Content validation study:

2016 key stage 2 reading and mathematics tests

An investigation into the approach to domain sampling for the new suite of national curriculum tests

October 2017

Ofqual/17/6286/1

Authors

This report was written by Paul E. Newton and Benjamin M. P. Cuff, from Ofqual's Strategy, Risk and Research directorate.

Acknowledgements

The authors wish to thank Liz Twist, Helen Claydon, Stephen Goodman, Pam Kaur, and Barbara Donahue from the Standards and Testing Agency (STA) for their helpful contributions to this project.

For their generous assistance in identifying subject experts, we would like to thank Alison Borthwick, Hilary Povey, Sue Gifford, Debbie Morgan, Matt Lewis, Alice Onion, Barbara Conridge, Paul Clayton, John Hickman, Louise Beattie, Tracy Parvin, and Louise Johns-Shepherd.

For their advice on methodological issues, we would like to thank Jeffrey Goodwin, Melanie Ehren, Nick Wollaston, Zeek Sweiry, and Ayesha Ahmed.

Finally, we wish to thank all those who acted as subject matter experts for this research, and those who reviewed earlier versions of this report.

Contents

Tables and Figures

Glossary of terms

Individuals / Organisations

SMEs – Subject matter experts

STA – Standards and Testing Agency

Cognitive domain strands for reading

Acs_Tgt_Inf – Accessibility of target information

Cmpl_Tgt_Inf – Complexity of target information

Tsk-Sp_Cmpl – Task-specific complexity

Tec_Knw_Rq – Technical knowledge required

Rsp_Str – Response strategy

Cognitive domain strands for mathematics

Dpth_Und – Depth of understanding

Comp_Cmpl – Computational complexity

Spt_Rea/Dat_Int – Spatial reasoning / data interpretation

Rsp_Str – Response strategy

Executive summary

New key stage 2 reading and mathematics tests, based on a revised national primary curriculum, were introduced in May 2016. In this report, we present findings from research investigating these new testing arrangements. We focused specifically on the approach adopted by the Standards and Testing Agency (STA) to sampling national curriculum learning outcomes within the new tests.

Content validation studies, such as this one, can help us to understand the extent to which tests effectively sample from the full range of learning outcomes set out in the relevant curriculum. Within the limitations of what can reasonably be assessed, sampling should be relevant and representative, so that pupils can gain marks for demonstrating mastery across the full curriculum and so that assessment can support effective teaching and learning of the full range of learning outcomes specified by the curriculum. This is particularly important for 'high-stakes' tests, such as these, which are heavily relied upon in performance measures.

Our conclusions

Although we have identified potential areas for improvement and further research, our findings provide support for the robustness of STA's approach to domain sampling for the new suite of national curriculum tests. Their approach compares favourably with approaches adopted for similar tests, internationally. 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 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 is particularly important evidence in relation to the cognitive domain because the domain strands were innovative for both mathematics and reading, having been introduced specifically for the new tests. Finally, the degree of consistency between STA's item ratings and those of our independent experts supported the conclusion that STA's ratings were plausible and, by extension, that the 2016 tests sampled relevantly and representatively. Some important questions remain however. In particular, the representation of problem solving in the mathematics *Test Framework* and, consequently, in the mathematics test. This area, and others set out below, could benefit from further study.

Our investigation

To underpin their new testing arrangements, STA developed *Test Framework* documents – one for reading and one for mathematics – which specified their approach to sampling learning outcomes. Each document included a sampling

blueprint for topics of study from the national curriculum (known as the 'content domain' blueprint) and a sampling blueprint for levels and types of thinking skills necessary for responding to test questions (known as the 'cognitive domain' blueprint).

STA specified two sets of content domain 'strands' – one for reading and one for mathematics – using categories like 'algebra' and 'geometry' (for mathematics) or 'explain the meaning of words' and 'retrieve and record information' (for reading). These categories were less clearly demarcated within the national curriculum programmes of study for reading than for mathematics; mathematics categories were taken directly from the national curriculum.

While content blueprints have a long pedigree in educational testing, and the practice of specifying content blueprints is well-established, this is less true for the cognitive domain. Indeed, there is far less agreement amongst testing professionals concerning how best to classify the cognitive domain, let alone how best to sample it. In short, this is still an active area of research. With the removal of levels from the national curriculum, STA adopted an innovative approach to classifying the cognitive domain, developing their own strands for test development purposes. Four strands were developed for mathematics ('depth of understanding', 'computational complexity', 'spatial reasoning and data interpretation', and 'response strategy'), and five strands were developed for reading ('accessibility of target information', 'complexity of target information', 'task-specific complexity', 'technical knowledge required', and 'response strategy').

The main aim of our research was to provide an independent evaluation of STA's approach to sampling learning outcomes from the key stage 2 programmes of study for reading and mathematics within the new suite of tests. This was achieved via three studies, which considered the topics and thinking skills that appeared to be tested by the 2016 test questions, in terms of their relevance and representativeness. Relevance concerned the extent to which tested learning outcomes could be traced back to the *Test Framework* and national curriculum; while representativeness concerned the extent to which the balance/weightings of learning outcomes tested corresponded to the balance/weightings of those learning outcomes in the *Test Framework* and national curriculum. The topics and thinking skills tapped by any particular test question are not necessarily obvious. For this reason, we invited independent subject matter experts (SMEs), for mathematics and reading respectively, to share their views on the topics and thinking skills tapped by questions from the 2016 papers. These experts were asked to pass judgements at the sub-question level (e.g. a separate judgement for 1a and 1b, where questions were sub-divided) which is why we discuss the following results in terms of 'items' rather than 'questions'.

In Study 1, we compared the STA's content domain (topic) classifications and cognitive domain (thinking skills) ratings, for each test item, against guidelines given

in the *Test Framework* documents. Findings showed that the 2016 test for reading was representative with regards to the content domain, ie the sampling of content domain strands was within the guidelines prescribed by the *Test Framework*. This was also true for mathematics; at least, in relation to the high-level weighting specification given in the *Test Framework*.

The content domain weightings for mathematics were actually specified at a very high level of aggregation in the *Test Framework* document; that is, at the level of 'Number, ratio and algebra' (which included five separable strands) versus 'Measurement, geometry and statistics' (which included four). This left open the question of how each of the nine strands ought to be weighted, independently. Additionally, because the programme of study was specified by key stage 2 year, this left open the question of how topics from each of the four years ought to be weighted. There is no official answer to either of these questions. However, by considering the number of requirements specified in the national curriculum – literally the number of bullet points associated with each strand and year – we were able to infer very rough (unofficial) weightings. In relation to this analysis, it appeared that the 2016 mathematics test allocated a lower than 'expected' weighting (a) to topics of study from Year 4 and (b) to the measurement and geometry strands. The test also appeared to allocate a higher than 'expected' weighting (a) to topics of study from Year 6 and (b) to the calculation strand. Because there is no official specification of expected weightings at the level of strands and years, it would be wrong to conclude that the 2016 enacted weightings were inappropriate. Indeed, although not specified in the *Test Framework* for mathematics, STA's internal criteria for test construction do specify that the upper key stage 2 topics should be weighted more heavily than lower key stage 2 ones, reflecting the assumption that the majority of pupils will be working at the levels set out in the Y6 programme of study by the end of key stage 2. Having said that, these observations do raise the question of whether greater detail concerning expected weightings ought to be provided in future iterations of the *Test Framework*.

Study 1 also indicated that the 2016 reading test was representative with regards to the cognitive domain. The mathematics test contained fewer than intended marks for high computational complexity demands, one more than intended for medium computational complexity demands, slightly more marks than intended for medium spatial reasoning/data interpretation demands, and one more than intended for low spatial reasoning/data interpretation demands. Because of their size, and because it is far harder to classify the cognitive domain (as opposed to the content domain) definitively, we do not consider these differences to be problematic.

In Study 2, two panels of subject matter experts (SMEs) – one panel for reading and one for mathematics – were asked to rate the appropriateness of STA's content domain classification for each test item. For all but 1 of the 39 items on the reading test, SMEs concurred with STA's content classification. They also generally agreed that the content domain strands, as outlined in the *Test Framework*, were an

appropriate representation of the range of topics presented in the national curriculum. For mathematics, there was somewhat less concurrence with STA's content classifications. For 28% of the mathematics items, SMEs did not rate STA's classification as being the "most appropriate". However, when those disagreements were investigated in more detail, they tended almost always to reflect a difference of opinion at the sub-strand level. In other words, SMEs almost always agreed with STA over the strand tested by an item. As such, the somewhat lower levels of concurrence for mathematics were simply a consequence of the fact that the mathematics items were classified into much finer content categories than the reading items. Results for both subjects therefore provide support for the relevance and representativeness of the content domain sampling process.

In Study 3, the same panels of experts were asked to rate the demands of each item according to each of the cognitive domain strands that were developed by the STA. The consistency with which SMEs were able to apply these rating scales supported the plausibility of STA's strands, as it would have been difficult to achieve this consistency if those strands did not reflect genuine demands relevant to the assessment of key stage 2 learning outcomes. For both reading and mathematics, mean SME ratings of item demands closely matched the ratings made by STA. In addition, there was a relatively even spread of cognitive demands across items sampling each of the different content areas for both reading and mathematics. The findings support the relevance and representativeness of the cognitive domain sampling process. Although, for reading, SMEs' ratings were statistically significantly higher than STA's for the technical knowledge required strand, and statistically significantly lower than STA's for the response strategy strand, the sizes of these differences were small and we do not consider them to be problematic.

While these findings provide general support for the robustness of STA's approach to test development, a number of detailed insights are worthy of comment. For the cognitive domain, SMEs in the reading group noted an apparent overlap between certain domain strands (in particular, complexity of target information, technical knowledge required, and task-specific complexity demands), suggesting that these may not be mutually exclusive. Mathematics SMEs suggested that it might have been more appropriate to separate the spatial reasoning/data interpretation cognitive domain strand into two strands. SMEs in both groups also identified a number of other types of demands that did not feature in the *Test Framework* documents. For reading, these included time pressure and how engaging the reading texts were. For mathematics, these included demands on working memory, language demands, and the degree of contextual knowledge required. None of these observations presents a major challenge to STA's cognitive domain modelling. Moreover, it is important to recognise that STA's models are primarily tools for test development, and therefore need to balance completeness against manageability. However, these observations do underline the importance of treating the cognitive domain as 'work in progress' and of continuing to research and develop it for future test models.

Finally, SMEs in the reading group identified an issue with the ordering of demands in the test papers. Specifically, it seemed as though a number of particularly high-demand items were located mid-way through the paper. This issue is explored in greater detail in a separate review of evidence on the accessibility of the 2016 reading test.

Although we concluded that the STA's approach to test development was robust, we also recognised certain caveats, and we identified certain questions that still remain. First, as acknowledged in the *Test Framework* documents, certain elements of the national curriculum cannot straightforwardly be tested (eg certain aspects of discussion for reading, certain aspects of measuring for mathematics) and will, necessarily, not be sampled within the national curriculum tests. This is not a criticism of the test development process. However, this will result in a certain amount of content underrepresentation, and this does present a risk to validity. This is also a potential risk to teaching and learning, as it may increase the likelihood that those untested curriculum requirements are not taught as effectively, or thoroughly, as the tested ones. These risks were highest for reading, as a substantial number of curriculum requirements could not straightforwardly be tested. Similar risks also arise when certain content strands are perceived to have an unduly low weighting; this perception seemed quite possible for reading, as the intended weighting range for certain strands included zero (eg 0 to 6%) and, for one strand, actually turned out to be zero in the 2016 test.

Second, as the national curriculum documents for English and mathematics do not specify weightings for the content and cognitive domains, it is unclear the extent to which the weightings prescribed by the *Test Frameworks*, and therefore those enacted within the 2016 tests, appropriately reflect the intentions of the curriculum designers. This highlights the role of each *Test Framework* in constructing a model of pupil proficiency in each subject area, via decisions on weightings and suchlike. Since it is possible for quite different models of pupil proficiency to be constructed from the same national curriculum requirements, this raises questions concerning the extent to which and the ways in which curriculum designers ought to contribute to the construction of the corresponding *Test Framework*. Indeed, it raises the parallel question of the extent to which assessment designers ought to be involved in the construction of the corresponding programme of study. In response to questions like these, STA explained that the *Test Frameworks* were shared with members of the national curriculum review team during 2012/2013. It is worth considering whether it would be useful to develop a more formal model of collaboration for future revisions of the national curriculum and *Test Frameworks*.

Third, some SMEs suggested that the separation of the content domain into individual strands seemed somewhat artificial in relation to real-world applications of mathematics, and that the three main aims of the national curriculum for mathematics – fluency, reasoning, and problem solving – may not be represented adequately by the *Test Framework*. This echoed concerns that have recently been

raised by various teaching and mathematics associations. Unfortunately, our investigation was not designed to address this issue directly; so this important question remains open.

1 Introduction

This report presents outcomes from a research project which was designed to investigate an aspect of the validity of national curriculum assessment arrangements, focusing particularly on the test design process for the new suite of national curriculum tests delivered for the first time in 2016. This section introduces Ofqual's approach to regulating national assessment arrangements, the new suite of national curriculum tests, and our research methodology, generally known as 'content validation'. Subsequent sections describe methods, results, and conclusions of 3 content validation studies.

1.1 Ofqual's approach to regulating national assessments

Ofqual has two statutory objectives in relation to national assessments, as specified by the Apprenticeships, Skills, Children and Learning Act 2009:[1]

> The assessments standards objective is to promote the development and implementation of regulated assessment arrangements which
>
> a) give a reliable indication of achievement, and
>
> b) indicate a consistent level of attainment (including over time) between comparable assessments.
>
> The public confidence objective is to promote public confidence in regulated qualifications and regulated assessment arrangements.

In other words, we regulate to promote the validity of national assessments and to promote confidence in assessment outcomes. By putting validity at the heart of what we do, we emphasise that assessment arrangements must ensure that statutory assessments (including national curriculum tests and teacher assessments) measure what they need to measure, to ensure that pupils' results are accurate and useful.

The present study focuses on the validity of new testing arrangements for key stage 2. More specifically, it focuses on the approach adopted by the Standards and Testing Agency (STA) to developing each year's test papers; three papers for mathematics and one for reading. Although this study provides only one source of evidence concerning the validity of the new testing arrangements for key stage 2, this kind of investigation, which is known as a 'content validation' study, is considered to be one of the most fundamental validation techniques for educational

[1] http://www.legislation.gov.uk/ukpga/2009/22/section/128

attainment tests. The present study scrutinises the 2016 tests for reading and mathematics, to provide insights into the effectiveness of STA's test development procedure. STA produce a range of tests, including biennial key stage 2 science sampling tests, key stage 1 tests and the phonics screening check. We focused this study on the key stage 2 reading and mathematics tests because they have a very high profile, being heavily relied upon in performance measures and having particularly 'high-stakes' for schools.

1.2 The new suite of national curriculum tests

With the introduction of the new national curriculum in September 2014, national assessment arrangements were also revised. National curriculum levels were removed, new teacher assessment frameworks were developed, and the national curriculum tests were redesigned. This included implementing a government commitment to "tough new standards" for literacy and numeracy in primary schools.[2] Other changes to the suite of national curriculum tests for key stage 2 included:

- no separate test for the highest achieving pupils (previously the Level 6 test);
- a small number of questions to stretch the highest achieving pupils;
- replacing the mental mathematics test with an arithmetic test that would focus on assessing fundamentals/fluency; and
- using scaled scores, and an expected standard threshold, to report pupil results.

In March 2014, STA published the first draft of *Test Framework* documents for the new reading and mathematics tests (for the most recent drafts, see STA, 2016a, 2016b).[3] These documents were designed to translate the new national curriculum into a test specification and were written principally for those involved in the test development process.

Each *Test Framework* document is like a blueprint that specifies how learning outcomes are to be sampled by the questions in each test. An essential warrant underpinning any validity claim is that tests are designed to measure the right 'thing' in the right 'way'. The right 'thing' in this instance is the range of learning outcomes in the national curriculum, for key stage 2 reading and mathematics, respectively. The right 'way' involves relevant and representative sampling, by giving due weight to the

[2] https://s3-eu-west-1.amazonaws.com/manifesto2015/ConservativeManifesto2015.pdf

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

various elements of each programme of study when constructing national curriculum tests.

In relation to the present study, it is important to note that the *Test Frameworks* prescribe a range of different question formats, including:

- selected-response questions (eg selecting the correct answer from four alternative options, for either mathematics or reading);
- short-answer constructed-response questions (eg recording the answer to a simple calculation question for mathematics, or writing a sentence to explain how a certain piece of information is known for reading); and
- extended constructed-response questions (eg writing a couple of sentences to explain a problem solution for mathematics, or writing a short paragraph to explain the appeal of a particular character for reading).

The 2016 test papers, as well as sample materials, are available for download from the STA website.[4]

1.2.1 Sampling the national curriculum

The *National Curriculum in England Framework Document* (DfE, 2014) specifies the statutory national curriculum, which sets out in programmes of study, on the basis of key stages, subject content for those subjects that should be taught to all pupils in state-funded schools.[5] End-of-key-stage national curriculum tests are designed to measure the degree to which pupils have mastered those programmes of study.

Within the national curriculum framework document, **subject content** tends to be specified in terms of what pupils need to be taught, for example: 'pupils should be taught to ask and answer questions about totalling and comparing categorical data'. These content requirements typically concern skills that can be applied across a range of problems or situations; which, in the context of testing, means in response to different kinds of task or question. As such, the tasks or questions that appear in a test may also differ in terms of the **cognitive demands** which they pose to pupils: both in terms of the *nature* of those cognitive demands; and in terms of the *level* of cognitive demand posed. For instance, certain totalling tasks or questions may pose different cognitive demands from other totalling tasks, or they may pose essentially the same cognitive demands but at different levels, or perhaps both. This helps to explain why a pupil who is able to solve straightforward totalling problems may not

[4] https://www.gov.uk/government/collections/national-curriculum-assessments-practice-materials

[5] https://www.gov.uk/government/collections/national-curriculum

yet be able to solve complex totalling problems. STA describes these cognitive demands as the "thinking skills and intellectual processes required for the [tests]" (STA, 2015b, p. 29).

Following this logic, national curriculum requirements can be specified in terms of two dimensions:

- the elements of subject content that are recorded within the programme of study for a particular key stage; and
- the nature and level of cognitive demands posed by the tasks or questions through which mastery of subject content elements is to be demonstrated.

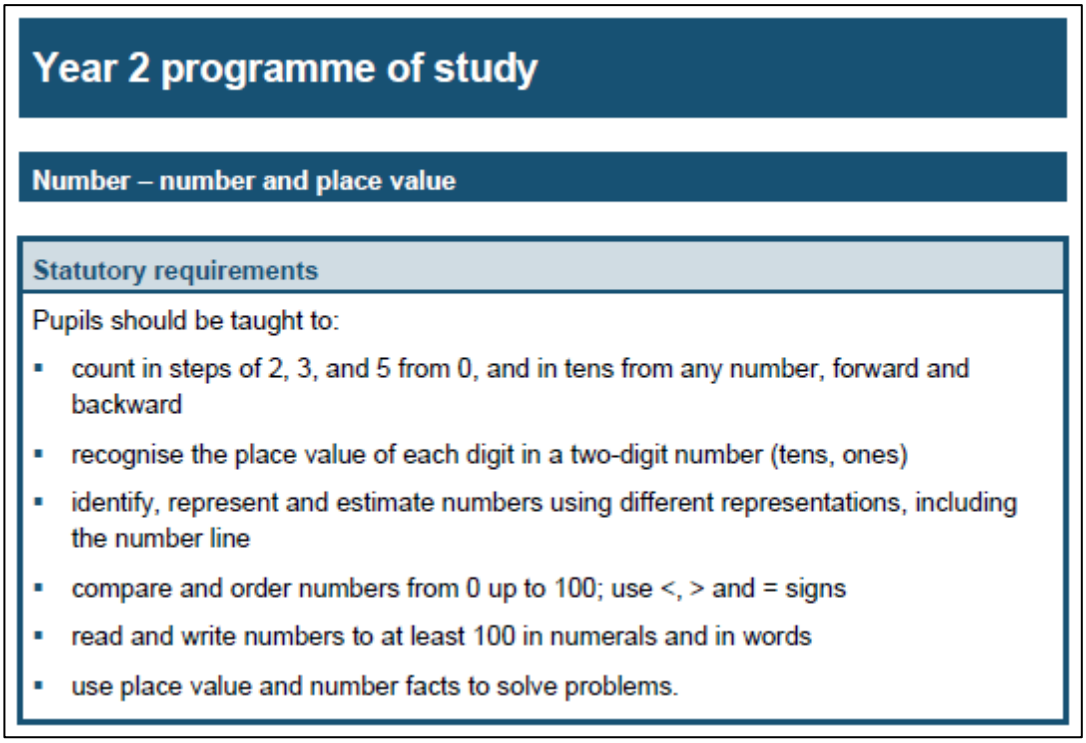
STA refers to these two dimensions as the **content domain** and the **cognitive domain** respectively.

