Accessibility of the 2016 key stage 2 national curriculum reading test: review of evidence

October 2017

Ofqual/17/6286/2

Contents

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 would like to thank Colin Watson, Liz Twist, Rebecca Evison, Barbara Donahue and Stephen Goodman from the Standards and Testing Agency, for providing data on pupils' performances on the 2016 reading test, and for explaining details of the test development process. We would also like to thank those who contributed to this report by providing insights into issues related to the accessibility of the 2016 reading test, including Paul Clayton, Tracy Parvin, Alex Grady, Amanda Hipkiss, Sue Denny, Lucy Radford and Matthew Barnes.

Executive summary

The first papers from a new suite of key stage 2 tests were taken by pupils in May 2016. As soon as the reading test had been sat, teachers began to express concerns over its accessibility. They were particularly concerned that the test may have been unduly hard to access for pupils with lower levels of attainment, including those with Special Educational Needs and Disabilities (SEND).

The following review of evidence ‘piggybacks’ on a broader investigation into the approach to domain sampling for the new suite of national curriculum tests (Newton & Cuff, 2017). It is essentially an appendix to that investigation, although the two reports have been written up separately to ensure that neither their foci nor their aims should be confused. The report on the main investigation presents a comprehensive evaluation of an aspect of test development for the new suite of national curriculum tests, focused upon the approach to domain sampling. The focus of this subsidiary report is upon one particular test – the 2016 reading test – and concerns that were raised with its accessibility. In writing this report, our intention is to represent those concerns, to contextualise them in relation to additional information and analysis, and to consider issues arising. In other words, although this review of evidence was never conceived as a comprehensive evaluation in its own right, and was therefore not designed to reach definitive conclusions, we anticipate that its outcomes will be useful as a record of events, and also as a stimulus for further clarification, research and development.

The two aims of this subsidiary review were to:

- improve our understanding of teachers’ concerns over the accessibility of the 2016 reading test, including the variety of potential causes of those concerns; and
- consider whether there may be any questions for the Standards and Testing Agency concerning the potential for enhancing test development procedures for future years.

This report collates and summarises a wide body of evidence, which originated from a variety of sources, including:

- submissions to the recent Committee Inquiry into Primary Assessment;
- test and item functioning data provided by the Standards and Testing Agency;
- our broader investigation into the approach to domain sampling for the new suite of national curriculum tests (our ‘content validation’ study);
- independent research and reports on the 2016 reading test;
- comments published on social media; and

- focus groups with stakeholders, conducted by Ofqual.

The 2016 reading test was designed differently from previous years. It was intended to differentiate between a wider range of pupils, and designed to accommodate a higher ‘expected standard’ threshold. Therefore, it was likely to result in a somewhat different kind of experience. The evidence reviewed in this report concerns how pupils, their teachers and other stakeholders experienced the 2016 test, and how they reflected on that experience. In considering the collated evidence, it is important to appreciate that we have made good use of perceptions of inaccessibility, including reflections on possible causes of inaccessibility. Not all of the conclusions reached by teachers and stakeholders will necessarily have been entirely accurate: some may have been partially accurate; others may have been inaccurate. Having said that, they all constitute evidence of teachers’ and other stakeholders’ concerns, and therefore help us to understand the 2016 test experience.

One of the most challenging issues to identify and address is the threat of differential validity; in particular, features of assessment tasks that inadvertently inflate or deflate results for particular subgroups of pupils. This includes the possibility that features of texts or questions render an assessment task somewhat inaccessible for certain pupil subgroups, but very much more inaccessible for others. Many of the perceptions and reflections presented in this report raise concerns like these; for example, a focus group of primary teachers from Teesside suggested that it might have been relatively harder for pupils from lower socio-economic backgrounds to engage with certain of the texts; similarly, our SEND representatives suggested that it might have been relatively harder for pupils with certain kinds of SEND to deal with abrupt transitions in response strategy demands across questions. Again, although these teachers and stakeholders have raised important concerns, the present report is unable to substantiate or refute them. So their views should be understood as sources of relevant evidence alongside other sources of relevant evidence; eg alongside views expressed by teachers and SEND experts during the test development process, and alongside statistical evidence on test and item functioning produced as part of the test review process. It is important to acknowledge that not all of these sources of evidence are entirely consistent. For instance, some of the most extreme views aired on social media suggested that even high-attaining students would have been flummoxed by the test; whereas evidence from the mark distribution demonstrated that many high-attaining students performed extremely well. In fact, the distribution of marks was very Normal (in a statistical sense) with the ‘typical’ pupil scoring around half marks and with similar numbers of pupils scoring very high marks as very low marks. The following observations illustrate the issues that emerged from our review:

