-5 Product code: STA/20/8514/e



Follow us on Twitter:
[@educationgovuk](https://twitter.com/educationgovuk)



Like us on Facebook:
facebook.com/educationgovuk