It is important to note, at this point, the distinction between levels of *demand* and levels of *difficulty*. Difficulty can be defined as the quantification of how easy or hard pupils found the test. Demands are part of the explanation for why it proved to be easy or hard. STA uses its cognitive domain strands (demands) to help determine how difficult a question may turn out to be. The predictive power of this approach relies upon the identification of all the different types of demands that are important to demonstrating mastery of the subject; these are the intended demands associated with each question. However, it is important to remember that questions are sometimes difficult because of unintended demands. For example, pupils may find a test item difficult due a poor choice of wording, or a distracting graphic, even when the item is intended to test only low-level cognitive skills. The distinction between these concepts should be borne in mind throughout reading this report.

National curriculum tests are intended to 'sample' from the national curriculum, by presenting a mixture of questions, each one designed to tap into a particular aspect, or combination, of knowledge, skills and/or understanding. To justify the claim that national curriculum testing procedures have validity requires (amongst other things) evidence that the tests sample the national curriculum fairly (that is, relevantly and representatively); not simply in terms of the content domain, but also in terms of the cognitive domain. It would be problematic if, for instance, the tests were representative in terms of the content domain, but sampled only low-level cognitive demands; or, if they sampled too many low-level cognitive demands for a particular content sub-domain, whilst sampling too many high-level cognitive demands for another.

Making sense of the requirement to sample the national curriculum content domain representatively is not entirely straightforward. The task is made easier because subject content elements are listed within the national curriculum framework document clearly and concisely under content sub-domain headings for each year

group. Figure 1 illustrates how these elements are listed for the sub-domain 'Number – number and place value' which is one of a set of sub-domains (also known as 'content strands') identified for key stage 2 mathematics.



Year 2 programme of study

Number – number and place value

Statutory requirements

Pupils should be taught to:

- count in steps of 2, 3, and 5 from 0, and in tens from any number, forward and backward
- recognise the place value of each digit in a two-digit number (tens, ones)
- identify, represent and estimate numbers using different representations, including the number line
- compare and order numbers from 0 up to 100; use <, > and = signs
- read and write numbers to at least 100 in numerals and in words
- use place value and number facts to solve problems.

Figure 1. Extract from the national curriculum framework (DfE, 2014, p. 116)

Although these elements are listed clearly and concisely, interpretation and judgement is still required in order to translate the curriculum framework into national curriculum tests, via the *Test Framework* documents. For instance, the curriculum provides no explicit indication of whether any of its statutory requirements is in any way more or less 'weighty' than any other. In addition, assessment tasks often draw upon a number of content domains simultaneously, even when they are primarily focused upon a single one. Indeed, it is sometimes impossible to classify an assessment task unambiguously in terms of the content element that it is primarily intended to test.

Making sense of the requirement to sample the national curriculum cognitive domain representatively is far more complicated. References to cognitive demands in the national curriculum are far more oblique than references to content requirements. More fundamentally, although the 'constructs' into which national curriculum content domains are de-constructed, are well-established, and have widespread currency – such as 'ratio' vs. 'algebra' for mathematics and 'summarise main ideas' vs. 'make inferences from the text' for reading – this is not true for the cognitive domains.

Despite the fact that attempts to analyse educational objectives by cognitive domain can be traced back a long way (eg Bloom, Englehart, Furst, Hill, & Krathwohl, 1956), the science of cognitive domain analysis remains under-developed; perhaps even more so when considering subject-specific cognitive demands.

For these reasons, few large-scale testing programmes produce blueprints which address cognitive domain sampling at the level of detail presented within STA's *Test Framework* documents.[6] STA's cognitive domain sampling approach is unique; having been developed specifically for the suite of tests that were administered for the first time in summer 2016. An example of the first of the four 'cognitive strands' for mathematics is presented in Figure 2.

Table 5: Depth of understanding

Strand	Rating scale			
	(low) 1	2	3	4 (high)
Depth of understanding	recall of facts or application of procedures	use facts and procedures to solve simple problems	use facts and procedures to solve more complex problems	understand and use facts and procedures creatively to solve complex or unfamiliar problems

Figure 2. Extract from the mathematics *Test Framework* (STA, 2015b, p. 29)

1.2.2 STA's content and cognitive domain strands

As noted before, STA's content domain strands are explicit within the national curriculum, whereas its cognitive domain strands were developed specifically for the purpose of test development.

Content strands

The national curriculum for **mathematics** specifies nine content strands (the letters in parentheses are the abbreviations used by STA, eg C = calculation):

[6] Recent work by PARComp_Cmpl, ETS, and Pearson on the *Development of Cognitive Complexity Measures for PARComp_Cmpl* provides an illustration of similarly innovative research and development (e.g. presentation to the Cognition and Assessment SIG Business Meeting at the annual meeting of the American Educational Research Association, Philadelphia, April 5, 2014, www.aera.net/LinkClick.aspx?fileticket=v1rGlSYcCgk%3D&portalid=38).

1. Number – number and place value (N)
2. Number – addition & subtraction, multiplication & division (C)
3. Number – fractions (F)
4. Ratio and proportion (R)
5. Algebra (A)
6. Measurement (M)
7. Geometry – properties of shapes (G)
8. Geometry – position and direction (P)
9. Statistics (S).

A programme of study is specified for each of the four key stage 2 years; and statutory teaching requirements are specified as discrete bullet points, under each of the strands (as illustrated in Figure 1). A total of 173 statutory requirements (bullet points) are specified for key stage 2 mathematics, summarised by year and content strand in the second and third columns of Table 1.

STA's *Test Framework* document for mathematics transforms these strands and requirements (bullet points) into a table, with the nine strands in rows and the four years in columns. The transformation is very faithful to the presentation in the national curriculum, albeit with a small number of changes, including:

- one example of two national curriculum requirements being collapsed into a single *Test Framework* requirement
- twelve examples of one national curriculum requirement being divided into at least two *Test Framework* requirements (half of these divisions being for Measurement).

In effect, the *Test Framework* represents the 173 national curriculum requirements as 195 requirements, summarised by year and content strand in the fourth and fifth columns of Table 1.

The *Test Framework* document identifies a small number of statutory teaching requirements that are difficult to assess fully in a paper-based format. These include mental mathematics requirements and requirements involving practical equipment (STA, 2015b, p. 28).

Table 1. Breakdown of statutory teaching requirements by year and strand

	National Curriculum		Test Framework	
	Number of bullet points	% of total	Number of bullet points	% of total
Year				
Year 3	33	19.1	44	22.6
Year 4	42	24.3	45	23.1
Year 5	49	28.3	53	27.2
Year 6	49	28.3	53	27.2
Strand				
Number	25	14.5	27	13.8
Calculation	39	22.5	39	20.0
Fractions	40	23.1	40	20.5
Ratio	4	2.3	4	2.1
Algebra	5	2.9	5	2.6
Measurement	27	15.6	44	22.6
Geometry	19	11.0	22	11.3
Position	6	3.5	6	3.1
Statistics	8	4.6	8	4.1

The national curriculum for **reading** is specified differently from mathematics. The programmes of study for years 3/4 and 5/6, respectively, begin by distinguishing 'word reading' from 'comprehension' and by noting that, while word reading cannot be overlooked during key stage 2, the focus should be upon comprehension. The statutory requirements for comprehension, as for mathematics, are specified as bullet points, although these are not sub-divided by strand. In effect, the major bullet points within the national curriculum framework document constitute eight strands (which are summarised below):

1. maintain positive attitudes to reading
2. understand what they read
3. retrieve, record and present information
4. participate in discussions about books
5. discuss and evaluate how authors use language
6. distinguish between statements of fact and opinion
7. explain and discuss their understanding of what they have read
8. provide reasoned justifications for their views.

Clearly, some of these strands would not be amenable to assessment via a conventional reading comprehension test (eg 'maintain positive attitudes to reading'). In fact, the strands presented in the *Test Framework* for reading (comprehension) focus heavily upon the national curriculum 'understanding' strand ('understand what they read'). 5 of the 8 *Test Framework* strands, presented below, correspond to minor bullet points within the national curriculum 'understanding' strand (ie 1 & 3-6):

1. give / explain the meaning of words in context
2. retrieve and record information / identify key details from fiction and non-fiction
3. summarise main ideas from more than one paragraph
4. make inferences from the text / explain and justify inferences with evidence from the text
5. predict what might happen from details stated and implied
6. identify / explain how information / narrative content is related and contributes to meaning as a whole
7. identify / explain how meaning is enhanced through choice of words and phrases
8. make comparisons within the text.

Whereas the *Test Framework* for mathematics is essentially a re-presentation of the national curriculum framework, the *Test Framework* for reading is more of a re-interpretation. Indeed, the eight strands also share some continuity with the seven

assessment focuses previously used by STA to represent the content domain for earlier national curriculum reading tests, to support test development and marking.[7]

Cognitive strands

In recent years, it has become more common for large-scale educational assessments to be developed on the basis of sampling blueprints which specify both content and cognitive domain strands, although few large-scale testing programmes produce blueprints which address cognitive domain sampling at the level of detail presented within STA's *Test Framework* documents. Internationally, one of the most common bases for defining the cognitive domain is Bloom's taxonomy of educational objectives for the cognitive domain (Bloom et al., 1956). Bloom's taxonomy proposes a hierarchy of thinking skills; from knowledge, at the bottom, through comprehension, application, analysis, synthesis, to evaluation, at the top. A blueprint might be constructed by creating a table which crosses content domain strands with the six cognitive domain strands, to ensure that questions are not all targeted at particular thinking skill levels (eg only knowledge and comprehension) but are targeted appropriately across all of the six thinking skill levels.

In the UK, it has become traditional for examination syllabuses/specifications to include a small number of assessment objectives, which typically specify a certain combination (ie percentage weighting) of:

- knowledge and understanding;
- practical skills; and
- application of knowledge, understanding and practical skills.

Although these objectives identify somewhat different kinds of cognitive demands, thereby resonating with Bloom's taxonomy, they are not hierarchical. In his work on sources of difficulty in assessment tasks, Alastair Pollitt also rejected the simplistic idea of a hierarchy of cognitive skills, attempting instead to identify the particular cognitive processes involved in responding to test and exam questions. With collaborators, he developed the CRAS rating scales (see Pollitt, Ahmed, & Crisp, 2007). Distilled from evidence in many subject areas, the CRAS scales identify four distinct strands, the last two of which are sometimes subdivided:

[7] Assessment focuses for English reading: AF1 Use a range of strategies, including accurate decoding of text, to read for meaning; AF2 Understand, describe, select or retrieve information, events or ideas from texts and use quotation and reference to text; AF3 Deduce, infer or interpret information, events or ideas from texts; AF4 Identify and comment on the structure and organisation of texts, including grammatical and presentational features at text level; AF5 Explain and comment on writers' uses of language, including grammatical and literary features at word and sentence level; AF6 Identify and comment on writers' purposes and viewpoints, and the overall effect of the text on the reader; AF7 Relate texts to their social, cultural and historical contexts and literary traditions.

1. **Complexity** – the number of elements that need to be kept in mind while answering, and that need to be related to each other.
2. **Resources** – the extent to which candidates are given all and only the information they need to complete a task, or are required either to supply it themselves or extract it from a source that also contains irrelevant information.
3. **Abstractness** – the extent to which a task deals with ideas rather than concrete objects or phenomena.
4. **Strategy** – the extent to which candidates are required to devise their own strategies for completing the task.

These strands identify generic cognitive demands, with relevance across a range of content domains. Figure 3 illustrates how the scales can be presented, and used to identify a low-to-high rating, for each of the strands, for each task/question.

	1	2	3	4	5
Complexity The number of components or operations or ideas and the links between them		Mostly single ideas and simple steps. Little comprehension e, except that required for natural language. Few links between operations.		Synthesis or evaluation is required. Need for technical comprehension. Makes links between cognitive operations.	
Resources The use of data and information		More or less all and only the data/information needed is given.		Student must generate or select the necessary data/information.	
Abstractness The extent to which the student deals with ideas rather than concrete objects or phenomena.		Mostly deals with concrete objects.		Mostly abstract.	
Task strategy The extent to which the student devises (or selects) and maintains a strategy for tackling the question.		Strategy is given. Little need to monitor strategy. Little selection of information required.		Students need to devise their own strategy. Students must monitor the application of their strategy.	
Response strategy The extent to which students have to organise their own response.		Organisation of response hardly required.		Must select answer content from a large pool of possibilities. Must organise how to communicate response.	

Figure 3. The CRAS Scales (extract from Pollitt et al., 2007, p. 186)

STA's cognitive domain strands were modelled on the CRAS scales, tailored to the particular demands of key stage 2 mathematics tests and reading tests, respectively.

STA identified four cognitive domain strands for mathematics:

1. depth of understanding (Dpth_Und)
2. computational complexity (Comp_Cmpl)
3. spatial reasoning and data interpretation (Spt_Rea/Dat_Int)
4. response strategy (Rsp_Str).

STA identified five cognitive domain strands for reading:

1. accessibility of the target information (Acs_Tgt_Inf)
2. complexity of the target information (Cmpl_Tgt_Inf)
3. task-specific complexity (Tsk-Sp_Cmpl)
4. response strategy (Rsp_Str)
5. technical knowledge required (Tec_Knw_Rq).

Appendices A and B explain what STA means by each of these cognitive domain strands, for mathematics and reading, respectively. This information comes primarily from the *Test Framework* documents, which contain the only published information on the cognitive domain strands. For the current project, additional insights were also gained from conversations with STA officials (past and present) responsible for test development. Although there are clear overlaps between STA's cognitive domain strands and the CRAS scales, such as the response strategy strand, there are also significant differences, which reflect their subject-specificity. STA's strands were derived from first principles, from internal research and from the wider literature; for example, recent work by Lumley, Routitsky, Mendelovits, and Ramalingam (2012) was particularly influential in crafting the strands for reading.

1.3 Content validation

Content validation is the generic name given to studies that investigate the relevance and representativeness of domain sampling during test development.[8] It involves independent experts judging test questions in terms of what they perceive them to be measuring. If judgements made by independent experts agree with those made

[8] Although, in recent years, the term 'alignment study' has become more popular.

(during the course of test development) by test development teams, then this provides important evidence in support of the overarching validity claim.

1.3.1 Approaches to content validation

Content validation is undertaken to provide evidence that a particular testing procedure has been effectively designed to ensure that each live test samples representatively from its target domain (or domains). It typically involves panels of subject matter experts (SMEs) scrutinising the questions that comprise a particular test and classifying each one in terms of the content area that they consider it to be testing. SMEs may also be invited to classify the same questions in terms of cognitive demands; although this is somewhat less common.[9]

Sometimes, content validation is undertaken during the test development process, to provide a semi-independent evaluation of judgements made by question writing teams. Indeed, under the current test development model, content ratings and cognitive domain references are reviewed by a test review group and by curriculum advisors, as part of the test development process.

Occasionally, particularly for very high profile tests, content validation is undertaken subsequent to live administration, independently of the test development process and of test development personnel. This is in order to provide a more public, arms-length evaluation. A good example of this kind of content validation is the study by Sireci, Robin, Meara, Rogers, and Swaminathan (2000), which provided an external evaluation of the 1996 grade 8 science tests for the North American National Assessment of Educational Progress survey.

A study that was built into the test development process would be likely to focus very pragmatically upon the degree to which those responsible for writing questions and compiling tests had succeeded in the task of developing the test *as that task was interpreted and operationalised by the test development agency*. A study of this kind would therefore be unlikely to question any of the decisions upon which the *Test Framework* was designed, eg the:

- content domain sampling model (the presentation and weighting of elements from the national curriculum framework);
- definition of the constructs underpinning the cognitive domain strands; and
- criteria for applying rating scale points within each cognitive domain strand.

[9] Even when extended to the cognitive domain, it tends to still be known as ‘content validation’.

On the other hand, a study that was not built into the test development process would be free to explore a wider range of assumptions and judgement calls, and to investigate more comprehensively and independently the degree to which the tests sampled representatively from the national curriculum. We decided to frame the rationale for the present content validation study as a more comprehensive and independent investigation than might be conducted within the development process for a particular test, inviting scrutiny of a wider range of assumptions and judgement calls.

We set clear parameters for the study, to enable its outcomes to be as useful as possible. Our ultimate point of reference was the national curriculum framework document (DfE, 2014) and its specification of key stage 2 mathematics and English. It should be noted that Ofqual has a remit to question statutory assessment arrangements (in particular, the validity of procedures for assessing the national curriculum), but has no remit to question the national curriculum per se. Second, as noted in the *Test Framework* documents, we accepted that there are elements of the national curriculum that are not readily amenable to testing. Similarly, of course, we acknowledged that only statutory elements of the national curriculum are tested. We recognised that we would need to bear such considerations in mind during our investigation.

1.3.2 Content validation in the UK

Although content validation is often described as one of the most fundamental validation techniques for educational attainment tests, there is no tradition of content validation research in the UK. It is not easy to locate studies of this sort within the 'local' literature on large-scale educational assessments, albeit with a few notable exceptions in relation to England (eg Clesham, 2013; Greatorex, Shaw, Hodson, & Ireland, 2013) and the Republic of Ireland (eg Cullinane & Liston, 2016).

Despite the relative absence of formal content validation studies, recent decades have seen the growth of a tradition of research in the UK into the features that make test and exam questions difficult; which overlaps with content validation in important ways, particularly in relation to sampling the cognitive domain and the idea of cognitive demands. Some of this work has oriented towards being able to predict the difficulty of test and exam questions (eg Crisp & Grayson, 2013; El Masri, Ferrara, Foltz, & Baird, 2017). Indeed, some of this work overlaps directly with the tradition of research into comparability of assessment standards (eg Pollitt et al., 2007). Other work in this tradition has oriented towards being able to understand the demands made by features of test and exam questions, to help question writers to write better questions (eg Ahmed & Pollitt, 2007; Crisp & Grayson, 2013; Greatorex, 2013; Pollitt, Ahmed, Baird, Tognolini, & Davidson, 2008; Pollitt, Entwistle, Hutchinson, & de Luca, 1985; Pollitt, Hughes, Ahmed, Fisher-Hoch, & Bramley, 1998; Sweiry, 2006; Vappula & Clausen-May, 2006).

1.3.3 The current research

Our content validation project included a desk-based analysis (Study 1) of evidence provided by STA, followed by an empirical investigation involving a content-strand rating exercise (Study 2) and a cognitive-strand rating exercise (Study 3). The overarching purpose of these studies was to investigate STA's approach to developing the test papers, focusing specifically upon its approach to sampling learning outcomes from the key stage 2 programmes of study, from Year 3 to Year 6, for reading and mathematics. Should the tasks which comprise the tests not align appropriately or adequately with the *Test Frameworks*, then this would raise concerns over the validity of the recently reformed national assessment arrangements. As teaching is often driven by assessments, a knock-on effect could be that the national curriculum would not be appropriately taught. It is therefore important to gather validity evidence to assess the likelihood of these risks.

Two general research questions provided a broad structure for the studies:

- Do items[10] in the key stage 2 tests embody content and cognitive demands that are consistent with the national curriculum, to elicit an appropriate, ie **relevant**, body of evidence of attainment?
- Do the key stage 2 tests effectively sample those content and cognitive demands, to elicit an adequate, ie **representative**, body of evidence of attainment?