- although teachers were anticipating a higher ‘expected standard’ threshold, concerns began to be expressed before that standard had been set, rather than directly in response to it;
- concerns expressed on social media had different emphases. Some commentators were mainly critical of the texts; others were mainly critical of the questions. Some described the test as generally very hard; whilst others described the test as particularly hard for certain groups of pupils, including pupils with English as an Additional Language (EAL) and pupils with SEND.
- conclusions expressed more formally by teacher and subject associations echoed concerns expressed on social media, including the following perceptions and reflections:
    - certain contexts within certain texts may have been especially challenging for large numbers of pupils to engage with, eg the ‘antiquated’ feel of the first text
    - particularly difficult questions that occurred earlier in the test may have demoralised lower-attaining pupils and prevented them from demonstrating their true level of attainment on later questions
    - too many pupils failed to complete the test in the allocated time.
- test and item functioning data seemed to support some of these concerns, for instance:
    - although the first few questions were answered correctly by the vast majority of pupils, the hardest questions on both the first and second texts occurred mid-way through their respective question sets
    - perhaps relatedly, omit rates began to rise considerably towards the second half of the questions on the second text
    - nearly one in twenty pupils were deemed not to have reached the first question relating to the third text, and around a quarter of all pupils were deemed not to have reached the end of the test
- readability analyses suggested that the 2016 reading booklet was harder to read than both the sample booklet and the 2017 booklet. However, the 2016 reading booklet was not necessarily more complex than booklets from previous years. Indeed, according to one analysis, the range of reading complexity levels across the (three) texts from 2016 was similar to the ranges observed across the (five) texts (split across the two tests) from the years 2012 to 2015.
- according to outcomes from our content validation study:

  - items in the 2016 test reflected appropriate levels of demand – both according to STA ratings and according to those of a group of independent experts – as judged in relation to the *Test Framework* blueprint
  - however, in terms of the technical knowledge required strand, our independent experts rated items from the 2016 test somewhat higher than STA had
  - items 15 to 21 (the second half of the second question set) were predominantly either high in mean demand, or high in difficulty, or both; suggesting an abrupt transition in the middle of this question set from low demand/difficulty items to high demand/difficulty items.

- insights from SEND representatives highlighted a number of question, text, and question-text-interaction factors that might have proved particularly challenging for pupils with SEND (and perhaps also for other pupils too), for instance:
  - although the first few questions were answered correctly by the vast majority of pupils, they were still felt to be insufficiently accessible for pupils with SEND (and this impact might have been compounded by variation in cognitive strategies required to answer them)
  - the fact that all of the texts were presented as continuous passages may have been particularly challenging for certain groups, e.g. pupils with visual and auditory impairments and pupils with autism
  - the language and vocabulary of the 2016 test was perceived to be technically very challenging, particularly so for pupils with SEND
  - time pressure was also considered to present an unnecessary barrier.

On the balance of evidence presented, it seems plausible that the combined impact from multiple ostensibly negligible challenges – stemming from both question and text factors – may have rendered the 2016 reading test unduly hard to access for at least some pupils. Unfortunately, aggregate impacts of this sort can be difficult, if not impossible, to detect statistically and can be equally hard to identify via test reviews. Although accessibility issues may have prevented the test from measuring reading comprehension accurately for particular pupils, and perhaps also for certain groups of pupils, we do not have sufficient evidence to be able to reach any definitive conclusion concerning which pupils might have been affected in this way, nor how many pupils, nor to what extent.

The business of test design/development is not a precise science, and it always involves trade-off and compromise. However, test designers/developers need to be able to provide reassurance that their processes are sufficiently rigorous, and that design/development decisions strike appropriate balances. The experience of 2016 raises a number of important questions for STA, in particular:

