Standards
& Testing
Agency

Appendix G: 2019 Validity framework

Key stage 2 English grammar, punctuation and spelling

January 2020

Contents

Summary

The validity frameworks are an 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) English grammar, punctuation and spelling (GPS) national curriculum test (NCT), based on the national curriculum programme of study (2014) for English at KS2.
b. The content domain is defined in the KS2 English GPS test framework (Section 4, pages 7–13).
c. The content domain sets out the elements of the programme of study that are assessed in the KS2 English GPS test. The content domain is derived from the English programme of study for writing – vocabulary, grammar and punctuation (Appendix 1: Spelling and Appendix 2: Vocabulary, grammar and punctuation).
d. STA grouped the elements from the curriculum into content domains, which are divided into subdomains. The areas covered under 'vocabulary' are the parts of the content domain that relate to words and word building.
e. Over time, the tests will sample from each area of the content domain.
f. 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 English GPS NCTs.
g. STA asked a panel of education specialists to review a draft of the content domain before it was finalised. The range of stakeholders that was involved in producing the content domain gives assurance that it is appropriate.
h. 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.

National curriculum reference	Explanation
English Appendix 2, Year 3, text: Introduction to paragraphs as a way to group related material Headings and subheadings to aid presentation	These statements are better suited to being assessed as part of teacher assessment of writing. They could be partially assessed in a test in terms of asking pupils how texts are organised (both in paragraphs and through headings); it would not show how well the pupils use this skill in their own writing, without a longer writing task.
English Appendix 2, Year 4, text: Use of paragraphs to organise ideas around a theme	This statement is better suited to being assessed as part of teacher assessment of writing. This could be partially assessed in a test in terms of asking pupils how texts are organised (both in paragraphs and through headings); it would not show how well the pupils use this skill in their own writing, without a longer writing task.
English Appendix 2, Year 5, text: Devices to build cohesion within a paragraph Linking ideas across paragraphs using adverbials of time, place and number or tense choices	These statements are better suited to being assessed as part of teacher assessment of writing. They could be partially assessed in a test in terms of asking pupils what devices are working to ensure cohesion; it would not show how well the pupils use cohesive devices in their own writing, without a longer writing task.
English Appendix 2, Year 6, text: Linking ideas across paragraphs using a wider range of cohesive devices: repetition of a word or phrase,	These statements are better suited to being assessed as part of teacher assessment of writing. This could be partially assessed in a test in terms of

grammatical connections, e.g. the use of adverbials and ellipses Layout devices	asking pupils what devices are working to ensure cohesion; it would not show how well the pupils use cohesive devices or manipulate their own writing for effect, without a longer writing task.

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 English GPS test. This is not a significant exclusion 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 KS2 English grammar, punctuation and spelling test is defined in the KS2 English grammar, punctuation and spelling test framework (Section 5: Cognitive domain, pages 18–22).
b. Before developing the cognitive domain, STA reviewed the domains for similar sorts of test. STA derived the cognitive domain for the English GPS test from

sources including the work of Bloom[1] (1956) and Hughes *et al.* (1998)[2] because these could be adapted to align closely with the types of questions used in the test. Furthermore, the work of Bloom is widely used and understood in the classroom and so is familiar to teachers; the work of Hughes *et al.* is widely used in considering the cognitive demand of examination questions.

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. This ensured that the cognitive domains published in the test frameworks are valid and usable.
f. Questions within the test are rated across four classifications (detailed in tables 3–6) to inform a judgement of their overall cognitive demand, as in table 2.

[1] Bloom, B., Engelhart, M., Furst, E., Hill, W., and Krathwohl, D. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company

[2] 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, Queen's University Belfast

Classification	Description	Ratings scale
Cognitive level	A three-point scale indicating the degree of cognitive complexity associated with the operation required by the question	1 (low) – 3 (high)
Response complexity	A four-point scale, subcategorising the selected and constructed question formats used for the test, according to their respective levels of demand	1 (low) – 4 (high)
Abstraction rating	An indication of the familiarity of the question's vocabulary and context for the test population	1 (low) – 3 (high)
Strategy support rating	An indication of the support offered within the question and the extent to which pupils need to organise and strategise their own response	1 (low) – 3 (high)

Table 2: Cognitive classifications

The cognitive level is classified within a three-point taxonomy as in table 3.