In other words, we decided to investigate whether items from each of the tests corresponded to elements of knowledge, skill and understanding that derived from the new national curriculum, and whether the tests reflected a proper balance of those elements. This involved a two-pronged approach, relying upon:

- evidence provided by STA (Study 1), and
- evidence from two rating exercises (Studies 2 and 3).

This stimulated a series of more specific research questions:

1. To what extent did the content domain strand balance in the 2016 tests reflect the intended content domain strand balance, given STA's classification of each item in terms of the content domain and the *Test Framework* blueprints? [Study 1 – representativeness]

[10] From hereon, the report will tend to refer to 'items' rather than 'questions' because many national curriculum test questions comprise subcomponents which test distinct elements of knowledge, skill and understanding. Each subcomponent is referred to as an item.

2. To what extent might independent experts agree with each other, and with STA, over the classification of each item in terms of the content domain? [Study 2 – relevance and representativeness]

3. To what extent did the cognitive domain strand balance in the 2016 tests reflect the intended cognitive domain strand balance, given STA's classification of each item in terms of the cognitive domain and the *Test Framework* blueprints? [Study 1 – representativeness]

4. To what extent might independent experts agree with each other, and with STA, over the classification of each item in terms of the cognitive domain? [Study 3 – relevance and representativeness]

An additional, higher level question was also posed:

5. To what extent might independent experts believe that STA's approach to sampling learning outcomes from the key stage 2 programmes of study, as set out in the *Test Framework* documents, is appropriate and adequate? [Study 2 and Study 3 – relevance and representativeness]

Although the current research was intended to serve as a public, arms-length evaluation, its purpose was not exclusively summative. In other words, it was not intended simply to provide a definitive, *post hoc*, yes or no judgement. First, the method is not capable of delivering absolutely definitive answers. Content validation is inherently judgemental, based on the opinions of independent subject matter experts (SMEs). If those SMEs are sufficiently qualified to be considered credible judges, then their judgements should be taken seriously. However, their judgements are still opinions, and these sit alongside other opinions, including those documented during the test development process.

Second, the study had considerable potential to fulfil a formative purpose, by providing information that may be used by STA during future test development iterations. As noted above, we are ultimately concerned with the validity of national curriculum testing arrangements; and testing procedures can be improved, from year to year, on the basis of feedback from evaluations such as those described in the present report. As the STA's approach to cognitive domain sampling is quite innovative, it seemed likely that this might benefit from the kind of independent evaluation that Ofqual could undertake.

2 Study 1 – Test representativeness of the content and cognitive domains

This study made use of information provided by STA, in the form of their own content / cognitive domain classifications of the items in the 2016 mathematics and reading tests, as well as information provided within the *Test Framework* and national curriculum documents. The subsections to follow present the findings of this study.

2.1 Content domains

2.1.1 Reading test

Although the *Test Framework* for reading specifies weightings for the content domain strands for test development purposes, the national curriculum does not, nor is it possible to infer a weighting from the degree to which each strand is elaborated within the curriculum. Therefore, although we are able assess the representativeness of the test in relation to the *Test Framework*, we are far less able to assess whether the framework, and therefore the test itself, appropriately reflects the intentions of the curriculum designers.

Table 2 shows the weightings prescribed by the *Test Framework* document, for each of the content domain strands, alongside the weightings enacted by STA through the items chosen for the 2016 test. A comparison of these figures shows that the sampling of the content domain strands by the 2016 test – according to the STA classification of the content tested by each item – was consistent with the guidelines prescribed by the *Test Framework*, ie, the test was representative within the design parameters laid down by STA. It is worth noting that this was despite the fact that no question was targeted primarily at strand 2h (the *Test Framework* allows for a 0-6% weighting for some content strands).

Table 2. Intended and enacted weightings of reading content domains

Content domain	Intended weightings (*Test Framework*)		Enacted weightings (2016 test)	
	Number of marks	Percentage of total mark	Number of marks	Percentage of total mark
2a - give / explain the meaning of words in context	5-10	10-20%	10	20%
2b - retrieve and record information / identify key details from fiction and non-fiction	8-25	16-50%	15	30%
2c - summarise main ideas from more than one paragraph	1-6	2-12%	1	2%
2d - make inferences from the text / explain and justify inferences with evidence from the text	8-25	16-50%	18	36%
2e - predict what might happen from details stated and implied	0-3	0-6%	3	6%
2f - identify / explain how information / narrative content is related and contributes to meaning as a whole	0-3	0-6%	1	2%
2g - identify / explain how meaning is enhanced through choice of words and phrases	0-3	0-6%	2	4%
2h - make comparisons within the text	0-3	0-6%	0	0%

2.1.2 Mathematics test

For the mathematics test, content domain strands are combined into 2 'content areas' within the *Test Framework* for the purposes of weightings. Table 3 shows that the weightings enacted in the 2016 tests were consistent with those prescribed by the *Test Framework* (ie the mathematics tests were also representative within the design parameters laid down by STA). Table 4 shows the same when broken down into individual test papers.

Table 3. Intended and enacted weightings of mathematics content domains (whole test)

	Intended weighting (*Test Framework*)		Enacted weighting (2016 tests)	
Content area and included strands	Number of marks	Percentage of total mark	Number of marks	Percentage of total mark
Number, ratio and algebra • Number, place value (N) • Addition, subtraction, multiplication, division, calculations (C) • Fractions, decimals and percentages (F) • Ratio and proportion (R) • Algebra (A)	83-93	75-85%	86	78%
Measurement, geometry and statistics • Measurement (M) • Geometry – properties of shapes (G) • Geometry – position and direction (P) • Statistics (S)	17-27	15-25%	24	22%

Table 4. Intended and enacted weightings of mathematics content domains (by paper)

Content area[a]	Intended weighting (*Test Framework*)		Enacted weighting (2016 tests)	
	Number of marks	Percentage of total mark	Number of marks	Percentage of total mark
Paper 1				
Number, ratio and algebra	40	100%	40	100%
Measurement, geometry and statistics	0	0%	0	0%
Paper 2				
Number, ratio and algebra	22-26	63-74%	23	66%
Measurement, geometry and statistics	9-13	26-37%	12	34%
Paper 3				
Number, ratio and algebra	22-26	63-74%	23	66%
Measurement, geometry and statistics	9-13	26-37%	12	34%

[a] Including the same strands as in Table 3

Similar to the documents for reading, although the *Test Framework* for mathematics specifies weightings for the content domain, the national curriculum does not. However, with respect to mathematics, rough weightings can, to a degree, be inferred from the national curriculum by considering the extent to which each strand is elaborated. For example, one can look at the number of bullet points provided for each content domain strand. By looking at the same for the *Test Framework*, and counting the number of items from each strand presented in the test, weightings given within the national curriculum can be compared with the *Test Framework*, and the test itself. The first 3 columns of Table 5 are taken from Table 1. The fourth column of Table 5 shows the enacted weightings of the 2016 test.

One might infer from Table 5 that the weightings given for the different years of study are comparable between the national curriculum and the *Test Framework*. However,

in the test, a lesser than 'expected' weighting is given to topics of study from year 4, and a greater than 'expected' weighting is given to topics of study from year 6, in comparison with the other 2 documents. Additionally, in the test, a greater than 'expected' weighting is given to the calculation strand, and a lesser than 'expected' weighting is given to measurement and geometry, compared to the other 2 documents. These findings are discussed in section 5.1.1.

Table 5. Curriculum, *Test Framework* and Test (enacted) weightings

	National Curriculum	*Test Framework*	2016 Test	
	% of bullet points	% of bullet points	% of items	
Year				
Year 3	19.1	22.6	16.4	(20,23,6)
Year 4	24.3	23.1	**11.8**	(13,6,17)
Year 5	28.3	27.2	28.2	(30,26,29)
Year 6	28.3	27.2	**43.6**	(38,46,49)
Strand				
Number	14.5	13.8	10.0	(3,14,14)
Calculation	22.5	20.0	**37.3**	(63,26,20)
Fractions	23.1	20.5	21.8	(30,14,20)
Ratio	2.3	2.1	5.5	(5,6,6)
Algebra	2.9	2.6	3.6	(0,6,6)
Measurement	15.6	22.6	**10.0**	(0,17,14)
Geometry	11.0	11.3	**6.4**	(0,9,11)
Position	3.5	3.1	1.8	(0,3,3)
Statistics	4.6	4.1	3.6	(0,6,6)

Note. Figures in parentheses in the 2016 Test column represent the percentage of items on each paper (1,2,3) rounded to the nearest whole number.

2.2 Cognitive domains

2.2.1 Reading test

For the first three of the five cognitive domain strands for reading, the *Test Framework* prescribes that there should be "questions across the range of demand 1 to 4, predominantly 2 to 4" (STA, 2015a, p. 13). Although all items were predominantly rated by STA as 2-4 for each strand in the 2016 test, and while ratings for accessibility of target information (Acs_Tgt_Inf) and task-specific complexity (Tsk-Sp_Cmpl) did span 1-4, the range of complexity of target information (Cmpl_Tgt_Inf) only spanned a range of 1-3 (again, as rated by STA).

The *Test Framework* is somewhat more prescriptive with regards to the intended weightings of response strategy (Rsp_Str) demands. Table 6 shows that the 2016 test was consistent with intended weightings.

Table 6. Intended and enacted weightings of response strategy demands

Response strategy rating	Percentage of total mark (*Test Framework*)	Percentage of total mark (2016 test)
1	20-40%	32%
2-3	40-70%	56%
4	6-24%	12%

For the fifth strand, technical knowledge required (Tec_Knw_Rq), the *Test Framework* prescribes that "the majority of questions will be at [levels 1 and 2]" (STA, 2015a, p. 13). Thirty of the thirty nine items on the 2016 test were rated by STA at levels 1 and 2, while the remainder (23%) were rated level 3.

As mentioned earlier, it may be problematic if too many low-level cognitive demands were sampled for a particular content sub-domain, whilst too many high-level cognitive demands were sampled for another. Figure 4 shows STA's cognitive demand ratings for items from each of the different content domain strands. Content strands 2c, 2e, 2f, and 2g were only assessed by one item each, thus these strands only sampled from one level in each cognitive domain strand. Of the other content strands, none sampled only high or low demands, although strands 2a and 2b mostly sampled only low Rsp_Str demands while strand 2d sampled higher Rsp_Str demands. See STA (2015a, p. 7) for what these strands refer to.

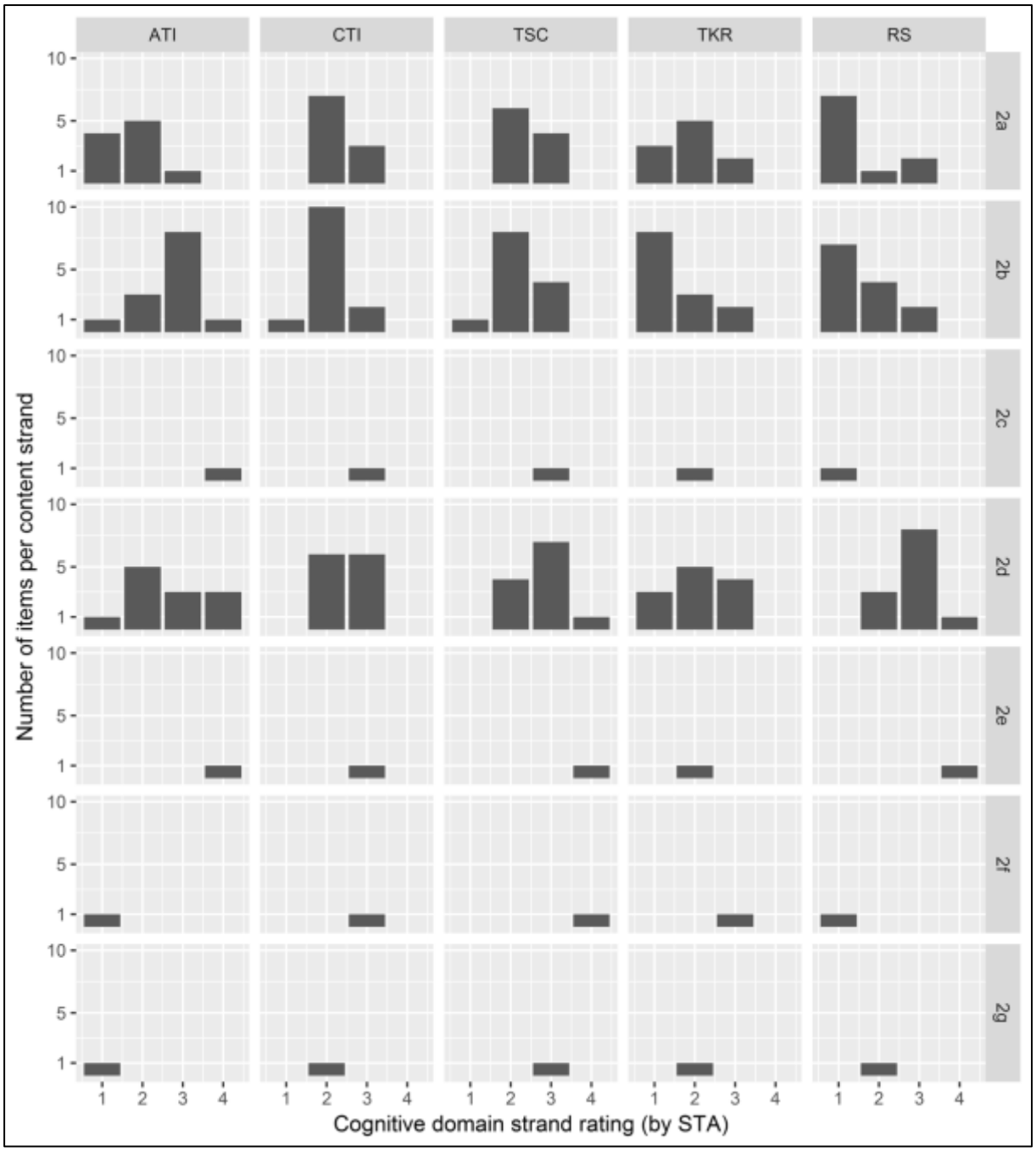


Figure 4. STA ratings of reading cognitive demands by content strand

Note. "ATI" = Accessibility of target information; "CTI" = Complexity of target information; "TSC" = Task-specific complexity; "RS" = Response strategy. "2a" = give / explain the meaning of words in context; "2b" = retrieve and record information / identify key details from fiction and non-fiction; "2c" = summarise main ideas from more than one paragraph; "2d" = make inferences from the text / explain and justify inferences with evidence from the text; "2e" = predict what might happen from details stated and implied; "2f" = identify / explain how information / narrative content is related and contributes to meaning as a whole; "2g" = identify / explain how meaning

is enhanced through choice of words and phrases; “2h” = make comparisons within the text.

2.2.2 Mathematics test

Table 7 shows the intended and enacted weightings for the 2016 mathematics test. Findings show that the test was consistent with the weightings prescribed in the *Test Framework* for depth of understanding (Dpth_Und) and response strategy (Rsp_Str) demands. However, it appears as though the test contained fewer than intended marks for high computational complexity (Comp_Cmpl) demands (ie ratings of 4), and one more than intended for medium Comp_Cmpl demands (ie ratings of 2-3). Similarly, the test contained slightly more marks than intended for medium spatial reasoning/data interpretation (Spt_Rea/Dat_Int) demands, and one more than intended for low Spt_Rea/Dat_Int demands. Given that enacted weightings fall only marginally outside intended weightings, this presents little cause for concern.

Table 7. Intended versus enacted weightings of mathematics cognitive domain strands

Strand and rating	Marks (*Test Framework*)	Marks (2016 tests)
Dpth_Und		
1	30-70	38
2-3	30-70	67
4	10-20	5
Comp_Cmpl		
1	0-30	6
2-3	60-100	**101**
4	10-20	**3**
Spt_Rea/Dat_Int		
1	60-80	**81**
2-3	0-20	**28**
4	0-10	1
Rsp_Str		
1	40-70	57
2-3	40-70	52
4	0-10	1

Figure 5 shows the sampling of cognitive demands within each mathematics content area, combining information across the three papers (note that this represents the number of items classified at each level, rather than the number of marks). Most content strands did not stand out as only sampling either high-level or low-level demands. The calculation (‘C’) and fractions (‘F’) content areas are noteworthy for having considerable numbers of items rated at the lowest demand level (1) for depth of understanding, spatial reasoning / data interpretation, and response strategy. This

was primarily due to the types of items presented within Paper 1 (ie straightforward addition, subtraction, multiplication, and division of integers and fractions/decimals). The number ('N') and statistics ('S') content areas are noteworthy for having comparatively few items rated at the higher demand levels (3 or 4). When presented in terms of items (rather than marks) it is worth noting that only three items received the highest rating for the depth of understanding strand, and only one item received the highest rating for each of the three remaining strands.

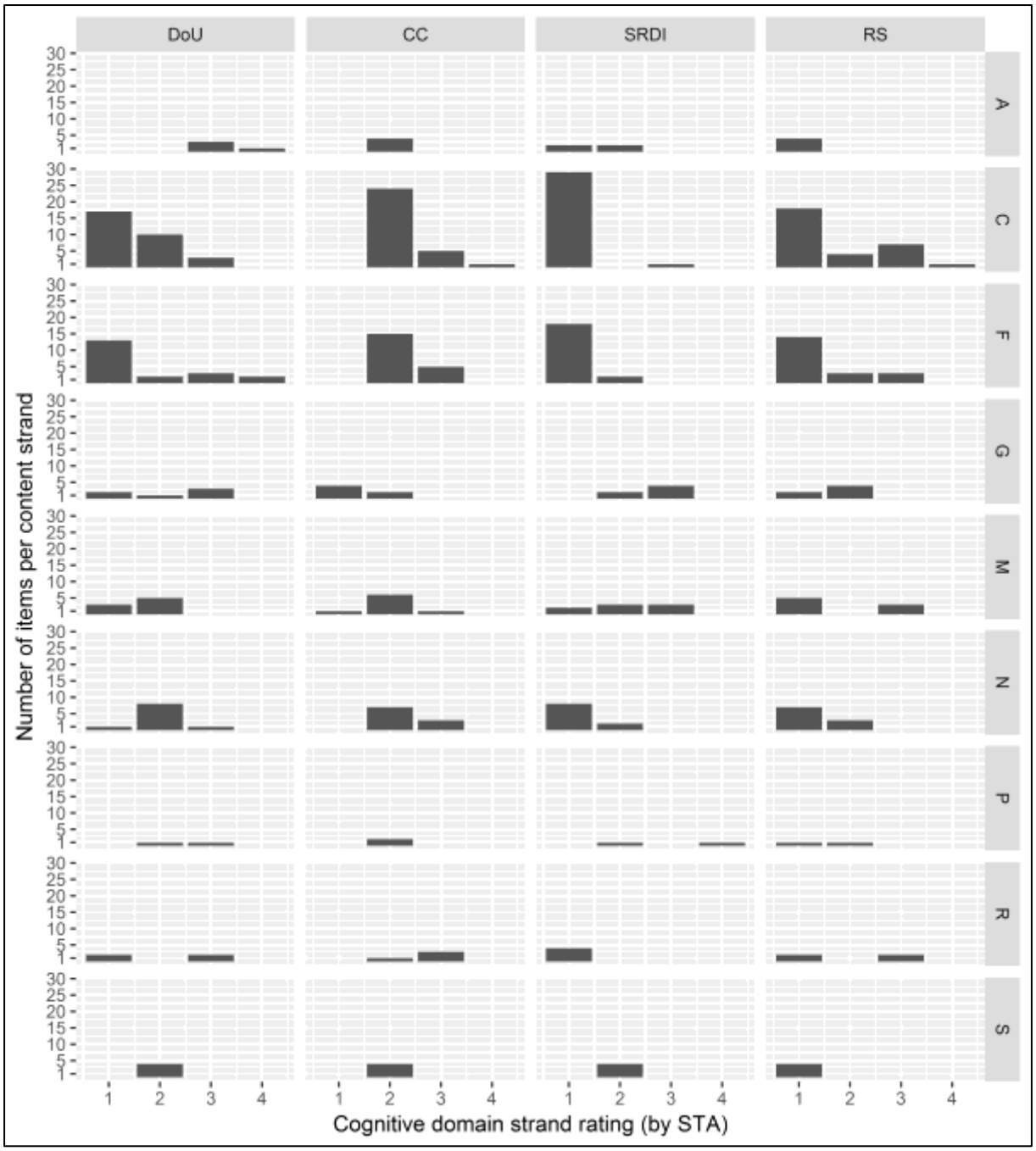


Figure 5. STA ratings of mathematics cognitive demands by content strand

Note. “DoU” = Depth of understanding; “CC” = Computational complexity; “SRDI” = Spatial reasoning / data interpretation; “RS” = Response strategy. “N” = Number – number and place value; “C” = Number – addition & subtraction, multiplication & division; “F” = Number – fractions; “R” =Ratio and proportion; “A” = Algebra; “M” = Measurement; “G” = Geometry – properties of shapes; “P” = Geometry – position and direction; “S” = Statistics.