1. **Are pupils given sufficient time to complete the reading test?** Although ‘reading at speed’ might be considered a legitimate part of the reading comprehension construct – an aspect of reading fluency – it is not immediately obvious what that ought to mean in terms of the percentage of pupils expected to reach the end of the test. Clarification on this issue would be useful.
2. **Are there ways in which test and item review processes can be made more rigorous?** The fact that STA did not foresee the intensity of concerns that teachers expressed with the 2016 reading test, nor certain of the particular concerns identified within the present report – despite having followed well-established processes which included consulting a teacher panel, a test review group, and an inclusion panel during the test development phase – raises the question of why, and of whether such concerns might have come to light had these panels been run differently, or if a different kind of review process had been adopted, or suchlike.
3. **Can alternative approaches be adopted to investigate potential biases more effectively for pupils from various groups (eg SEND, EAL, socio-economic)?** The potential for bias within reading tests can be harder to spot than for tests in many other subject areas. This is partly because demands related to reading load, task context, etc, are clearly construct-irrelevant in relation to these other subject areas, and their potential for causing construct-irrelevant score variance, ie bias, is fairly obvious. In other words, it is clear, in principle, how bias might occur; even though, in practice, it might be tricky to determine whether it actually has occurred. Conversely, for reading tests, demands related to reading load, task context, etc, are either construct-relevant, or less obviously construct-irrelevant, which makes their potential for causing bias far less obvious; even though bias is still a very real threat. Potential biases are also harder to spot for the key stage 2 reading test because it comprises sets of questions linked to common passages, rendering outcomes from DIF (Differential Item Functioning – see page 26) analyses harder to interpret. These issues raise the question of whether additional steps can be taken to address problems, like these, which are specific to the key stage 2 reading test.

1 Background

The first papers from a new suite of key stage 2 tests were taken by pupils in May 2016. As soon as the reading test had been sat, teachers began to express concerns over its accessibility. They were particularly concerned that the test may have been unduly hard to access for pupils with lower levels of attainment, including those with Special Educational Needs and Disabilities (SEND).

After these concerns had come to light, Ofqual discussed the matter with officials from the Standards and Testing Agency (STA), who explained that the reading test had been reviewed for fairness at various stages during the test development and trialling period. For instance, STA convened a series of two-day meetings involving a variety of panels, which reviewed the three texts and their related items. This included: (i) a teacher panel, made up of teachers from a range of school types from different geographical areas; (ii) a test review group, made up of teachers, head teachers and local authority representatives from across the country; and (iii) an Inclusion Panel, including SEND group representatives, inclusion service heads, head teachers experienced in working with pupils with SEND, educational psychologists, and others. In addition, STA's analysis of the technical 'functioning' of the test and of the individual questions and sub-questions ('items') that comprised it had provided no reason to conclude that the test was unduly hard to access for its target cohort of pupils. (STA shared this data with Ofqual and relevant analyses will be discussed in subsequent sections.)

Independently of these concerns, as part of our ongoing regulation of statutory assessments, Ofqual had set up an investigation into the approach adopted by the STA to developing the new suite of tests. Known as a 'content validation' study, it involved independent experts judging items from the key stage 2 reading and mathematics tests in terms of the subject content and thinking skills that they appeared to have been designed to assess. Our project began in 2016 and its research component was completed during February 2017 (Newton & Cuff, 2017).

Although not designed specifically to investigate accessibility issues, it seemed likely that our content validation study might be able to shed additional light on teachers' concerns with the 2016 reading test; given its focus on sampling both subject content and thinking skills. We therefore decided to invite representatives from stakeholder groups, particularly SEND groups, to review outcomes from our investigation, and to see if they were able to provide additional insights into teachers' concerns.

This report 'piggybacks' on our content validation study. We decided to write it up separately – as a review of evidence – so as not to be confused with, nor to distract from, the main report's important outcomes and conclusions. As such, it was not designed as a comprehensive evaluation in its own right, and there was no assumption that it would necessarily reach a definitive conclusion concerning the accessibility of the 2016 reading test. Having said that, the decision to prepare a separate review of evidence provided a good opportunity to collate and summarise a wide body of evidence related to those concerns.

The two aims of our subsidiary review were to:

- improve our understanding of teachers' concerns over the accessibility of the 2016 reading test, including the variety of potential causes of those concerns; and
- consider whether there may be any questions for the Standards and Testing Agency concerning the potential for enhancing test development procedures for future years.

We begin by explaining the structure of the new reading test and by describing reactions to the 2016 test administration. The main body of the report presents evidence from a variety of sources. Our report ends with a tentative conclusion concerning the accessibility of the reading test in 2016, and raises a number of questions for further consideration by STA.