Cognitive level	**Knowledge and comprehension** (low cognitive demand)	**Application and analysis** (mid cognitive demand)	**Synthesis and evaluation** (high cognitive demand)
Explanation	Remembers learnt information and demonstrates an understanding of the facts. Identifies linguistic features and understands their use.	Applies knowledge to given linguistic contexts. Can categorise and analyse examples of language.	Compiles component ideas or proposes alternative solutions. Makes comparisons and judgements about the uses of language and punctuation.
Example question stems	*What is the name of the punctuation mark below?* *Circle two … in this sentence.*	*Complete the sentence below with a … that makes sense.* *Categorise these into ...* *Re-write the sentence below.*	*What would be the effect of replacing this … with …?*

Table 3: Cognitive level

The response complexity is considered within a scale that ranges from closed to extended response formats, subcategorised into a number of types and in table 4.

Response format	**Selected response**	**Constructed response: data transformation**	**Constructed response: prompted**	**Constructed response: independent**
Explanation	Selecting the correct response or identifying a feature from a given field of data.	Transforming a given word, phrase or sentence.	Inserting a word or phrase within a given target sentence, following a specific prompt.	Open response, without a prompt or frame within which to write.
Example question stems	*Put a tick to show ...* *Circle all the … in the sentence below.*	*Re-write the sentence below, changing it to [past] tense.* *Replace the underlined words with a [contraction].*	*Add an [adjective] to complete the sentence.*	*Write a statement about …* *Explain why a … is needed in the sentence below.*

Table 4: Response strategy

The abstraction rating is an indicator of the familiarity of the question for the test population. It takes into account the concreteness or abstractness of the concepts involved and the likely familiarity of the vocabulary and context for the test population as in table 5.

Abstraction rating	**1** (low abstraction demand)	**2** (mid abstraction demand)	**3** (high abstraction demand)
Description	The vocabulary and context can reasonably be assumed to be highly familiar to the majority of children taking the test.	The vocabulary and context may fall outside the child's immediate personal experience but are nonetheless familiar through coverage in the primary curriculum, children's literature or the media.	The vocabulary and context will be the least familiar and are likely to be outside the direct experience of those sitting the test.
Examples of contexts or vocabulary	School-based situations Domestic and family scenarios Food Animals Items of clothing Colours Public transport Hobbies	Topics covered in other primary curriculum subjects, e.g. science and nature, significant periods of history. Visits, e.g. school trips, parks, libraries, transport, beaches.	Low-frequency spellings/vocabulary. Appropriate adult scenarios, e.g. workplaces that children rarely encounter.

Table 5: Abstraction rating

The strategy support rating indicates the extent to which the pupil must arrive independently at an understanding of the question requirements, response method and answer format as in table 6.

Strategy support rating	**1** (high support rating)	**2** (mid support rating)	**3** (low support rating)
Description	Indicates questions that provide the child with a high level of support. This may be an exemplar response that fully models the process and answer format required, and that can effectively be transposed to the child's own response.	Indicates questions including a partial level of support. This may be an explanation of some technical terminology included in the question, or an example to follow which partially shows the method or expected result, but is not fully transferrable to the child's own response.	Indicates questions that do not include any support, and in which the child is therefore required to interpret the vocabulary, method and expected answer format independently.

Table 6: Strategy support rating

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?

322 items were available for the 2019 KS2 English GPS test construction.