3 Study 2 – Content domain ratings

3.1 Participants

With support from STA and from national subject associations[11], we recruited a group of subject matter experts (SMEs) who would take part in both Study 2 and Study 3. A full list, along with their job roles, can be found in Appendix C. These SMEs were recruited on the basis of their level of expertise in reading / mathematics, and on their availability to attend a group meeting in person for Study 3. For both reading and mathematics, 12 SMEs took part in the project; 6 of these had previously marked the 2016 key stage 2 tests (recruited via STA), and 6 had not (recruited via the subject associations). This number exceeds the typical sample size used in similar studies (which has been reported as 3-10 SMEs by Herman, Webb, & Zuniga, 2007). The purpose of this sampling approach was to gain a wide range of viewpoints and experiences, rather than to test for differences between markers and non-markers.

One mathematics expert, and two reading experts were also recruited to pilot the method described in the following section, before materials were sent to the main group. These pilot SMEs did not also take part in the main study, but their feedback on the task was used to hone the task instructions and response spreadsheets.

3.2 Methodology

SMEs from both groups completed Study 2 remotely (ie individually and at home). They were posted physical copies of the *Test Framework*, the 2016 reading / mathematics test, the mark scheme, and an extract from the national curriculum. They were also emailed a response spreadsheet.

For each item on the test, SMEs were posed three questions, for which they were required to record their responses on the response spreadsheet:

1. Do you agree that the content domain reference allocated to this item by the STA is the **most appropriate** classification, from the full list provided by STA in Table 3[12] of the *Test Framework* document?
    a. yes – it's the <u>most</u> **appropriate**

[11] Mathematical Association, British Society for Research in Learning Mathematics, National Centre for Excellence in the Teaching of Mathematics, National Association of Mathematics Advisers, National Association for Teachers of English, National Association of Advisers in English, United Kingdom Literacy Association, Centre for Literacy in Primary Education.

[12] Table 2 for reading; Table 3 for maths.

b. it's reasonably appropriate, but *not* the most **appropriate**

c. it's somewhat **appropriate**

d. no – it's **inappropriate**

2. If you believe that any of the **alternative** content domain references are at least **reasonably appropriate** for this item, from the list provided by STA in Table 3, then please record no more than two of them, in order of appropriateness (i.e. most appropriate first).

3. If you believe that **none** of the available content domain references is even **reasonably appropriate** for this item, from the list provided by STA in Table 3, then please briefly explain why.

After recording their responses, they were asked to return the spreadsheet by email. Shortly after the deadline for returning the completed spreadsheets (2 days for reading; 6 days for mathematics), SMEs attended a group day in person. Most of this day was concerned with Study 3, but the SMEs were also asked to comment on Study 2 during some focus group discussions. The quantitative results presented within the following sections shall therefore also be complemented with qualitative insights where appropriate, which were gained during these focus group discussions.

3.3 Results

3.3.1 Reading test

Figure 6 shows the number of SMEs who agreed that the content classification made by STA, for each item on the reading test, was the most appropriate. For all but one item (97% of items in the test), the majority of SMEs did agree that STA's classification was the most appropriate classification to make.

Question 10 was the only item where more than half of SMEs did not agree that STA's classification was the most appropriate (strand '2b' – 'retrieve and record information / identify key details from fiction and non-fiction'). However, most of these SMEs indicated that STA's classification was still 'reasonably appropriate'. Of the 7 that did not rate STA's classification as being the most appropriate, 5 suggested strand '2d' as an alternative ('make inferences from the text / explain and justify inferences with evidence from the text'). During the focus group discussions, one SME suggested that this disagreement may have been due to an overlap between the content and cognitive domains for this item:

> "I really struggled with [Question 10], and could almost have put it in a few different places… But looking at the cognitive domain today has made me even more aware of why I struggled… Had I gone very much into cognitive processes when I was trying to unpick it, and does [the] answer

> I've given in the content domain reflect more the cognitive processes?... I think I've maybe got myself into a whole cognitive domain discussion."

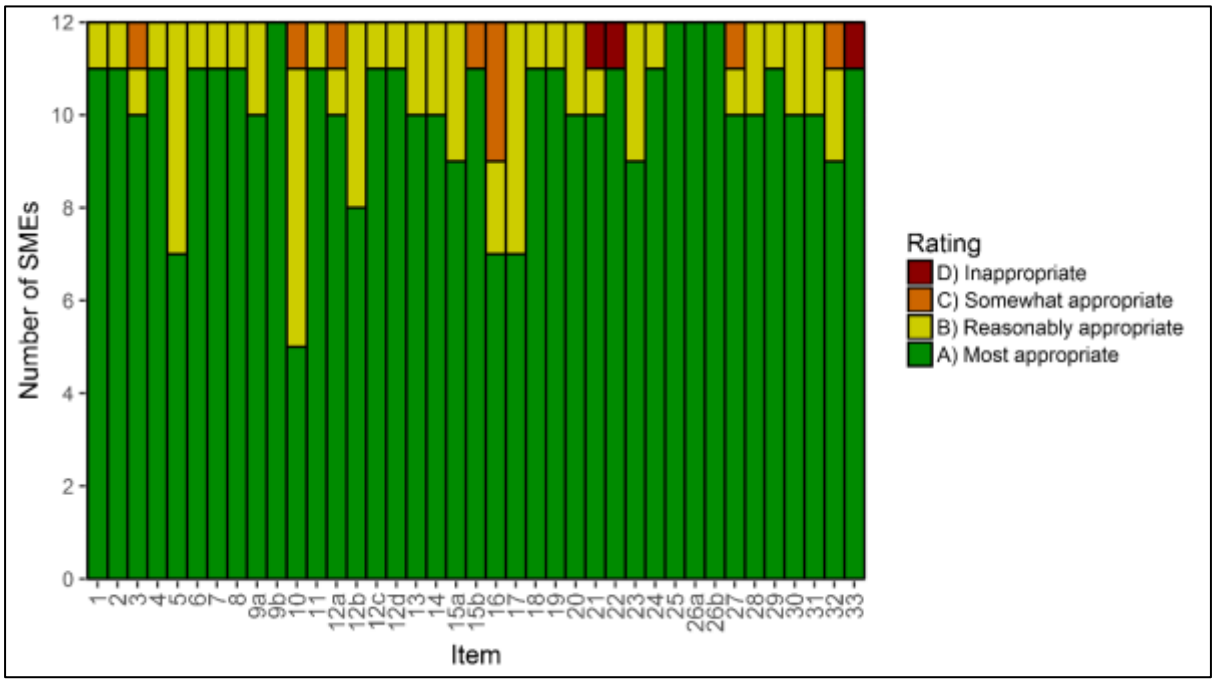


Figure 6. Content domain ratings for reading

When asked for their views on the reading content domain strands more generally, some SMEs suggested that domain '2f' was quite difficult to understand ('identify / explain how information / narrative content is related and contributes to meaning as a whole'). One pointed out that none of the items on the test related to domain '2h' ('make comparisons within the text'), and another commented that none of the strands outlined in the *Test Framework* covered 'authorial intent or authorial viewpoint', which is part of the national curriculum. These comments aside, SMEs generally agreed that the content domain strands as outlined in the *Test Framework* were an appropriate representation of the range of content domain strands relevant to the subject.

3.3.2 Mathematics test

Figures 7-9 show the content domain strand classification ratings for the 3 mathematics papers. In contrast to the reading group, more than half of the mathematics SMEs responded that the STA classification was not the most appropriate for a total of 25 items (28% of all mathematics items). However, in all but 2 cases, the most commonly suggested alternative classifications for each item were from different sub-strands of the same strand that was classified by the STA,

suggesting that the SMEs did agree that the STA had selected the most appropriate strand when classifying most items.

As an example, Questions 16 and 17 from Paper 1 were classified as '5F8' by the STA ('read, write, order and compare numbers with up to three decimal places').[13] The most common alternative classification suggested by the SMEs for each of these items was '5F10' ('solve problems involving numbers up to three decimal places'). One comment made during the focus groups may suggest that disagreement between sub-strands of the fractions ('F') strand may have been due to difficulties in fully understanding how each sub-strand is being defined.

> "[5F8] in particular came up quite a lot for me, and all it says in the content domain is 'read, write, order and compare numbers with up to three decimal places'. And that must also cover addition and subtraction of decimals, but nowhere is addition and subtraction of decimals mentioned at all."

More than half of SMEs did disagree with the STA over the most appropriate sub-strand and strand for Questions 13 and 19 from Paper 3. These had been classified by the STA as belonging to the 'fractions, decimals and percentages' and 'number and place value' strands respectively. However, the majority of SMEs classified these items as belonging to the 'calculations' strand. These particular items were not commented upon during the focus groups.

[13] It is worth mentioning that the lack of explicit reference to decimal calculations in the national curriculum may have made this exercise more difficult, in terms of knowing how to reference such questions. For the arithmetic paper, most areas of disagreement related to this content area.

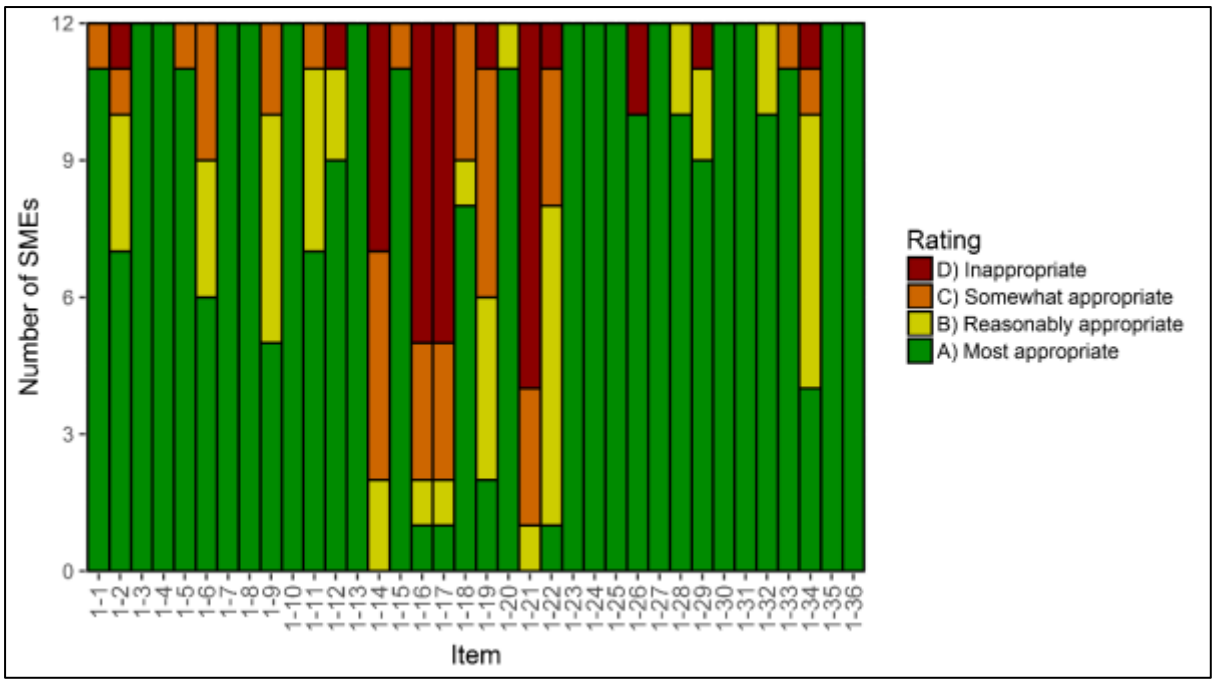


Figure 7. Content domain ratings for mathematics Paper 1

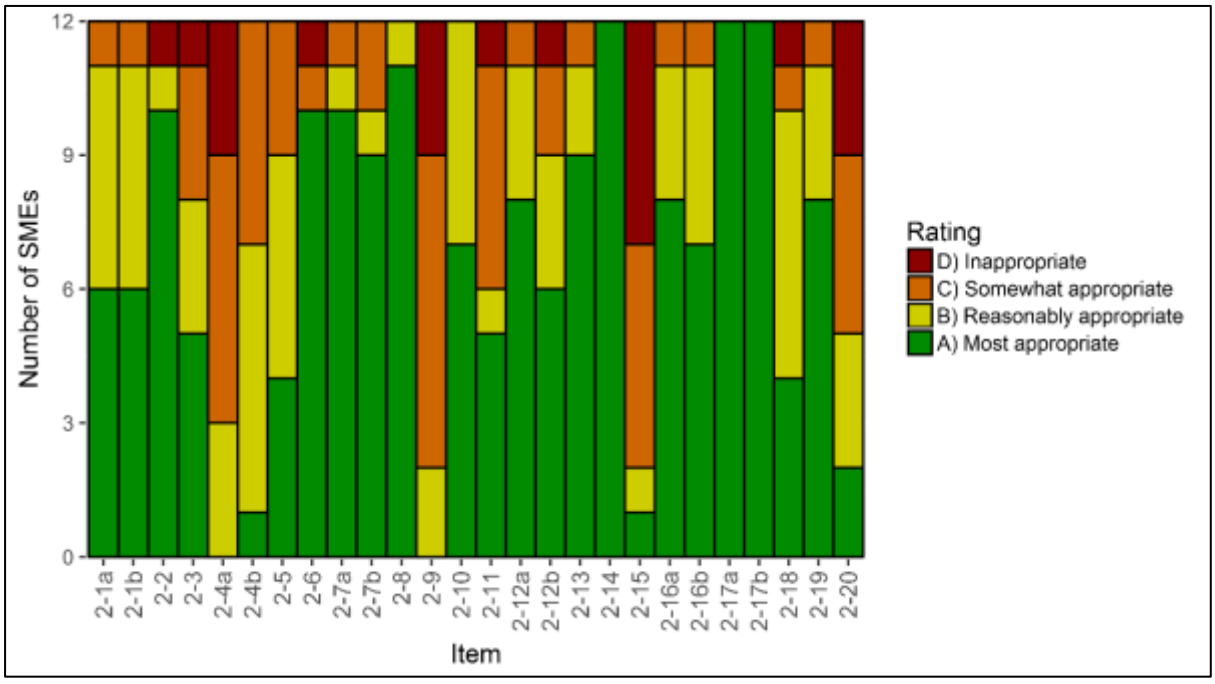


Figure 8. Content domain ratings for mathematics Paper 2

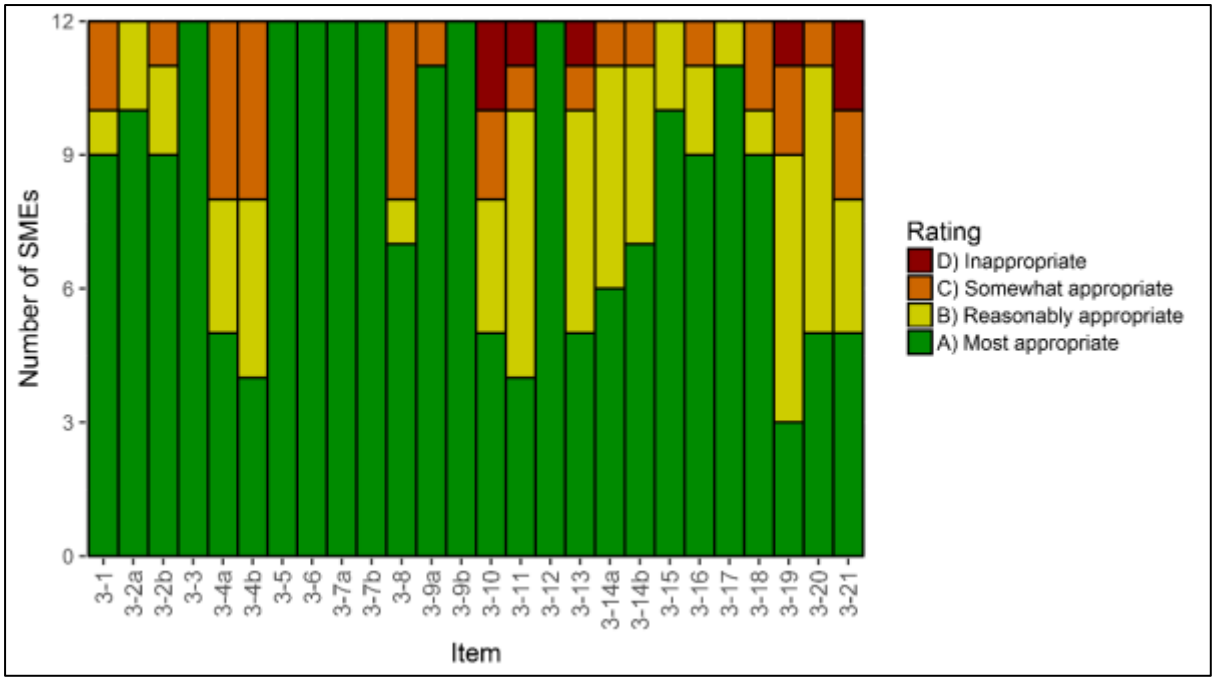


Figure 9. Content domain ratings for mathematics Paper 3

During the mathematics focus groups, some SMEs suggested that there may have been too much repetition of certain strands within the 2016 tests (ie that they were over-sampled). Calculation items were one example. Other areas of the curriculum / framework were perceived by some to be under tested (eg 'measurement' – consistent with the findings of Study 1; Table 5). Some SMEs also suggested that the level of certain items did not appear to match the descriptions given in the framework.

> "I think that there was some repetition of content when I went through the content domain of paper one… I was marking off a criteria for one question and then two questions later it appears again, and you just think, 'why have you put that question in as well, haven't we just tested that?'"
>
> "I also felt there wasn't a lot of Year 4 coverage. You know, the test is to cover the whole key stage 2, isn't it? There wasn't a lot that I could contribute to Year 4"
>
> "There was no measure - use a protractor or use a ruler or any type of measure or anything, was there?"
>
> "The last mention of inverse was Year 4, but there's no way you could say [that] that question was a Year 4 objective. There's no way you'd give that to a Year 4 child. It was a top end Year 6, if not even higher."

Some SMEs questioned the appropriateness of the separation of content domains within the *Test Framework*. In particular, it was suggested that this may have led to

an overlap between several of the domains, and that the nature of this separation does not reflect the reality of real-world mathematics.

> "[There's] a real problem if you separate out measure from number, because most of measure is either testing your understanding of number or your understanding of calculation."
>
> "Also there's a bit of a problem around fractions, because fractions are also part of the number system, they're also part of multiplication and division… so you've got this weird thing where what we're doing is compartmentalising mathematics, not in the way mathematics works."

4 Study 3 – Cognitive domain ratings

4.1 Participants

The same participants from Study 2 also took part in Study 3. However, one mathematics SME was unable to take part in Study 3 due to illness, meaning that there were 12 SMEs for reading, but 11 for mathematics.

As before, the following methodology was remotely piloted by the same 3 SMEs (1 for mathematics; 2 for reading) in advance of the main study. Training materials and task instructions were amended based upon their feedback.