2 The reading test

The key stage 2 reading test consists of a booklet containing three reading texts (12 sides of A4) plus a separate question and answer booklet (20 sides of A4). Pupils have a total of one hour to read the three texts and to respond to the questions. They are permitted to approach the test as they choose, eg working through one text and answering related questions before moving on to the next. The administration instructions for 2016 explained that the "least-demanding text will come first, with the following texts increasing in level of difficulty" (STA, 2016, p. 2).

Three versions of the new reading test are now in circulation: the sample test (published June 2015 [1]); the 2016 test (published May 2016 [2]); and the 2017 test (published May 2017 [3]). These tests all accommodate the full cohort of pupils (prior to 2016, there was a separate 'Level 6' test for high-attaining pupils). Tables 1 and 2 summarise key information on these tests, which helps to establish the context for the present report.

Table 1. Summary of information on the reading booklets and texts.[4]

Sample test	**2016 test**	**2017 test**
Space Tourism – non-fiction 767 words: 295+224+248	The Lost Queen – fiction 385 words	Gaby to the Rescue – fiction 601 words
Giants (poem) – fiction 171 words	Wild Ride – fiction 780 words	Swimming the English Channel – non-fiction 705 words: 531+96+78
The Lost World – fiction 725 words	The Way of the Dodo – non-fiction 641 words	An Encounter at Sea – fiction 623 words
1663 words in total	1806 words in total	1929 words in total

[1] https://www.gov.uk/government/publications/2016-key-stage-2-english-reading-sample-test-materials-mark-scheme-and-test-administration-instructions

[2] https://www.gov.uk/government/publications/key-stage-2-tests-2016-english-reading-test-materials

[3] https://www.gov.uk/government/publications/key-stage-2-tests-2017-english-reading-test-materials

[4] Two of the non-fiction texts were split into three sub-sections (hence X words = A+B+C).

All three versions share a common structure, with questions on three discrete texts (two fiction, one non-fiction). However, the texts within each booklet differ somewhat, both in terms of structure and in terms of genre. The non-fiction texts tend to be presented via discrete chunks of information; although this was not true for the 2016 Dodo text, which was presented as a continuous passage.

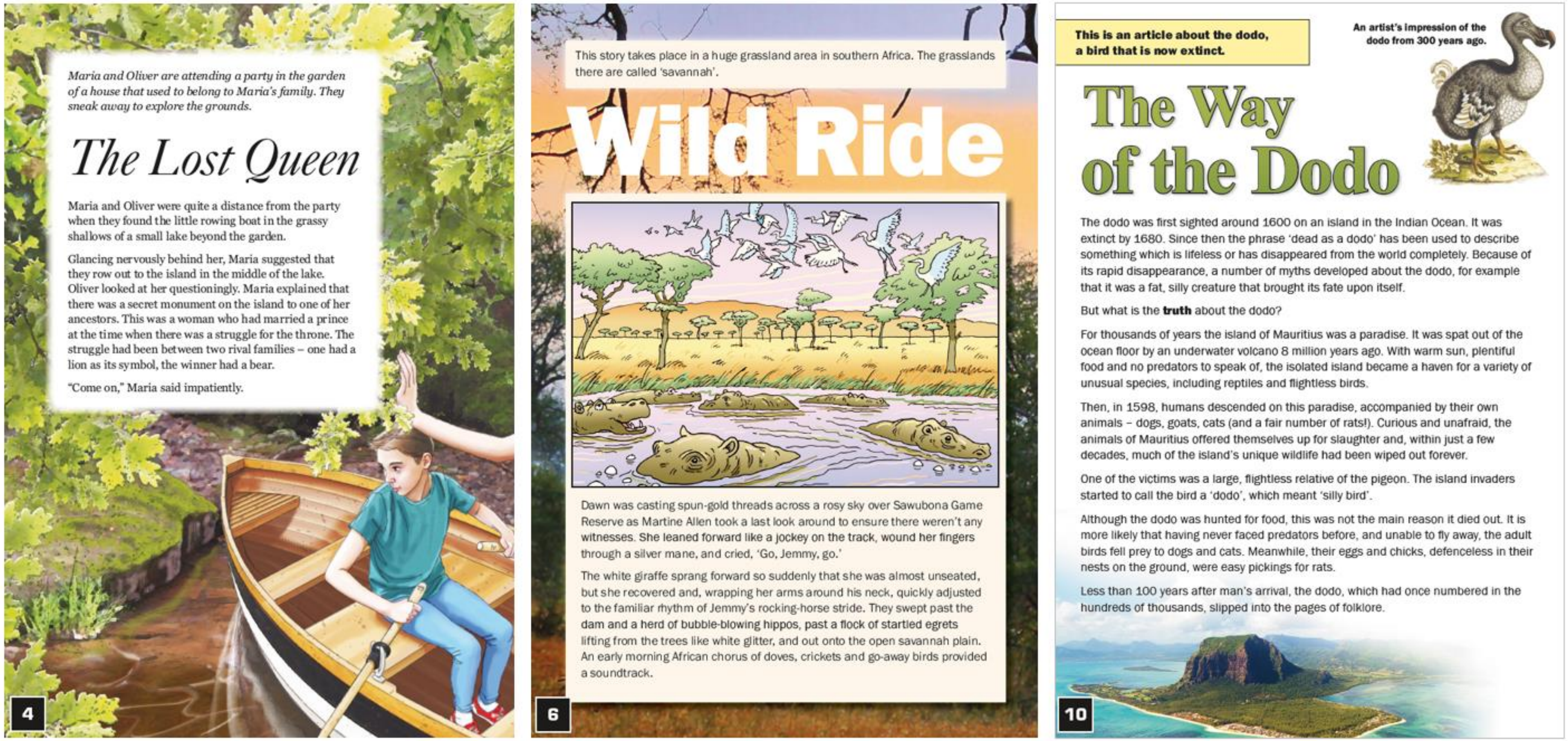
Table 2. Summary of information on the reading test items.