There were 212 marks available for Paper 1: questions, which covered the content domain and cognitive domains as shown in tables 7–11:

Content domain area	Number of marks available
Grammar	120
Punctuation	57
Vocabulary	35

Table 7: Content domain coverage of items available for Paper 1: questions

Cognitive level	Number of marks available
Knowledge and comprehension	104
Application and analysis	99
Synthesis and evaluation	9

Table 8: Cognitive level coverage of items available for Paper 1: questions

Response complexity	Number of marks available
Selected response	154
Constructed response: data transformation	19
Constructed response: prompted	25
Constructed response: independent	14

Table 9: Response complexity coverage of items available for Paper 1: questions

Abstraction rating	Number of marks available
1	115
2	81
3	0
n/a (no linking context)	16

Table 10: abstraction rating coverage of items available for Paper 1: questions

Strategy support rating	Number of marks available
1	0
2	117
3	95

Table 11: Strategy support rating coverage of items available for Paper 1: questions

There were 110 marks available for Paper 2: spelling, which covered the content domain as shown in table 12:

Content domain reference	Number of marks available
S38	9
S39	2
S40	5
S41	5
S42	4
S43	4
S44	4
S45	3
S46	8
S47	9
S48	4
S49	3
S50	3
S51	4
S52	3
S53	3
S54	3
S55	5
S56	6
S57	5
S58	4
S59	6
S60	3
S61	5

Table 12: Content domain coverage of items available for Paper 2: spelling

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

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

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

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

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

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

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

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

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

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.

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. 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 English GPS test meets the requirements of the test framework as follows:

	Target	2019	Previous range
Grammar marks	25–35	27	28–30
Punctuation marks	10–20	16	15–17
Vocabulary marks	3–7	6	5

Table 13: Desired range and number of marks for each content domain area over time

Teachers, subject experts, markers, inclusion experts and independent curriculum advisers reviewed the test at expert review 3 on 10 October 2018. Their comments are summarised below:

a. The test provides good coverage of the curriculum and there is a good balance of question types and contexts.
b. The demand of the paper is comparable to previous years.
c. The test is accessible to the majority of pupils taking the standard version of the test, but the requirement for correct spelling in the short answer paper still troubled one panellist.
d. The questions are clear and unambiguous.
e. The majority of the panel thought that the questions were technically accurate.
f. They did not find any strong enemies (questions that are too similar for inclusion in the same test).

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

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

Qualitative data is collected during the TPT, 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 psychometrician and the deputy director for assessment development attended item finalisation for the 2019 KS2 English grammar, punctuation and spelling 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 used 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 English GPS test at expert review 3. This group included teachers and inclusion, curriculum, assessment and English 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. The panel did not identify any items in the 2019 KS2 English GPS test that needed to be swapped.

TDRs collate 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 KS2 English GPS.

STA held a one-day mark scheme finalisation meeting for the 2019 KS2 English GPS test. At this meeting, an expert group of senior markers reviewed the live test and mark scheme and responses from 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 English grammar, punctuation and spelling test.
b. STA gives item writers a clear brief to use the relevant parts of the national curriculum document for KS2 English GPS 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 20 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 English GPS test covers the full range of abilities within the key stage. The test is made up of a range of different cognitive domains, as specified in the test framework. The 2019 KS2 English GPS test meets the desired coverage of all strands of the cognitive domain, as set out in the test specification. The easier cognitive domains (knowledge and comprehension) make up 66% of the test, while application and analysis and synthesis and evaluation make up 31% and 3% of the test respectively.
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 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 English grammar, punctuation and spelling 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 English GPS 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. One item in the 2016 TPT, two items in the 2017 TPT and one item in the 2018 TPT were flagged as having moderate DIF and were excluded from selection for the 2019 test. 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 minimal change in the item's difficulty.
g. For the 2019 KS2 English grammar, punctuation and spelling braille test, the modifier used standard modification to minimally change the format of items.

Sometimes the modifiers are unable to modify an item in a way that maintains its original construct. None of the items in the 2019 KS2 English GPS braille test required modifications that changed the construct of the question and STA did not have to replace any of the items.

h. None of the items in the 2019 KS2 English GPS modified large print test (MLP) required modifications that changed the construct of the question. The modifier was able to modify all of the items, mostly using standard modifications to minimally change the format, and STA did not have to replace any items.