4.2 Methodology

As the cognitive domain rating exercise was conceptually more difficult than the content domain rating exercise, SMEs were asked to attend a group day in person to complete Study 3. Separate days were held for reading and mathematics.

One of the decisions that needed to be made for this study was how much training SMEs should be given on STA's cognitive domain strands. One option would have been to give them no training, and ask them to as a group produce their own strands and rating scales, which might or might not bear much resemblance to those developed by STA (and then to rate the 2016 questions in these terms). However, to develop a new analytical framework would be a research exercise in itself, and would therefore not have been feasible within the scope of this study. Another possibility would be to fully train them in how STA conceived each of the strands and rating scales (and then to rate the 2016 questions in those terms). Arguably, though, this would not add a great deal of value, because our SMEs would not be able to exercise sufficient independence.

To secure an appropriate balance, SMEs were given enough training so that they could understand what STA meant by each strand, but were not given specific training on how exactly to apply the scales. In other words, we trained them on what is meant by each strand, and what might be considered to be high or low demand in each strand, but did not give specific training on what should be rated as a 1, 2, 3, or 4 on each rating scale (note, in contrast, that STA has developed specific criteria for applying these rating points consistently). Our approach allowed SMEs to exercise their independent expert judgements concerning levels of demand, but without deviating from the meaning of each strand as intended by STA. The materials for the training sessions can be found in Appendix A and Appendix B.

After being trained, SMEs were split into 2 groups (with each group containing equal numbers of markers and non-markers) and were asked to practise applying each of the rating scales on a small number of practice items. Practice items were taken

from the 2015 sample assessment materials.[14] For the reasons already discussed, the purpose of this exercise was not to standardise between SMEs on their exact ratings of these practice items, but rather to foster a common understanding of what is meant by each strand. SMEs were encouraged to discuss each item, and to suggest whether and why each item might be rated high or low on each domain, but discussions of specific numerical ratings were discouraged. Both groups then came together to share any points of clarification.

After the training session was completed, SMEs were asked to independently rate each item on the 2016 test against each cognitive domain strand. They were instructed to use STA's definition for each of the cognitive domain strands ("Do your best to understand exactly what STA means by demand for each strand") but to use their own sense of how demanding each item is ("Don't necessarily try to replicate STA's application of the 4-point scale… Use the scale to indicate your opinion of the level of demand of the item"). This approach allowed the SMEs to maintain a degree of independence. Due to time limitations, ratings were only made for Papers 2 and 3 in mathematics.

Our final training slide ended with the task instructions in full (expressed in essentially the same way for mathematics and reading):

> Please apply this rating scale by considering the range of questions that you consider might reasonably be included in a key stage 2 reading test – given current national curriculum programmes of study – and the degree to which those questions might vary in terms of level of demand for each strand.
>
> - what STA means by ATI, CTI, TSC, RS and TKR[15]
>
> - your opinion of the level of demand of the item (for each strand)
>
> - given your opinion of what an NC test for Y6 pupils might reasonably include

While these instructions seemed unproblematic for the mathematics SMEs, one of the reading SMEs raised a question concerning the meaning of 'reasonably' which precipitated further discussion. The idea of 'reasonableness' was included in the instructions with specific reference to the national curriculum for reading, ie reasonable in relation to how the national curriculum is specified. However, in a sense, the SMEs were also being asked to judge what **they** considered to be reasonable, which made the task feel more value-laden, particularly in the context of a test that many teachers (during the summer) had criticised as unreasonably demanding, particularly for children with special educational needs. This led to

[14] https://www.gov.uk/government/collections/national-curriculum-assessments-practice-materials
[15] Accessibility of target information, complexity of target information, task-specific complexity, response strategy, and technical knowledge required.

discussion of the likelihood that an item judged to be ‘high demand’ for certain children might be judged ‘low demand’ for others. Following this helpful discussion, it was suggested that SMEs’ judgements should be referenced: to national curriculum requirements; and to the cohort of children – as a whole – for whom the tests were designed.

After completing their ratings, the SMEs were once again separated into 2 groups (a different combination to before), and focus group discussions were held (mean length = 57 minutes). In these groups, the SMEs were asked to discuss their confidence in applying the scales and whether any of the cognitive domain strands made less or more sense than others, including whether they perceived any overlap between the strands. They were also asked about their thoughts on the demands of the papers. Finally, they were asked about the content domain rating task and their thoughts on the various content domain strands outlined in the *Test Framework* (for Study 2). A whole group discussion was then held to discuss any outstanding issues (mean length = 7 minutes).

The following sections present the findings of the cognitive domain rating exercise, complemented with qualitative findings from the focus groups where appropriate.

4.3 Results

4.3.1 Reading test

Inter-rater consistency

The first question that we needed to explore was whether there was any consistency between the ratings provided by our experts. This was to determine the extent to which they could be said to hold similar views on the items which they rated, and on the rating scales. As already noted, they were not standardised to apply the scales in the same way, so evidence of inter-rater consistency would indicate the extent to which they reached similar conclusions independently. It was important to consider this question before considering the extent to which they agreed with STA’s ratings.

Table 8. Rater consistency statistics for reading[16]

Strand	Cronbach alpha	Unweighted % agree.	% of items for which >½ SMEs gave exact same rating	Weighted % agree.	Gwet's AC_2 Coefficient (95% CI)
Acs_Tgt_Inf	.93	46.08%	56%	80.77%	.56 (.49 – .62)
Cmpl_Tgt_Inf	.84	38.42%	59%	78.37%	.54 (.49 – .59)
Tsk-Sp_Cmpl	.92	45.57%	64%	82.39%	.61 (.57 – .66)
Tec_Knw_Rq	.83	36.83%	51%	75.82%	.47 (.41 – .54)
Rsp_Str	.97	60.72%	90%	88.18%	.76 (.71 – .81)

Table 8 shows the inter-rater consistency statistics for the reading SMEs. The columns presented within this table can be interpreted as follows:

1. The five cognitive domain strands for reading.
2. Cronbach's alpha is a measure of internal consistency, providing estimates of the consistency with which the SMEs applied each of the rating scales. Values above 0.8 are commonly considered to indicate good internal consistency (all scales exceeded this benchmark).
3. Unweighted percentage agreement statistics are the propensities for perfect agreements between all possible pairs of SMEs' ratings.
4. The percentage of items for which more than half of SMEs gave the exact same rating gives another indication of the propensity for SME agreements.
5. One of the common criticisms of unweighted agreement statistics is that all disagreements are treated the same, regardless of the size of disagreement. In reality, smaller differences in opinion are often considered to be more acceptable than larger differences. To give an example from the current context, a disagreement of 1 scale point (eg the difference between a rating of 2 and a rating of 3) might be considered to be a reasonably acceptable difference of opinion, whereas a disagreement of 3 scale points (eg the difference between a rating of 1 and a rating of 4) would indicate poor inter-

[16] Acs_Tgt_Inf = accessibility of target information; Cmpl_Tgt_Inf = complexity of target information; Tsk_Sp_Cmpl = task-specific complexity; Tec_Knw_Rq = technical knowledge required; Rsp_Str = response strategy.

rater agreement. Weighted percentage agreement statistics acknowledge the fact that small disagreements might be considered to be acceptable in some instances. In effect they allow one to widen the definition of 'agreement' to give partial credit for these smaller, more acceptable differences in opinion. For the purposes of the current analyses, 100% weighting was given to perfect agreements, 75% weighting was given to disagreements of 1 scale point (in effect, recasting them as '75% agreement'), 25% weighting was given to disagreements of 2 scale points, and 0% weighting was given to disagreements of 3 scale points.

6. A common criticism of percentage agreement statistics (both unweighted and weighted) is that they do not take the possibility of chance agreement into account (ie through random guessing). To address this, Gwet's AC_2 coefficients are presented, which estimate levels of agreement that cannot be attributed to chance. This particular coefficient was chosen in favour of other similar coefficients (eg Cohen's Kappa or Krippendorff's alpha) due to its more robust calculation of chance agreement (see Gwet, 2008). The AC_2 statistics reported here, which apply the same agreement weightings as before, would all be considered to indicate at least 'moderate' agreement according to commonly cited benchmarks (eg Altman, 1991; Landis & Koch, 1977). SMEs 'substantially' agreed in their Rsp_Str ratings (according to Altman, 1991).

During the focus groups, several SMEs stated that they would have appreciated more training on the cognitive domains in order to become more confident in applying the rating scales. Some also stated that although it was easy to distinguish between high and low demands, it was difficult to distinguish between ratings of 2 and 3 (one suggested that a 3-point rating scale may have been easier to apply). Despite these concerns, the coefficients presented in Table 8 do suggest that the SMEs were quite consistent with one-another in making their judgments, even though they were instructed to apply their own standards during the task.

Table 9. Mean cognitive demand ratings for reading (markers vs. non-markers)

Strand	Non-markers	Markers
Acs_Tgt_Inf	2.26	2.16
Cmpl_Tgt_Inf	2.54	2.44
Tsk-Sp_Cmpl	2.75	2.49
Tec_Knw_Rq	2.53	2.29
Rsp_Str	1.91	1.77

Mean SME ratings for each strand were separated between markers and non-markers, to see whether there were any differences in how these two groups were applying the rating scales. Table 9 shows that ratings were reasonably comparable between the two groups (statistical tests were not conducted due to small group sample sizes).

Levels of agreement between ratings made by the SMEs and STA

Table 10 shows the mean SME and STA ratings of the test for each cognitive domain strand, along with the differences between those means.

Table 10. Mean cognitive demand ratings for reading (SMEs vs. STA)

Strand	Mean STA rating	Mean SME rating	Difference	Gwet's AC_2 Coeff. (95% CI)
Acs_Tgt_Inf	2.41	2.21	-0.20	.72 (.62 - .82)
Cmpl_Tgt_Inf	2.33	2.49	+0.16	.86 (.79 - .93)
Tsk-Sp_Cmpl	2.56	2.62	+0.06	.83 (.75 - .91)
Tec_Knw_Rq	1.87	2.41	**+0.54*** **	.68 (.56 - .80)
Rsp_Str	2.00	1.84	**-0.16***	.82 (.74 - .91)

* $p < .05$; *** $p < .001$ in a one sample t-test (difference from 0)
Note. Gwet's AC_2 in this instance is the agreement between the mean SME rating for each item (rounded to the nearest whole number) and the STA rating for each item. The same agreement weightings as before have been applied.

One-sample t-tests (difference from 0 disagreement) showed that the mean of SME ratings for Tec_Knw_Rq (2.41) was statistically significantly higher than the mean of STA ratings (1.87) (t (38) = 5.23, $p < 0.001$). The mean of SME ratings for Rsp_Str (1.84) was statistically significantly lower than the mean of STA ratings (2.00) (t (38) = 2.22, $p = 0.03$). Although these differences were statistically significant, it should be noted that the sizes of these differences were small (less than 1 scale point difference). All other differences were non-significant. When the same weightings as before were applied, Gwet's AC_2 coefficients showed that there was good agreement between SME and STA ratings, again supporting the notion that disagreements between SMEs and STA were small. Graphs are presented in Appendix D that compare SME and STA ratings on an item level for each strand.

During the focus groups, some SMEs commented on the high demands for Tec_Knw_Rq, and the low demands for Rsp_Str:

> *Technical knowledge required*
>
> "There were words in there that they couldn't even have figured out from the context. So yeah, I think it was very demanding for 10-11 year-olds."
>
> "I think this paper particularly [was] very outside of some children's experiences. Now other people will say that shouldn't be a problem; we want the bar to be high… but actually for some children I think that was probably a significant barrier in this test"
>
> *Response strategy*
>
> "[Rsp_Str – high demand rating] says [in the training materials], 'answers are extended and require students to fully structure and organise their own responses', and really there's only those two questions that really require that. So, I suppose for me that was lacking somewhat."

Comparisons with the findings of Study 1

Further analyses were conducted to determine whether the differences between SMEs' and STA's ratings had any impact on the results of Study 1: ie intended versus enacted weightings. For Acs_Tgt_Inf, Cmpl_Tgt_Inf, and Tsk-Sp_Cmpl, the *Test Framework* states that there should be "questions across the range of demand 1 to 4, predominantly 2 to 4" (STA, 2015a, p. 13). Mean SME ratings of Acs_Tgt_Inf demands (rounded to the nearest whole number) for each item ranged 1-4 (as did STA ratings for Acs_Tgt_Inf). Mean SME ratings of Cmpl_Tgt_Inf demands ranged 2-4 (the range of STA ratings was 1-3). Mean SME ratings of Tsk-Sp_Cmpl demands ranged 2-4 (the range of STA ratings was 1-4). All these strands were predominantly rated in the 2-4 range by the SMEs (Acs_Tgt_Inf = 87% of items; Cmpl_Tgt_Inf, Tsk-Sp_Cmpl = 100% of items).

As noted in Study 1, the reading *Test Framework* is more prescriptive with regards to Rsp_Str demands. Table 11 shows that SME ratings for Rsp_Str were still within the guidelines specified by the *Test Framework*.

We have already seen, from Table 10, that the mean of all SME ratings for Tec_Knw_Rq (2.41) was statistically significantly higher than the mean of STA ratings (1.87). Despite this, the majority of the mean SME ratings for Tec_Knw_Rq, across all items, were still rated either level 1 or level 2, consistent with the *Test Framework* specification. Having said that, this was only just the case. Twenty one of the thirty nine item ratings were at level 2, seventeen at level 3, and one at level 4; so, in fact, none of the mean SME ratings for Tec_Knw_Rq was actually at level 1. Indeed, when considered in terms of marks, rather than items, this tipped the balance further. From this perspective, the majority of *marks* on the 2016 reading test were associated with a rating of level 3 (50%) and level 4 (2%), for the

Tec_Knw_Rq strand, according to the mean ratings of our SMEs. (We discuss the possibility of the Tec_Knw_Rq strand being slightly more demanding than anticipated by STA in the accompanying review of evidence.) Although this finding is interesting, it is important to remember that we did not ask our SMEs to attempt to replicate the STA's approach to rating items, as specified in the *Test Framework*. Our SMEs' judgements reflect a more intuitive impression of the level of demand associated with each item, with no attempt to standardise those judgements across SMEs.

Figure 10 shows the sampling of mean cognitive demands for items testing different content areas, as rated by the SMEs. The same conclusions can be drawn as in Study 1 (ie that there is a roughly even spread of demands across the different content strands, with the possible exception of Rsp_Str demands – items from content strands 2a and 2b mostly only sample low level demands, while items from strand 2d sampled some higher level demands).

Table 11. Intended and enacted weightings of Rsp_Str demands (with SME ratings)

		Percentage of total mark (2016 test)	
Response strategy rating	% of total mark (*Test Framework*)	STA ratings	SME ratings[a]
1	20-40%	32%	28%
2-3	40-70%	56%	60%
4	6-24%	12%	12%

[a] Based on mean rating for each item, rounded to the nearest whole number

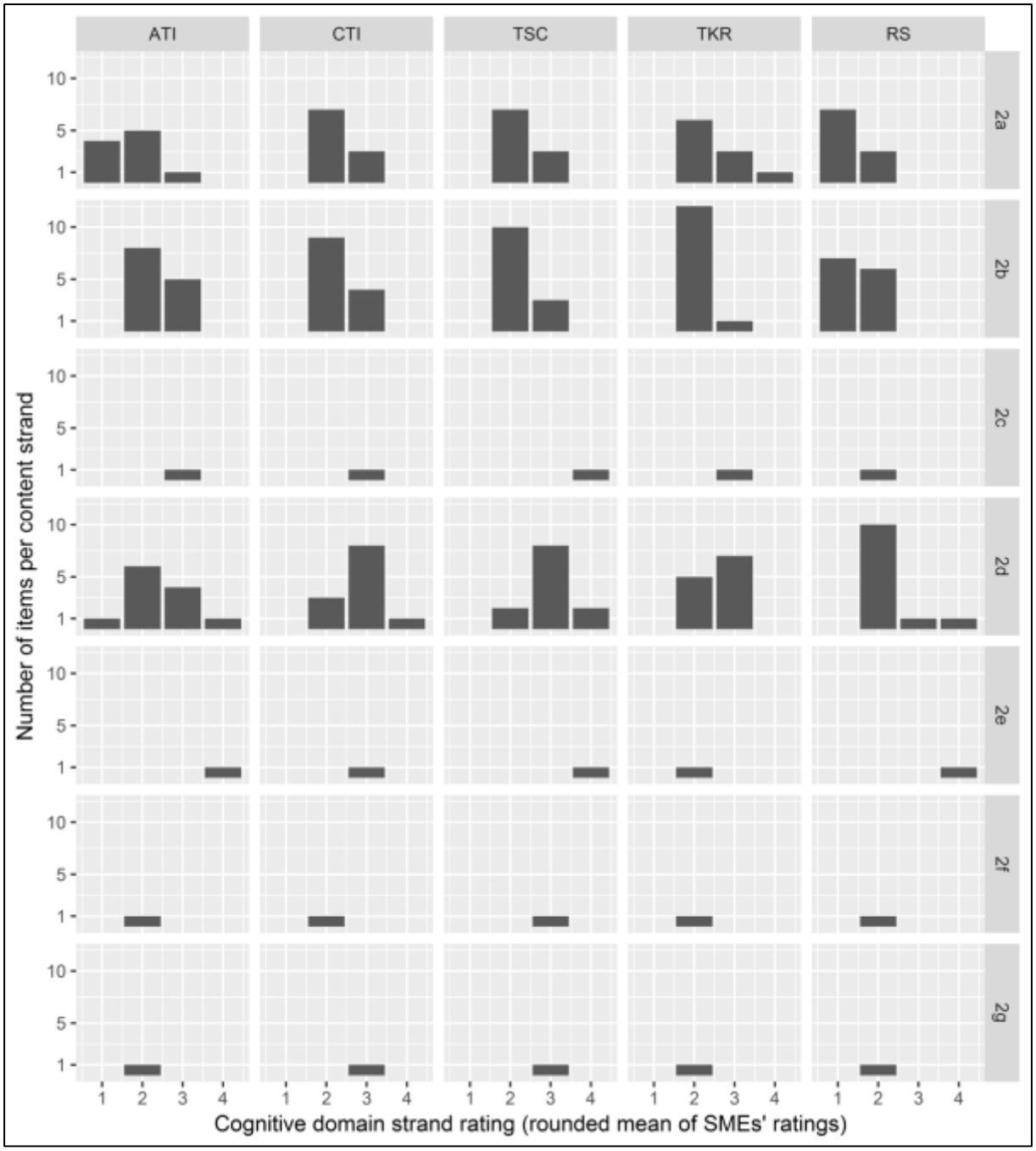


Figure 10. SME ratings of reading cognitive demands by content strand

Note. "ATI" = Accessibility of target information; "CTI" = Complexity of target information; "TSC" = Task-specific complexity; "RS" = Response strategy. "2a" = give / explain the meaning of words in context; "2b" = retrieve and record information / identify key details from fiction and non-fiction; "2c" = summarise main ideas from more than one paragraph; "2d" = make inferences from the text / explain and justify inferences with evidence from the text; "2e" = predict what might happen from details stated and implied; "2f" = identify / explain how information / narrative content is related and contributes to meaning as a whole; "2g" = identify / explain how meaning

is enhanced through choice of words and phrases; “2h” = make comparisons within the text.

Issues arising from the discussion

During the focus group discussions, our SMEs were invited to express their views on two general issues: the appropriateness of the cognitive domain strands, and the levels of demands in the 2016 test.

Although SMEs acknowledged the usefulness of this approach for test development, several noted that there was a degree of overlap between some domains. This made it difficult to apply the rating scales independently from one another. For example, several comments were made regarding the overlap between the Cmpl_Tgt_Inf and Tec_Knw_Rq domains, with SMEs often finding it difficult to separate these domains. Tsk-Sp_Cmpl was also considered by some to be too broad in focus.

> “I think where I had the discrepancy was between the complexity of the target information [and] the technical knowledge [required], mainly based around vocabulary… It was those two in particular where I felt it was a bit clouded.”
>
> “[Tsk-Sp_Cmpl] includes so many different things. It feels like that covers an awful lot… [So] you were doing a bit of a best fit, because there’s a lot in that one.”

The SMEs identified a number of other types of demands that they felt were not covered by any of STA’s strands. For example, several commented on the length of the test, and the amount of text that pupils needed to read within the time allotted to them. Some also noted that the texts were not very engaging for pupils, which may have imposed demands on motivation.

> “That is a significant time pressure I think on an 11 year-old to do that, even if they are a fluent reader and are able to orientate themselves around the text as we’ve spoken about, I think an hour is a quite limited amount of time.”
>
> “I just wondered if [STA] do any kind of maths around how many minutes that it actually takes [to read the texts].”
>
> “The Lost Queen [is] a story that you’re left halfway through, so it’s OK for a reading test, but… part of a reading test must be about engaging the reader in order for them to want to go back and sufficiently to dig deep into it.”

Several SMEs also raised concerns about the order in which items were presented within the test. Specifically, there was the suggestion that some of the more

demanding items were presented too early on in the paper, discouraging pupils at a relatively early stage. In response to these comments, Figure 11 was plotted. This shows the mean demand for each item (ie averaged across all 5 cognitive domain strands) in the order in which they were presented in the paper. The same graphs for each individual cognitive domain are presented in Appendix E. It is important to note that this computation of item 'mean demand' provides only a crude indication of its 'overall' intended/judged level of demand, because there are no guarantees concerning how different demands will interact. But it seems quite likely that there will be some relationship between 'mean' and 'overall' demand; and therefore that there will also be some relationship between 'mean demand' and the difficulty of each item. This issue is explored in greater detail in a separate review of evidence on the accessibility of the 2016 reading test.

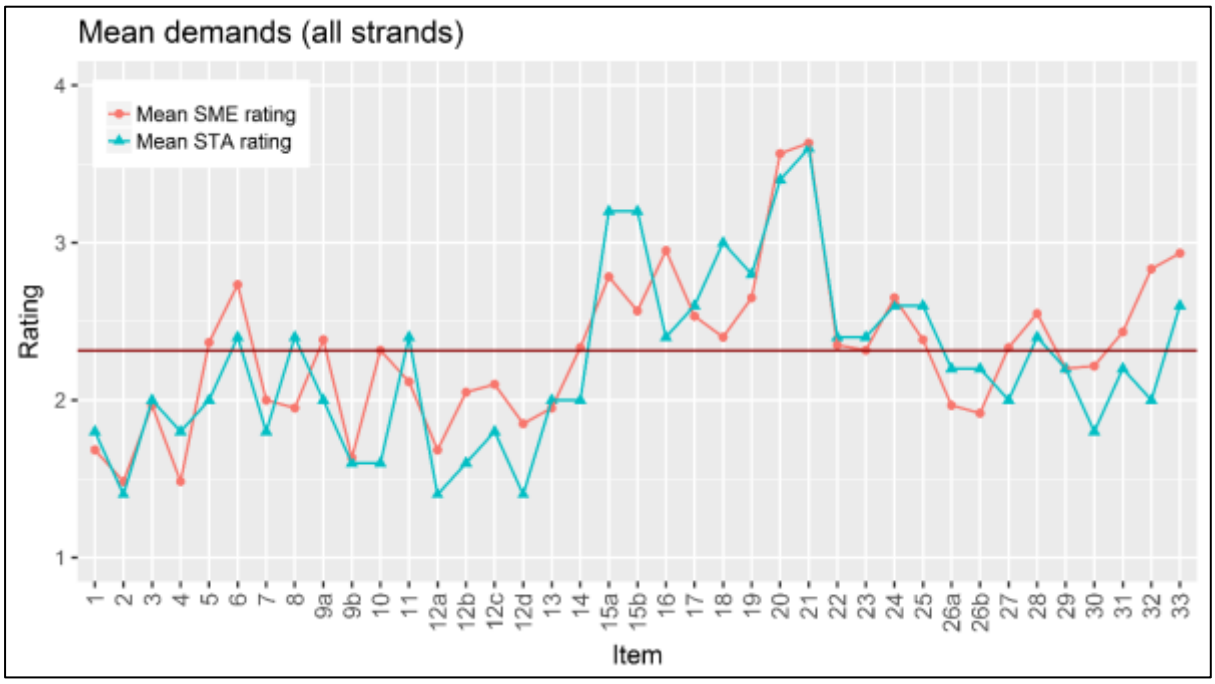


Figure 11. Mean cognitive domain ratings for each item on the reading test

Note. The horizontal red line indicates the grand mean of SME ratings for the paper.

It is also important to note that reading paper was divided into three sections: Questions 1-11; Questions 12-22; and Questions 23-33 (each relating to a different reading text). It is interesting to see, from Figure 11, how the middle section demonstrates a rapid transition from the lowest mean demand items to the highest mean demand items. Questions 20 and 21, towards the end of the second section demonstrate far higher mean demand than any of the questions in the third section, consistent with the following quote:

> "Why put question 20 and 21 in the middle? Give it to them when they're fresh, or give it to them right at the end when only your really gifted and talented kids are going to get there!"

4.3.2 Mathematics papers

Inter-rater consistency

During the focus groups, the mathematics SMEs said that they felt sufficiently confident in their understanding and application of the rating scales. However, as with the reading group, several noted a difficulty in distinguishing between ratings of 2 and 3, which is interesting, bearing in mind that they had been asked to apply the scale however they wished to (rather than in relation to an external point of reference).

Table 12. Rater consistency statistics for mathematics[17]

Strand	Cronbach alpha	Unweighted % agree.	% of items for which >½ SMEs gave exact same rating	Weighted % agree.	Gwet's AC_2 Coeff. (95% CI)
Paper 2					
Dpth_Und	.92	41.54%	62%	79.13%	.52 (.42 – .62)
Comp_Cmpl	.94	55.94%	96%	87.10%	.73 (.69 – .77)
Spt_Rea/Dat_Int	.95	64.55%	85%	88.13%	.79 (.67 – .90)
Rsp_Str	.88	49.02%	73%	81.49%	.63 (.50 – .76)
Paper 3					
Dpth_Und	.93	46.16%	73%	81.78%	.59 (.50 – .68)
Comp_Cmpl	.96	63.63 %	92%	89.56%	.77 (.69 - .86)
Spt_Rea/Dat_Int	.83	50.77%	88%	80.23%	.60 (.48 - .73)
Rsp_Str	.90	42.94 %	77%	79.86%	.57 (.46 - .69)

Similar to the findings for reading, although some SMEs stated that they would have felt more confident with more training, statistics did suggest that they applied the rating scales quite consistently (Table 12). All internal consistency statistics (Cronbach's alpha) exceeded commonly cited benchmarks for 'good' consistency. Gwet's AC_2 coefficients again suggested that the SMEs agreed to at least a 'moderate' degree, and SMEs 'substantially' agreed upon their ratings for Spt_Rea/Dat_Int (for Paper 2) and Comp_Cmpl (for both Paper 2 and 3) (according to the thresholds described by Altman, 1991). As with reading, there were no substantial differences in the ratings made by markers and non-markers (Table 13).

[17] Dpth_Und = depth of understanding; Comp_Cmpl = computational complexity; Spt_Rea/Dat_Int = spatial reasoning / data interpretation; Rsp_Str = response strategy.

Table 13. Mean cognitive demand ratings for mathematics (markers vs. non-markers)

Strand	Non-markers	Markers
Dpth_Und	2.23	2.24
Comp_Cmpl	2.11	2.24
Spt_Rea/Dat_Int	1.66	1.72
Rsp_Str	1.69	1.95

Levels of agreement between ratings made by the SMEs and STA

Table 14 shows the mean STA and SME ratings of the test for each cognitive domain strand, along with the difference between these values. None of the differences between the SME and STA ratings were statistically significant on either paper ($p > .05$). Agreement between SME and STA ratings was good for all strands on both papers. Graphs are presented in Appendix D that compare SME and STA ratings on an item level for each cognitive domain.

Figures 12 and 13 show SME and STA ratings according to the order of items presented within the test papers. These graphs also show good agreement between the SME and STA ratings. The ordering of items across the papers in terms of 'mean demand' does not appear to be particularly noteworthy (see Appendix E for the same graphs for each domain strand).

Table 14. Mean cognitive demand ratings for mathematics (SMEs vs. STA)

Strand	Mean STA rating	Mean SME rating	Difference	Gwet's AC_2
Paper 2				
Dpth_Und	2.46	2.21	-0.25	.74
Comp_Cmpl	2.27	2.18	-0.09	.82
Spt_Rea/Dat_Int	1.62	1.60	-0.02	.89
Rsp_Str	1.65	1.73	+0.08	.81
Paper 3				
Dpth_Und	2.15	2.26	+0.11	.68
Comp_Cmpl	1.92	2.15	+0.23	.81
Spt_Rea/Dat_Int	1.73	1.76	+0.03	.67
Rsp_Str	1.88	1.89	+0.01	.80

Note. Gwet's AC_2 in this instance is the agreement between the mean SME rating for each item (rounded to the nearest whole number) and the STA rating for each item. The same agreement weightings as before have been applied.

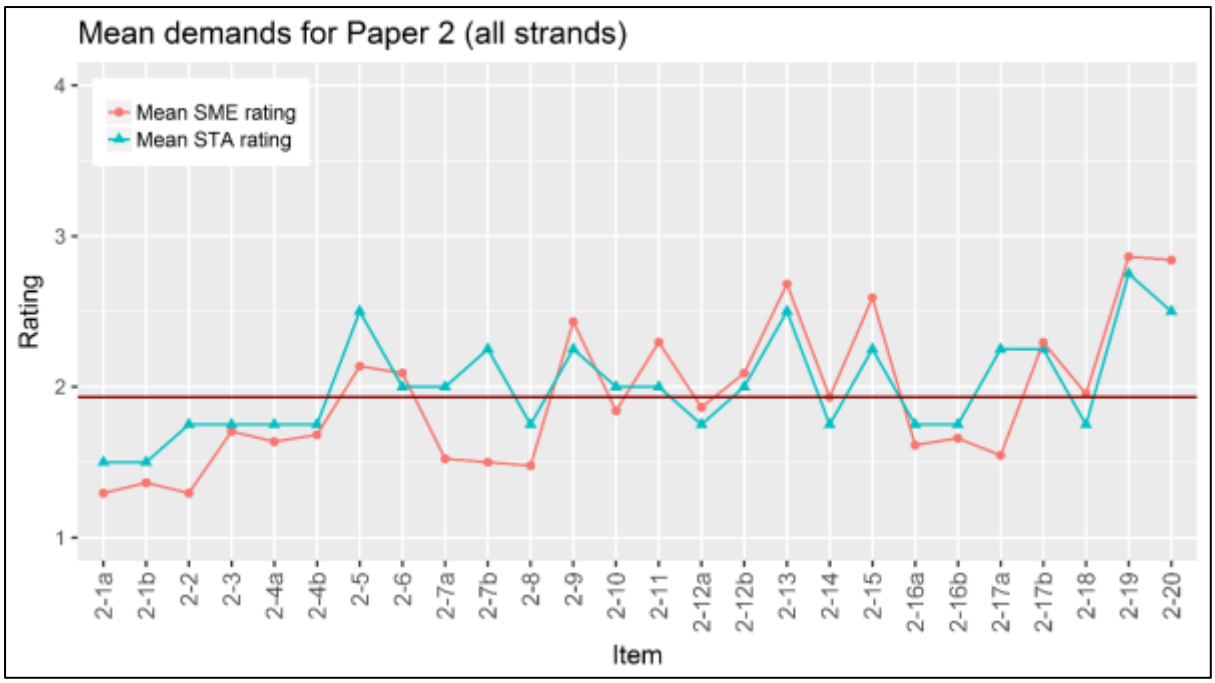


Figure 12. Mean cognitive domain ratings for each item on mathematics Paper 2

Note. The horizontal red line indicates the grand mean of SME ratings for the paper.

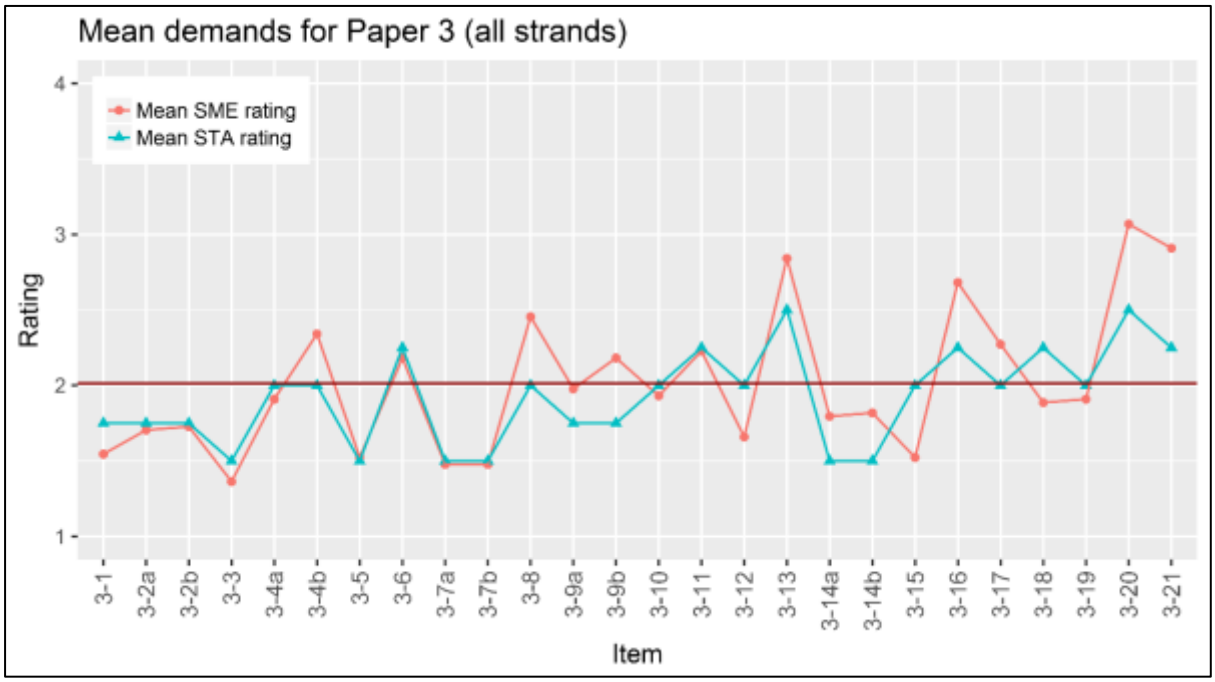


Figure 13. Mean cognitive domain ratings for each item on mathematics Paper 3

Note. The horizontal red line indicates the grand mean of SME ratings for the paper.

Comparisons with the findings of Study 1

Because the mathematics SMEs were only asked to make ratings for Papers 2 and 3, owing to time limitations, we were unable to compare the weightings of demands as rated by the SMEs against those presented within the *Test Framework* and the national curriculum (ie to replicate Tables 4 and 5).

Figure 14 shows the sampling of mean cognitive demands for items testing different content areas, as rated by the SMEs (Papers 2 & 3 only). As with the STA ratings, most content strands did not appear to sample only high-level or only low-level demands. Although, similarly to the STA ratings, for strand 'N' ('number and place value'), nearly all items were rated as having low level demands across all four strands (ie rounded mean ratings of 1 or 2). SMEs' ratings of demands for the calculation ('C') and fractions ('F') content areas appear higher in comparison with STA's ratings shown in Figure 5. However, this is primarily due to the fact that SMEs did not provide demand ratings for Paper 1, which tended to sample lower level demands for these content areas.

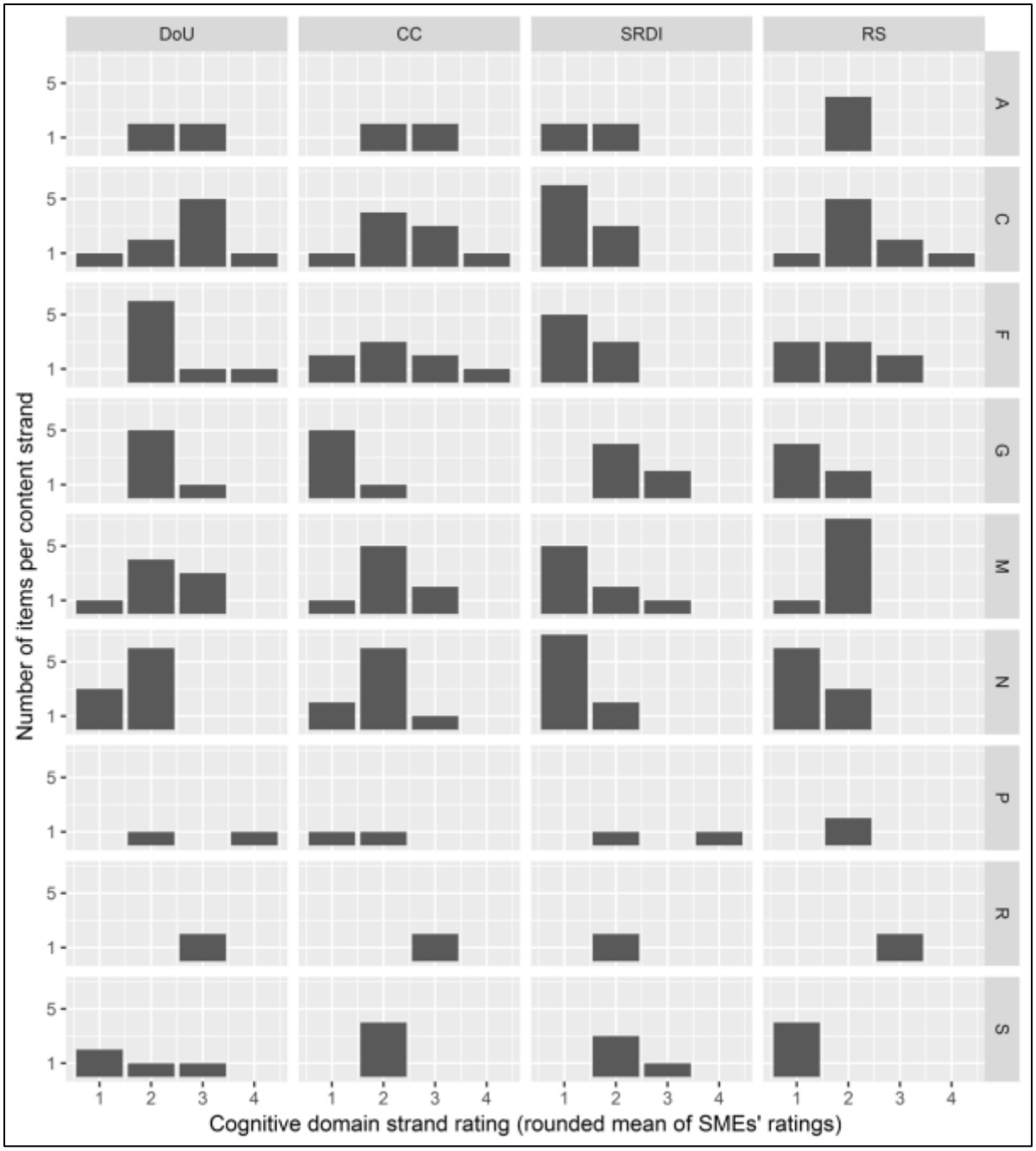


Figure 14. SME ratings of mathematics cognitive demands by content strand

Note. “DoU” = Depth of understanding; “CC” = Computational complexity; “SRDI” = Spatial reasoning / data interpretation; “RS” = Response strategy. “N” = Number – number and place value; “C” = Number – addition & subtraction, multiplication & division; “F” = Number – fractions; “R” =Ratio and proportion; “A” = Algebra; “M” = Measurement; “G” = Geometry – properties of shapes; “P” = Geometry – position and direction; “S” = Statistics.

Issues arising from the discussion

When asked to comment on STA's cognitive domain strands, the mathematics groups generally felt that they were appropriate and easy to understand. The only exception to this was the suggestion that it may have been better to separate the spatial reasoning and data interpretation elements of Spt_Rea/Dat_Int.