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

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

The following list demonstrates how STA interrogates pupil responses.

a. STA collects pupil responses for the KS2 English grammar, punctuation and spelling 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 as to 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 to 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 English GPS 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. 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, psychometricians, independent curriculum advisers, 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 STA TDRs, psychometricians and the independent curriculum advisers.
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, TDRs and psychometricians 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 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 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 English GPS test, one question (question 36) had a marking change. The TDR re-marked the item and found three scripts where the mark would change from 0 to 1. This change is minimal, 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 English GPS test using coding frames that were used in the trialling process. STA uses coding frames to capture the range of responses that pupils give, both creditworthy and non-creditworthy. This allows TDRs to understand how effective an item is and to identify any issues that could affect the accuracy of marking.
b. TDRs draft initial coding frames, which are refined during expert review and trialling. A range of stakeholders reviews the coding frames before they are used. This group includes the STA curriculum advisers, 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. TDRs may 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 English GPS test was constructed, TDRs used the coding information and pupil responses from the TPT to draft mark schemes. To maintain the validity of the data collected from the TPT, STA makes only minor amendments between the TPT coding frame and the live mark scheme. TDRs may refine the wording of the mark scheme or the order of the marking points for clarity and may include exemplar pupil responses from the script archive.
g. STA holds a mark scheme finalisation meeting, composed of TDRs, STA psychometricians, independent curriculum advisers, 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 English GPS test. Working with TDRs, STA psychometricians and the external curriculum advisers, 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 English grammar, punctuation and spelling test, this UAT highlighted that none 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 scheme 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, 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. TDRs and psychometricians consider the DIF values during 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. 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 English grammar, punctuation and spelling test framework (Section 7, pages 31–32).

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. 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. 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 which 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?

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 disabilities or SEND. They are based primarily on normal classroom practice and the available access arrangements are, in most cases, similar to those for other tests such as GCSEs and A levels.
c. STA publishes guidance on gov.uk about the range of access arrangements available to enable pupils with specific needs to take part in the KS1 tests. Access arrangements can be used to support pupils who have difficulty reading; who have difficulty writing; with a hearing impairment; with a visual impairment; who use sign language; who have difficulty concentrating; and who have processing difficulties.
d. The range of access arrangements available includes: early opening to modify test materials (for example, photocopying on to coloured paper); additional time; compensatory marks for pupils with a hearing impairment and unable to access the spelling test; scribes; transcripts; word processors or other technical or electronic aids; readers; prompters; rest breaks and written or oral translations.
e. Headteachers and teachers must consider whether any of their pupils will need access arrangements before they administer the tests.
f. Schools can contact the national curriculum assessments helpline or 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. The external marking agency carries out a range of checks to ensure that only markers who demonstrate acceptable marking accuracy and consistency mark NCTs.
b. 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)[3] for any one item is outside of 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.
c. The external marking agency undertakes 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.
d. 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.
e. 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 English GPS AMD bands and validity thresholds were set by the Marking Programme Leader and agreed with the Test Development Manager and the psychometrician. These were based on the complexity of the items and the number of marks available.

[3] 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.

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

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.

TDRs, psychometricians and senior staff discussed these methods at the STA test development subprogramme board. The STA Technical Advisory Group (comprising external experts in the field of test development and psychometrics) agreed that they were appropriate.

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 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. STA predicted the threshold of the expected standard at project board 3 to be 36. The final threshold for the KS2 GPS test was 36.

The evidence above confirms that TPT data accurately predicts 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 KS2 GPS is 80–120, and all scaled scores were represented in the 2019 test. 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 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 then used in a summed score likelihood IRT model to produce a summed score conversion table, which is then used to produce the raw score 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 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 the 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?

STA had a communication plan to inform stakeholders of the changes that were taking place prior to the introduction of the new national curriculum tests (and scaled scores) in 2016. This included speaking engagements with a range of stakeholders at various events and regular communications with schools and LAs 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?

STA monitors media coverage on a weekly basis, including coverage of NCTs and scores. During test week, STA monitors social media, 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: 978-1-78957-496-8 Product code: STA/20/8513/e

Follow us on Twitter:
@educationgovuk

Like us on Facebook:
facebook.com/educationgovuk