> "I think it's a bit strange to lump the geometry together with the data interpretation actually… I'd rather they'd be two separate things."

As with the reading group, the mathematics SMEs also identified a number of other types of demands, which they did not feel were captured by any of STA's cognitive domain strands. Some commented on the time pressures imposed by the tests. Some also commented on the degree of working memory required to solve certain questions. However, this may fall under STA's computational complexity strand. Some noted that certain questions required contextual knowledge that some pupils may not be familiar with. Comments were also made that the language of the question can impose demands that are not necessarily relevant to mathematical ability. The use of pictures might also affect demands when they do not help the pupil to answer the question. The following quotations illustrate these issues.

> "We actually sat down and we did the paper. And it was very difficult to do that paper in the time… Nothing was difficult… [but] it was really difficult to get it done [within the allotted time], even though [we are] really good at maths."
>
> "That problem was intrinsically simple mathematically, but there was so much in it… There are links to short-term memory and also working and processing… It was harder for them this year than the year before."
>
> "When you're dealing with children whose life experience is so very vastly different, when they are presented with problems of houses of 130-something thousand and they don't own a house and their parents don't and their grandparents don't, they don't understand that concept… what's simple for the vast majority of the population is not simple for everybody."
>
> "If [the question] becomes too convoluted, they just give up. They might actually be able to solve that quite competently in the classroom situation, but in a test… they just think, 'I can't waste time. I'm bamboozled by it, I'll just try the next one.'"
>
> "The picture was superfluous to the question really, it didn't add any value… Children are being on one hand told to look for any clues visually that might be there… and other times the picture's a bit random".

Finally, SMEs were asked about how well they felt the STA's cognitive domains covered the national curriculum. Although they thought that the domains were an

appropriate representation of the national curriculum, some noted that the core aims of the national curriculum for mathematics appeared to be missing from the cognitive domain strands.

> "It was a shame that the strands did not reflect the underpinning aims of the new national curriculum: fluency, reasoning, problem solving. Because it would be lovely if they had… Those are the three things that drive the national curriculum, but they're kind of lost in this: they're sort of all lumped together."

5 General Discussion

The overarching purpose of this study was to investigate STA's approach to developing national curriculum test papers, focusing specifically upon its approach to sampling learning outcomes from the key stage 2 programmes of study, from Year 3 to Year 6, for reading and mathematics. It matters that this sampling is **relevant**: that the tested learning outcomes can be traced back to the statutory national curriculum. And it matters that this sampling is **representative**: that the balance of learning outcomes tested corresponds to the balance of those learning outcomes in the statutory national curriculum (ie that learning outcomes are neither under- nor over-represented). If a national curriculum test failed to assess learning outcomes related to a particular content strand, then a pupil could score maximum marks on the test despite having failed to master any of the learning outcomes related to that strand. The inference that ought to follow from scoring maximum marks – that a pupil has thoroughly mastered the national curriculum – would be incorrect. More importantly, if, over time, successive tests repeatedly failed to assess learning outcomes related to that strand, then teachers would pick up on this regularity and would be likely to adjust their teaching accordingly. Indeed, this might occur even if the tests noticeably under-weighted those learning outcomes in a predictable manner. Before long, we might find that teachers no longer taught that content strand during key stage 2.

The evidence that we gathered suggested that STA's approach to sampling learning outcomes is robust. Given STA's interpretation of the national curriculum framework document, the *Test Framework* documents appeared to translate national curriculum teaching requirements into plausible blueprints for testing. The degree of consistency with which our independent experts rated items from the 2016 tests supported the conclusion that the way in which STA specified the content domain strands and the cognitive domain strands was plausible. This is particularly important evidence in relation to the cognitive domain because the domain strands were innovative for both mathematics and reading, having been introduced specifically for the new tests. In addition, the degree of consistency between STA item ratings and those of our independent experts supported the conclusion that STA's ratings were plausible and, by extension, that the 2016 tests sampled content and cognitive domains relevantly and representatively. In the following sections we shall consider in more depth the relationships between the tests, the *Test Frameworks*, and the national curriculum.

5.1 Relationships between the tests, the *Test Frameworks* and the national curriculum

5.1.1 Relevance and representativeness of test items to the *Test Frameworks*

Through the three studies reported here, we were able to assess the relevance and representativeness of the 2016 reading and mathematics tests to the *Test Frameworks*, both in terms of content domain classifications and levels of cognitive demands.

Starting with the content domain classifications, the results of Study 1 suggested that the 2016 reading test was appropriately representative, as the proportions of marks allocated for each content domain were all within the guidelines prescribed by the *Test Framework*. SMEs in Study 2 agreed with STA's classification for all but one item in the reading test, with the classification of that item still being rated as at least "reasonably appropriate" by each SME, thus supporting both the relevance and representativeness of the reading test items.

Findings for mathematics were less clear. In Study 1, we showed that the different content areas were appropriately represented within the test papers, under the guidelines prescribed by the *Test Framework.* However, the *Test Framework* does not specify intended weightings at the individual content strand level nor at the programme year level; and, as such, there are no official intended weighting specifications at these levels. When the enacted weightings were judged against a simple frequency count of requirements in the *Test Framework* – which closely mirrored a simple frequency count of requirements in the national curriculum – the test seemed to over-weight calculation and to under-weight (to a lesser extent) measurement and geometry. Similarly, it seemed to over-weight Year 6 content and to under-weight Year 4 content. In fact, although not specified in the *Test Framework* for mathematics, STA's internal criteria for test construction do specify that the upper key stage 2 topics should be weighted more heavily than lower key stage 2 ones, reflecting the assumption that the majority of pupils will be working at the levels set out in the Y6 programme of study by the end of key stage 2.

Of course, these findings rely on the assumption that weightings can be appropriately inferred from bullet-point frequency counts. This may not be case – we have no information on what the curriculum designers might have intended, nor even if they considered this issue in any depth – so conclusions should be drawn cautiously. However, anecdotal evidence that teachers have begun to identify and respond to perceived weighting patterns has already begun to emerge, as the

following quotation from the ongoing UCL Institute of Education[18] investigation into test preparation strategies in mathematics reveals:

> Prior to the introduction of scaled scores, teachers would talk about gradually building up the level of difficulty when teaching specific mathematical content areas, such as 'number sense and calculation', 'data handling' or 'shape and space'. Level 3, 4 and 5 test items on past key stage 2 test papers would help them understand the hierarchical nature of mathematics and how to introduce children to, for example, increasingly more difficult calculations (e.g. moving from one step to multistep problems, or from adding and subtracting whole numbers to adding and subtracting decimals). Resources such as Test Base would allow them to access available questions according to content area and difficulty level and they could simply download relevant questions when teaching a specific skill. Now that the levels have been removed, some of the teachers tell us that they just focus on getting all students to perform at level 5 in number and calculation as this is where most of the marks on the test are given and some hardly teach shape and space at all.[19]

Observations like these raise the question of whether more nuanced weightings ought to be included within future iterations of the *Test Framework*. The fact that the national curriculum does not provide any insight into weightings also highlights the role of the *Test Framework* in constructing a model of pupil proficiency in each subject area, via decisions on these matters. Since it is possible for quite different models of pupil proficiency to be constructed from the same national curriculum requirements, this raises questions concerning the extent to which and the ways in which curriculum designers ought to contribute to the construction of the corresponding *Test Framework*. Indeed, it raises the parallel question of the extent to which assessment designers ought to be involved in the construction of the corresponding programme of study. In response to questions like these, STA explained that the *Test Frameworks* were shared with members of the national curriculum review team during 2012/2013. It is worth considering whether it would be useful to develop a more formal model of collaboration for future revisions of the national curriculum and *Test Frameworks*.

In Study 2, SMEs for mathematics disagreed with STA's content classification of items on a greater number of occasions (nearly a third of all items) than for reading, but SMEs did agree in most cases with STA on the same strand (disagreements were mostly over the most appropriate sub-strand), thus supporting item relevance and representativeness at the content strand level, but not quite so categorically at

[18] See http://highstakestesting.co.uk/. The first author of the present report, Paul Newton, worked on this project, prior to joining Ofqual in 2014.
[19] https://ioelondonblog.wordpress.com/2016/07/19/life-after-levels-is-the-new-year-6-maths-test-changing-the-way-teachers-teach/

the level of sub-strands. The fact that mathematics SMEs disagreed to a greater extent than the reading SMEs is unsurprising, given the greater number of content sub-strands to choose from.

Moving on to the cognitive domain ratings, the findings of Study 1 showed that the proportions of marks allocated in the 2016 reading test for the different levels of each cognitive domain strand were within the guidelines prescribed by the *Test Framework*, thus supporting the representativeness of the 2016 test. The mathematics test appeared to have been somewhat less representative, as the marks allocated for some levels of two cognitive domain strands (computational complexity and spatial reasoning / data interpretation) were slightly misaligned with the guidelines prescribed by the *Test Framework*. Fewer marks than intended were allocated to high computational complexity demands, and more marks than intended were allocated to low and middle spatial reasoning / data interpretation demands. However, it should be noted that enacted weightings fell only marginally outside the intended weightings. Because of their size, and because it is far harder to classify the cognitive domain (as opposed to the content domain) definitively, we do not consider these discrepancies to be problematic. Although the reading SMEs rated technical knowledge required demands for the test statistically significantly higher than STA in Study 3, and response strategy demands statistically significantly lower than STA, the sizes of these differences were small. No statistically significant differences were found for mathematics. This again provides general support for the relevance and representativeness of the 2016 tests for both subjects.

5.1.2 Relevance and representativeness of the *Test Frameworks* to the national curriculum

By reflecting on some of the issues raised in the focus group discussions, we can also consider the appropriateness of the *Test Frameworks* when judged against the national curriculum for reading and mathematics; focusing specifically upon the plausibility of the content domain strands and cognitive domain strands.

With respect to the content domain, SMEs in both groups generally accepted STA's approach, and few had comments to make about the content domain strands, reflecting the fact that the content domains for reading and mathematics are far more established than are the cognitive domains. In fact, the only point of note was that mathematics SMEs raised some concern that the separation of content domain strands was somewhat artificial in relation to real-world applications of mathematics. We cover this point below in Section 5.2.3.

Further observations can be made about the cognitive domain strands. The results of Study 3 supported the plausibility of these strands, as SMEs in both the reading and mathematics groups demonstrated fairly good levels of agreement both between themselves and with the ratings made by STA (this degree of consistency and agreement would have been difficult to achieve if the strands developed by STA did

not reflect genuine demands relevant to the assessment of key stage 2 learning outcomes). However, in the focus group discussions, SMEs identified a degree of overlap between strands, which made it more difficult to apply the rating scales. For example, those in the reading group noted a degree of overlap between complexity of target information and technical knowledge required demands, suggesting that these strands may not be mutually exclusive for test items. The task-specific complexity strand was also noted to be quite broad in focus, again making it difficult to rate items independently from other domain strands. For mathematics, SMEs suggested that the spatial reasoning / data interpretation domain would be more appropriate if spatial reasoning and data interpretation elements were separated, as the demands of test items tended to come from one element or the other.

Other types of demands that do not feature in the *Test Frameworks* were also identified in the 2016 tests by our SMEs. For example, English SMEs discussed how demands for reading can be imposed by time pressures and questioned how engaging the reading texts were for pupils. Both of these factors appear to be construct-relevant (although, perhaps more so for the first than the second) despite not being overtly captured via the cognitive domain strands. As such, it is unclear from the published documentation if/how these demands ought to be managed during the test development process. Mathematics SMEs identified a number of additional types of demands. For example, some mathematics items require greater retention of information within working memory than others, thus imposing different demands on pupils (cf Ashcraft & Krause, 2007). Demands can also be imposed when candidates are required to have a degree of contextual knowledge in order to answer a question (cf Vappula & Clausen-May, 2006), and different levels of demands can again be imposed depending on the language of the question (cf Ferrara, Svetina, Skucha, & Davidson, 2011). An important consideration is whether these additional types of demands should be considered to be construct-*relevant* (ie relevant to the skills, knowledge and understanding that the tests aim to assess) or construct-*irrelevant*. The former might be targeted more explicitly in the *Test Framework* documents, and during test development, than the latter (ie by applying weightings in a similar manner to the current domain strands). However, it is still important to consider to potential effects of the latter during test development, to avoid compromising validity. For example, the inclusion of distracting pictures was noted to be one such source of construct-irrelevant demands by some mathematics SMEs.

Finally, another factor which is not considered within the *Test Framework* documents is the ordering of cognitive demands across the test papers. Both qualitative (in the focus groups) and quantitative (in Figure 11) findings suggested that there was a spike in demands mid-way through the reading test, which may have been challenging for some pupils and may even have negatively affected their chances of success on later questions (eg by limiting their time, or by reducing their motivation). A smoother ‘ramping’ of demands across the test might help to ameliorate potential

problems of this sort. This is therefore an important consideration for test development, and future iterations of the *Test Frameworks* might usefully include some guidance on ramping. It is important to acknowledge, however, that issues of ramping are particularly challenging for the reading test, because it comprises questions relating to different texts. This means that there are issues of ramping to consider within each respective set of questions as well as across the entire test. Effective ramping is also tricky to achieve when, for practical reasons, it is not possible to pre-test a large pool of questions for each text.

5.2 Further consideration of the relationship between the *Test Frameworks* and the national curriculum

To locate findings from our project within a broader context, it is helpful to foreground the following three observations and associated questions:

1. Even accepting the robustness of STA's approach to sampling learning outcomes, we still have to acknowledge that certain learning outcomes could not straightforwardly be tested – so how significant a threat to validity is this?
2. The weighting of national curriculum learning outcomes is an interpretative exercise, which is under-determined by the text of the national curriculum framework document – so how faithful are STA's weightings to the intentions of the authors of the national curriculum?
3. The translation of national curriculum learning outcomes into *Test Framework* specifications is a modelling exercise – so how faithful is STA's modelling to national curriculum requirements, including national curriculum aims?

Although none of these questions can be answered definitively, they all deserve further consideration.

5.2.1 Untested learning outcomes

The first issue relates to inevitable omissions from the *Test Framework* documents of parts of the national curriculum that are not easily testable in a paper-based format. This was true for a small number of content areas for mathematics, but was more significant for reading. The programmes of study for key stage 2 reading identify 8 content strands, several of which are not straightforwardly amenable to testing, including 'maintain positive attitudes to reading' and an additional 3 strands (from the upper key stage 2 programme of study) which explicitly identify discussion-related learning outcomes:

- participate in discussions about books that are read to them and those they can read for themselves, building on their own and others' ideas and challenging views courteously

- discuss and evaluate how authors use language, including figurative language, considering the impact on the reader
- explain and discuss their understanding of what they have read, including through formal presentations and debates, maintaining a focus on the topic and using notes where necessary.

Whereas the national curriculum for reading covers both word reading and comprehension, the *Test Framework* focuses intentionally upon comprehension. In fact, it is based primarily upon just 1 of the 8 national curriculum strands ('understand what they read') from which 5 of the 8 *Test Framework* content strands are drawn.[20] In addition, this content strand contains discussion-related learning outcomes, which are not straightforwardly amenable to testing:

- checking that the book makes sense to them, discussing their understanding and exploring the meaning of words in context
- asking questions to improve their understanding.

Included within submissions to the recent Education Select Committee inquiry into primary assessment were concerns over the degree to which the programmes of study for key stage 2 reading were unamenable to testing, including this comment from the National Literacy Trust[21]:

> The National Curriculum now incorporates requirements to support children's enjoyment of reading, as well as decoding and comprehension skills, but assessment at the end of primary school covers only children's comprehension skills (particularly deduction and inference).

Given these omissions, the *Test Framework* for reading cannot be said to completely align with the national curriculum; as acknowledged within the *Test Framework* itself. Although there are good reasons for not testing certain teaching requirements, this raises the risk that those requirements might not be taught. In previous years, such claims would have been countered by reassurance that untested aspects of the national curriculum would be assessed as dimensions of statutory teacher assessment judgements. Currently, though, this is not necessarily true, as the

[20] By way of contrast, the *Test Framework* for maths seems to be far more comprehensive in its coverage. It excludes only a few of the 173 national curriculum requirements; for instance, mental mathematics requirements (cited in 7 of 173 statements) and requirements involving practical equipment.

[21] http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/primary-assessment/written/41847.pdf

interim teacher assessment frameworks[22] do not claim to cover all of the content of the national curriculum, and this includes certain of the untested teaching requirements.

In addition, some of the content domain strands for reading that *are* amenable to testing have an intended weighting of between 0% and 6% in the *Test Framework*, implying that it might be acceptable for any particular test not to sample one or more of those (testable) content domain strands at all. Indeed, the enacted weighting of strand '2h' ('make comparisons within the text') in the 2016 test actually was 0%. Several other content strands for reading were only assessed by one item each, which made it impossible to sample a range of levels for each cognitive domain strand within those content strands. While, to some extent, this is unavoidable within a relatively short test, it emphasises the importance of monitoring content / cognitive domain sampling over time, to ensure that the same patterns do not occur year-on-year. Again, the risk is the potential backwash impact upon teaching and learning, if, for instance, it became apparent that certain strands were never tested with high-level cognitive demand questions.

Of course, problems like these could be ameliorated, to some extent, by making the test longer. Instead of a single, 50-mark reading test, we might envisage two, 50-mark reading tests. This would both improve the sampling of (testable) learning outcomes and increase the reliability of test results. However, this is not recommended, as any increase in validity would have to be weighed-up against an increase in testing impact upon pupils, testing burden upon teachers, test development and administration costs, and other potential consequences. Incidentally, when we transition from issues of validity to issues of impact upon teaching and learning, the regulatory 'baton' is passed from Ofqual to Ofsted. The research question would then be framed in terms of whether there is any shortfall between the intended curriculum and the taught curriculum.

5.2.2 Weighting learning outcomes

It is important to emphasise that, while various weightings are given in the *Test Framework* documents to the different content domain strands, such weightings do not come explicitly from the national curriculum. It is difficult to establish, therefore, the extent to which the operational decisions in the *Test Frameworks* appropriately reflect the intentions of the curriculum designers. In Table 1, we presented some comparisons between the national curriculum and *Test Framework* for mathematics, which suggested that the weightings of the *Test Framework* appeared to be fairly consistent with the national curriculum. However, this relied on the assumption that

22 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/538415/2017_interim_teacher_assessment_frameworks_at_the_end_of_key_stage_2_150716_PDFa.pdf

weightings can be accurately inferred from the number of bullet points given for each content strand, which may or may not be the case. For reading, such comparisons were not possible, due to the manner in which the *Test Framework* has re-interpreted the national curriculum. Therefore, although the 2016 tests seem to align fairly well with the *Test Framework* documents for both reading and mathematics, it is difficult to know whether the weightings given to different content areas within the *Test Frameworks* (and the tests themselves) effectively align with the curriculum designers' intentions for each content strand (if, indeed, the curriculum designers gave any thought to this issue). Similarly, the national curriculum does not state intentions in relation to the cognitive domain strands and, whilst we know there was interaction between curriculum designers and test developers during the design of the *Test Frameworks*, it is again difficult to know whether the weightings for different levels of cognitive demands found in the tests and *Test Frameworks* appropriately reflect the intentions of curriculum designers.

5.2.3 Modelling learning outcomes

Theoretically speaking, any evaluation of educational testing needs to bear in mind that, despite thousands of years of philosophical inquiry, there is still no consensus concerning what the things that we routinely claim to assess – dimensions of knowledge, skill and understanding – actually **are**; let alone unambiguous criteria for attributing them to learners. The point of a test is to manufacture situations through which such attributions can be made as unambiguously as possible. Yet, since our models of knowledge, skill and understanding in mathematics and reading are, at best, incomplete, we cannot expect to remove ambiguity entirely. Furthermore, to make both teaching and assessment tractable, we tend to decompose these highly complex constructs – reading proficiency and mathematics proficiency – into distinct elements; for example, via the content and cognitive domain strands from the *Test Frameworks*. This decomposition is pragmatically extremely useful, because it helps to ensure broad coverage, both when teaching and when testing. Yet, it carries the risk of oversimplification, fragmentation and, ultimately, misrepresentation. Exactly such concerns have been expressed recently in relation to the new mathematics curriculum and, by extension, to the new mathematics test, in a recent report from the National Union of Teachers (Little, 2016, p. 24)

> 'A high-quality mathematics education [should provide] a foundation for understanding the world, the ability to reason mathematically, an appreciation of the beauty and power of mathematics, and a sense of enjoyment and curiosity about the subject.'
>
> These laudable aims appear in the preamble to the Maths programme of study of the 2014 primary curriculum. It goes on to emphasize the importance of solving problems and the development of conceptual understanding (DfE 2014).

> Sadly none of this is carried through into the main document. Where the preamble talks about 'a highly interconnected discipline', the main body of the document is a list of disparate skills and knowledge.

To contextualise this quotation, it is worth quoting at length from the Aims section of the national curriculum framework document for mathematics (DfE, 2014, p. 108):

> The national curriculum for mathematics aims to ensure that all pupils:
>
> - become **fluent** in the fundamentals of mathematics, including through varied and frequent practice with increasingly complex problems over time, so that pupils develop conceptual understanding and the ability to recall and apply knowledge rapidly and accurately.
> - **reason mathematically** by following a line of enquiry, conjecturing relationships and generalisations, and developing an argument, justification or proof using mathematical language
> - can **solve problems** by applying their mathematics to a variety of routine and non-routine problems with increasing sophistication, including breaking down problems into a series of simpler steps and persevering in seeking solutions.
>
> Mathematics is an interconnected subject in which pupils need to be able to move fluently between representations of mathematical ideas. The programmes of study are, by necessity, organised into apparently distinct domains, but pupils should make rich connections across mathematical ideas to develop fluency, mathematical reasoning and competence in solving increasingly sophisticated problems.

The above quotation from Little (2016) raises important questions. The programmes of study for mathematics at key stage 2 are decomposed into 173 discrete elements in the national curriculum framework. These are transformed into 195 elements, distributed across nine content strands, in the *Test Framework*. But is this decomposition plausible? The fact that our independent experts almost entirely agreed with STA over the primary classification of each item into one of the nine content strands clearly provides some support for the plausibility of those categories, as a basis for organising teaching and testing.[23]

Yet, a more penetrating question concerns the extent to which the decomposition process threatens the assessment of fluency, mathematical reasoning and problem solving, which are central to the mathematics proficiency construct, according to the national curriculum. Written submissions to the recent Education Select Committee

[23] This is not to say that each item in the test assessed only one content strand; merely that it was possible to identify a primary content strand fairly unambiguously.

inquiry into primary assessment raised exactly this concern, alongside others related to STA's approach to sampling learning outcomes. A selection of comments from the Mathematical Association (MA) with the Association of Teachers of Mathematics (ATM) joint submission[24] helps to illustrate this (numbers have been added, for ease of reference):

> 1. The intended curriculum content is adequately assessed by the 2016 mathematics tests.
> 2. The 2016 tests are not consistent with the aims of the new curriculum: fluency and conceptual understanding, reasoning and problem solving.
> 3. In the 2016 tests, teachers consider over half of the written arithmetic test (40 marks) would in fact be better done mentally, but the curriculum no longer includes 'mental methods as a first resort'. This seems a retrograde step since mental maths builds up both conceptual understanding and fluency with number as a foundation for future mathematics learning, including algebra.
> 4. Fluency is far more than quick and accurate use of an algorithm, whether appropriate or not: it also requires highly-valued flexible and generalised use of procedures and their inverses. The arithmetic paper does not currently test these aspects at all, and much of the reasoning papers is simply procedural. If children are to be taught deep fluency, test papers must be seen to value it.
> 5. Teachers estimate that only about 25% of the 'reasoning' papers (total 70 marks) comprise items that require mathematical thinking (i.e. problem solving and reasoning): the rest represent routine operations. The design of the tests needs to move to better reflect the aims of the curriculum, with an improved balance between 'fluency' in core skills (currently representing 75% of the 'reasoning' papers as well as all of the arithmetic paper), and problem solving/reasoning.
> 6. Further, items on these latter papers often require a number of steps for solution, yet there are usually just two marks, meaning that a partially correct solution often receives no credit at all. The approach to marking these tests therefore needs re-thinking so that evidence of conceptual understanding is rewarded.

[24] http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/primary-assessment/written/41625.pdf

Similar concerns were raised in submissions from the Advisory Committee on Mathematics Education (ACME)[25] and from the National Association of Mathematics Advisers (NAMA)[26] including this additional comment from the NAMA submission:

> 7. Neither the tests nor the teacher assessment standards reflect key constructs in key stage 1 and key stage 2 mathematics, nor do either ensure pupils' coverage of the whole curriculum. The tests necessarily focus on many small bits of mathematics, and standards give priority to particular elements of maths. This can lead to teachers teaching small bits of maths in a fragmented way rather than helping children to make connections.

It is interesting to note that the MA/ATM submission accepted that the curriculum content was adequately assessed (comment 1), whilst going on to question the adequacy of sampling for fluency, conceptual understanding, reasoning, and problem solving (comment 2). To some extent, their concerns over fluency related to the removal of the mental mathematics paper (comment 3), which meant that there was no mechanism through which to assess these requirements. Yet, they also referred to the notion of 'deep fluency' (comment 4), suggesting that the test papers did not value it. Subsequent comments expanded concerns over the weighting of problem solving, reasoning and conceptual understanding (comments 5 and 6). The comment from NAMA illustrates similar concerns, noting the ideal of an integrated discipline, as well as negative impacts upon teaching and learning when curriculum and test frameworks fail to represent this interconnectedness adequately (comment 7).

The national curriculum framework document does seem to present mixed messages concerning the interconnectedness of the key stage 2 programmes of study for mathematics, and concerning the centrality of those thinking skills – mathematical reasoning, problem solving and conceptual understanding – that might scaffold this integration. Perhaps in recognition of the threat of fragmentation, 34 of the 173 national curriculum content requirements specifically incorporate the word 'problem' to highlight the importance of situating content demands in problem solving contexts. This translated into 37 of the 195 elements of the *Test Framework* for mathematics, spread across 7 of the 9 content strands. Eight of these 37 elements were actually tested in the 2016 test papers; predominantly in multi-mark questions on Papers 2 and 3. This accounted for 20 of the 70 available marks from Papers 2 and 3 (29%).

---

[25] http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/primary-assessment/written/42301.pdf

[26] http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/primary-assessment/written/42216.pdf

Beyond explicit reference to problem solving within curriculum framework statements, the purpose of the cognitive domain strands from the *Test Framework* was to highlight related thinking skills. Evidence from Study 3 is therefore of some relevance. Very few items from the 2016 tests were given the highest rating, by STA, for any of the cognitive domain strands: a rating of four was associated with just 5 of the available 110 marks for the depth of understanding strand; and with only 1 of the available 110 marks for both the spatial reasoning / data interpretation strand and the response strategy strand. The lowest rating was given far more frequently: 38 marks for depth of understanding; 81 for spatial reasoning / data interpretation; and 57 for response strategy. Once again, these enacted weightings were more-or-less within the intended weighting ranges for each of the strands. Yet, whether the intended weighting ranges sufficiently targeted the higher ratings is unclear; because the national curriculum framework is unclear on this matter. Moreover, even if subsequent tests included more items targeting higher ratings, it is still not clear that this would necessarily address the specific concerns related to mathematical reasoning, problem solving and conceptual understanding that the mathematics associations have raised.

What this discussion highlights is the possibility of a disjunction between the holistic aims of the national curriculum (which clearly foreground fluency, mathematical reasoning, problem solving, and an integrative approach to mathematics) and its atomistic 'bullet point' presentation of teaching requirements (which may obscure fluency, mathematical reasoning and problem solving). It also highlights the possibility that STA's interpretation of the national curriculum framework document – expressed in the *Test Framework* for mathematics – may similarly not capture those aims adequately, particularly in relation to problem solving. Consequently, although results from the present investigation concluded that the *Test Framework* documents appeared to translate national curriculum teaching requirements into plausible blueprints for testing – given STA's interpretation of the national curriculum framework document – that conclusion might not follow from an alternative interpretation.

Bearing this in mind, it is worth noting how the *Test Framework* distinguishes between rating points 3 and 4 on the depth of understanding cognitive domain strand:

3 use facts and procedures to solve more complex problems

4 understand and use facts and procedures **creatively** to solve complex or **unfamiliar** problems [taken from Table 5 of the mathematics *Test Framework*, emphasis added].

If, as might be argued, the essence of problem solving involves being able to generalise conceptual understanding creatively, to solve unfamiliar problems in which the method of approach is not immediately obvious, then the acknowledged

lack of test items rated 4 on the depth of understanding strand might well be interpreted as an underrepresentation of problem solving, when judged in relation to the national curriculum Aims statement. This seems to be at the heart of concerns expressed by the mathematics associations. Their overriding concern is that the relative absence of 'genuine' problem-solving questions (and marking criteria) is having a negative backwash impact upon teaching and learning, encouraging teachers to adopt an unduly procedural approach to mathematics that actively frustrates the ultimate objective of cultivating generalizable, conceptual understanding. Whilst the present research is unable to make a determination on this point, the question remains worthy of further study.

5.3 Study limitations

Although our evidence suggested that STA's approach to sampling learning outcomes is robust, this needs to be put in context, to ensure that it is not over-interpreted. The following paragraphs identify important caveats that should be borne in mind.

An approach to sampling learning outcomes is just one component of an approach to test development, which is just one component of an overarching procedure for testing. As such, even questions that appear to sample effectively, reflecting appropriate content and cognitive demands, may still fail to elicit evidence of genuine knowledge, skill and understanding (and therefore not sample effectively, in practice). This can occur, for instance, if questions are confusingly worded, or set in obscure contexts. It would require a different kind of study to investigate threats to validity such as these. Consequently, a conclusion concerning the robustness of STA's approach to domain sampling only considers a single 'link' in the argument 'chain' concerning the validity of national curriculum testing arrangements. That said, it is a very important link.

Due to the inescapably subjective nature of the kind of judgments upon which our conclusions are based, and due to unavoidably limited number of SMEs who provided these judgements, our conclusions should not (and could not) be considered definitive. It is possible that somewhat different patterns of results might have been found, had we recruited a different set of SMEs. In particular, future research might consider whether professionals who work with important subgroups of pupils, eg SEND or EAL pupils, share similar views to those whom we sampled.

We deliberately chose to include both SMEs who were extremely familiar with the new tests (senior markers) and SMEs who were at some remove (non-markers). Diversity has been identified, in the literature, as important to this kind of research (Davis-Becker & Buckendahl, 2013). Although their response patterns appeared to be reasonably comparable, it is still possible that the judgements of either groups may have been skewed by their different experiences. For instance, when discussing

their approach to rating the response strategy strand for mathematics, some of the markers observed that questions which asked for working did not necessarily require evidence of working for the award of full marks (assuming the correct answer was provided). This highlighted a potential tension between what the question demanded and what the mark scheme demanded. Our studies were framed in terms of what the questions demanded, rather than the mark schemes. However, the fact that this potential for tension was only identified in relation to a particular type of mathematics question led us to believe that this was probably not a major concern.

Several SMEs stated that they would have felt more confident in their cognitive domain strand ratings had they received more thorough training. Indeed, had the length of our training been increased, it is possible that this might have improved their understanding of those strands, and this might have increased the consistency between their ratings. Having said that, their ratings were quite consistent, which is all the more interesting bearing in mind that they were asked to apply the rating scales as they (each) saw fit, rather than as they imagined the STA might.

It is possible that social desirability effects (as noted by Davis-Becker & Buckendahl, 2013; Sireci, 1998) may have influenced the degree of agreement between ratings by our SMEs and those made by STA[27]. For instance, in Study 2, SMEs may have felt pressure to agree with the classifications made by STA. Similarly, in Study 3, they may have attempted to guess the STA cognitive domain ratings, rather than applying their own judgements (despite being given instruction to the contrary). Nevertheless, although such effects are possible in studies like these, we have no reason to believe that our results were significantly affected.

Finally, it is worth noting that our research was undertaken at a certain level of granularity. Had it been undertaken at a different level of granularity, it is possible that different issues might have been identified. For instance, the UCL Institute of Education research project mentioned in Section 5.1.1 is currently investigating the kinds of questions that have appeared in key stage 2 mathematics tests over the years; to see whether certain kinds of question may seem (to teachers) to be more common than others. This study is being undertaken at a level of granularity that enables teachers to identify, for example, '2-step money problems' or 'missing number (sequence completion) problems'. It is possible to imagine a situation in which the mathematics tests appeared to sample representatively across content and cognitive domains, at the level of granularity of the Ofqual study, whilst also appearing predictably restricted in the kinds of questions asked at a lower level of analysis.

[27] Sireci (1998, p. 303): "a socially desirable response set is one in which respondents try to portray themselves in a favorable light or respond to rating tasks in a way they think the investigator wants them to respond. This confound is a serious threat to the validity of item congruence and relevance ratings, because SMEs are well aware of the content areas the test is supposed to measure."

5.4 Conclusions

Our evidence suggested that STA's approach to sampling learning outcomes is robust. It compares favourably with approaches adopted for similar tests, internationally.

Although STA's approach is robust, this does not mean that it achieves perfect sampling. Just as there is no such thing as perfect assessment, there is no such thing as perfect domain sampling, especially as far as relatively short tests are concerned. The key stage 2 reading test is clearly limited in its coverage of the content of the national curriculum for reading; albeit for reasons that are well-understood and largely unavoidable.

STA's approach to weighting learning outcomes is necessarily interpretive, as the national curriculum does not specify weightings. While the approach adopted is plausible – particularly for mathematics – the extent to which it reflects the intentions of those who wrote the national curriculum is unknown; or, at least, undocumented. It is worth noting that the *Test Frameworks* were developed subsequently to, and independently of, the national curriculum framework documents; although the *Test Frameworks* were reviewed by members of the national curriculum review team during the development phase. For future revisions of the national curriculum and *Test Framework* documents, it would be worth considering whether it would be useful to develop a more formal model of collaboration between those responsible for specifying the curriculum and those responsible for specifying its assessment; to make certain that curriculum intentions, including high level aims, are faithfully translated.

STA's innovative approach to modelling the cognitive domain, for both reading and mathematics, is plausible. Our independent experts were able to appropriate STA's cognitive domain strands and to rate items from the 2016 tests according to these strands with a reasonable degree of consistency. However, despite the plausibility of this approach, prominent mathematics associations have argued that the design of the mathematics test may fail to do justice to the aims of the mathematics curriculum, which are framed in terms of fluency, mathematical reasoning, and problem solving. Although results from the present investigation concluded that the *Test Framework* documents appeared to translate national curriculum teaching requirements into plausible blueprints for testing – given STA's interpretation of the national curriculum framework document – that conclusion might not follow from an alternative interpretation. Representatives from mathematics associations have explained to Ofqual how the blueprints seem less plausible to them, given their own interpretation of the national curriculum Aims statement. Unfortunately, it is not

possible to arbitrate this issue on the basis of evidence from the present study. This important question remains open.

Although it is important to acknowledge that the purpose of *Test Framework* documents is to support consistent standards of test development over time – which means that it is important that they remain stable – it will also be important for STA to monitor their effectiveness and to continue to explore their conceptual underpinnings. This is particularly true for the cognitive domains, which are their most innovative feature, and therefore the most likely to require some refinement when revisions to the *Test Frameworks* are next made. On this point, it is worth returning to the distinction between demands and difficulty, discussed earlier. While the cognitive domain strands seemed, to our SMEs, to be plausible, further insight into their plausibility can be gained by modelling relationships between item demand ratings and item difficulty statistics empirically; for instance, using multivariate regression analyses to determine the variance in difficulty explained by each of the cognitive domain strands. Research of this kind will help to ensure that STA's approach to test development remains world class.

References

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. *Assessment in Education: Principles, Policy & Practice*, *14*, 201–232. http://doi.org/10.1080/09695940701478909

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, UK: Chapman and Hall.

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*, 243–248. http://doi.org/10.3758/BF03194059

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The classification of educational goals. Handbook I: Cognitive domain.* London: Longmans, Green and Co.

Clesham, R. (2013). *Good Assessment by Design: An International Comparative Analysis of Science and Mathematics Assessments*. London: Pearson.

Crisp, V., & Grayson, R. (2013). Modelling question difficulty in an A level physics examination. *Research Papers in Education*, *28*, 346–372. http://doi.org/10.1080/02671522.2012.673005

Cullinane, A., & Liston, M. (2016). Review of the Leaving Certificate biology examination papers (1999–2008) using Bloom's taxonomy – an investigation of the cognitive demands of the examination. *Irish Educational Studies*, *35*, 249–267. http://doi.org/10.1080/03323315.2016.1192480

Davis-Becker, S. L., & Buckendahl, C. W. (2013). A Proposed Framework for Evaluating Alignment Studies. *Educational Measurement: Issues and Practice*, *32*, 23–33. http://doi.org/10.1111/emip.12002

DfE. (2014). *National Curriculum in England Framework Document: December 2014*. London, UK: Department for Education. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal*, *28*, 59–82. http://doi.org/10.1080/09585176.2016.1232201

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice*, *30*, 3–15. http://doi.org/10.1111/j.1745-3992.2011.00218.x

Greatorex, J. (2013). *Context in Mathematics Examination Questions*. Cambridge: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/131388-context-in-mathematics-examination-questions.pdf

Greatorex, J., Shaw, S., Hodson, P., & Ireland, J. (2013). Using scales of cognitive demand in a validation study of Cambridge International A and AS level Economics. *Research Matters: A Cambridge Assessment Publication*, *15*, 29–37.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. http://doi.org/10.1348/000711006X126600

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education*, *20*, 101–126. http://doi.org/10.1207/s15324818ame2001_6

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*, 159–174. http://doi.org/10.2307/2529310

Little, G. (2016). Mathematics: conceptual understanding or counting by the rules? In NUT (Ed.), *The Mismeasurement of Learning: How tests are damaging children and primary education. Reclaiming Schools: The Evidence and the Arguments.* (pp. 24–25). London, UK: National Union of Teachers.

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012, April). *A framework for predicting item difficulty in reading tests*. Paper presented at American Educational Research Association (AERA) Meeting.

Pollitt, A., Ahmed, A., Baird, J.-A., Tognolini, J., & Davidson, M. (2008). *Improving the quality of GCSE assessment*. London, UK: Qualifications and Curriculum Authority.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 166–206). London, UK: Qualifications and Curriculum Authority.

Pollitt, A., Entwistle, N., Hutchinson, C., & de Luca, C. (1985). *What makes exam questions difficult?* Edinburgh, UK: Scottish Academic Press.

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H., & Bramley, T. (1998). *The effects of structure on the demands in GCSE and A Level questions*. London, UK: Qualifications and Curriculum Authority.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*, 299–321. http://doi.org/10.1207/s15326977ea0504_2

Sireci, S. G., Robin, F., Meara, K., Rogers, H. J., & Swaminathan, H. (2000). An External Evaluation of the 1996 Grade 8 NAEP Science Framework. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the Nation's Report Card: Research from the Evaluation of NAEP*. National Academic Press. Retrieved from http://www.nap.edu/catalog/9751.html

STA. (2015a). *English reading test framework: National curriculum tests from 2016*. Coventry, UK: Standards and Testing Agency. Retrieved from https://www.gov.uk/government/publications/key-stage-2-english-reading-test-

framework

STA. (2015b). *Mathematics test framework: National curriculum tests from 2016*. Coventry, UK: Standards and Testing Agency. Retrieved from https://www.gov.uk/government/publications/key-stage-2-mathematics-test-framework

Sweiry, E. (2013, October). *A Framework for the Qualitative Analysis of Examinee Responses to Improve Marking Reliability and Item and Mark Scheme Validity*. Paper presented at the 39th Annual Conference of the International Association for Educational Assessment. Tel Aviv.

Vappula, H., & Clausen-May, T. (2006). Context in maths test questions: Does it make a difference? *Research in Mathematics Education*, *8*, 99–115. http://doi.org/10.1080/14794800008520161

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346