	1 mark	**2 marks**	**3 marks**
Sample			
Space Tourism	11	4	0
Giants	9	1	0
The Lost World	13	2	1
2016 test			
The Lost Queen	9	2	1
Wild Ride	10	3	2
The Way of the Dodo	12	0	0
2017 test			
Gaby to the Rescue	13	1	0
Swimming the English Channel	11	3	0
An Encounter at Sea	10	1	2

All three versions had at least seven multi-mark items: the sample and 2016 tests had eight; the 2017 test had seven. The sample test had a single 3-mark item, relating to the final text; while the 2017 test had two three-mark items, also relating to the final text. The 2016 test had three three-mark items, relating to the first and second texts.

Visually, the booklets are presented attractively and professionally. Figure 1 illustrates pages from the 2016 test. Figure 2, also from the 2016 test, illustrates the clarity of presentation of questions and response spaces.

Figure 1. Extracts from the 2016 test – reading booklet.



Maria and Oliver are attending a party in the garden of a house that used to belong to Maria's family. They sneak away to explore the grounds.

The Lost Queen

Maria and Oliver were quite a distance from the party when they found the little rowing boat in the grassy shallows of a small lake beyond the garden.

Glancing nervously behind her, Maria suggested that they row out to the island in the middle of the lake. Oliver looked at her questioningly. Maria explained that there was a secret monument on the island to one of her ancestors. This was a woman who had married a prince at the time when there was a struggle for the throne. The struggle had been between two rival families – one had a lion as its symbol, the winner had a bear.

"Come on," Maria said impatiently.

This story takes place in a huge grassland area in southern Africa. The grasslands there are called 'savannah'.

Wild Ride

Dawn was casting spun-gold threads across a rosy sky over Sawubona Game Reserve as Martine Allen took a last look around to ensure there weren't any witnesses. She leaned forward like a jockey on the track, wound her fingers through a silver mane, and cried, 'Go, Jemmy, go.'

The white giraffe sprang forward so suddenly that she was almost unseated, but she recovered and, wrapping her arms around his neck, quickly adjusted to the familiar rhythm of Jemmy's rocking-horse stride. They swept past the dam and a herd of bubble-blowing hippos, past a flock of startled egrets lifting from the trees like white glitter, and out onto the open savannah plain. An early morning African chorus of doves, crickets and go-away birds provided a soundtrack.

This is an article about the dodo, a bird that is now extinct.

An artist's impression of the dodo from 300 years ago.

The Way of the Dodo

The dodo was first sighted around 1600 on an island in the Indian Ocean. It was extinct by 1680. Since then the phrase 'dead as a dodo' has been used to describe something which is lifeless or has disappeared from the world completely. Because of its rapid disappearance, a number of myths developed about the dodo, for example that it was a fat, silly creature that brought its fate upon itself.

But what is the **truth** about the dodo?

For thousands of years the island of Mauritius was a paradise. It was spat out of the ocean floor by an underwater volcano 8 million years ago. With warm sun, plentiful food and no predators to speak of, the isolated island became a haven for a variety of unusual species, including reptiles and flightless birds.

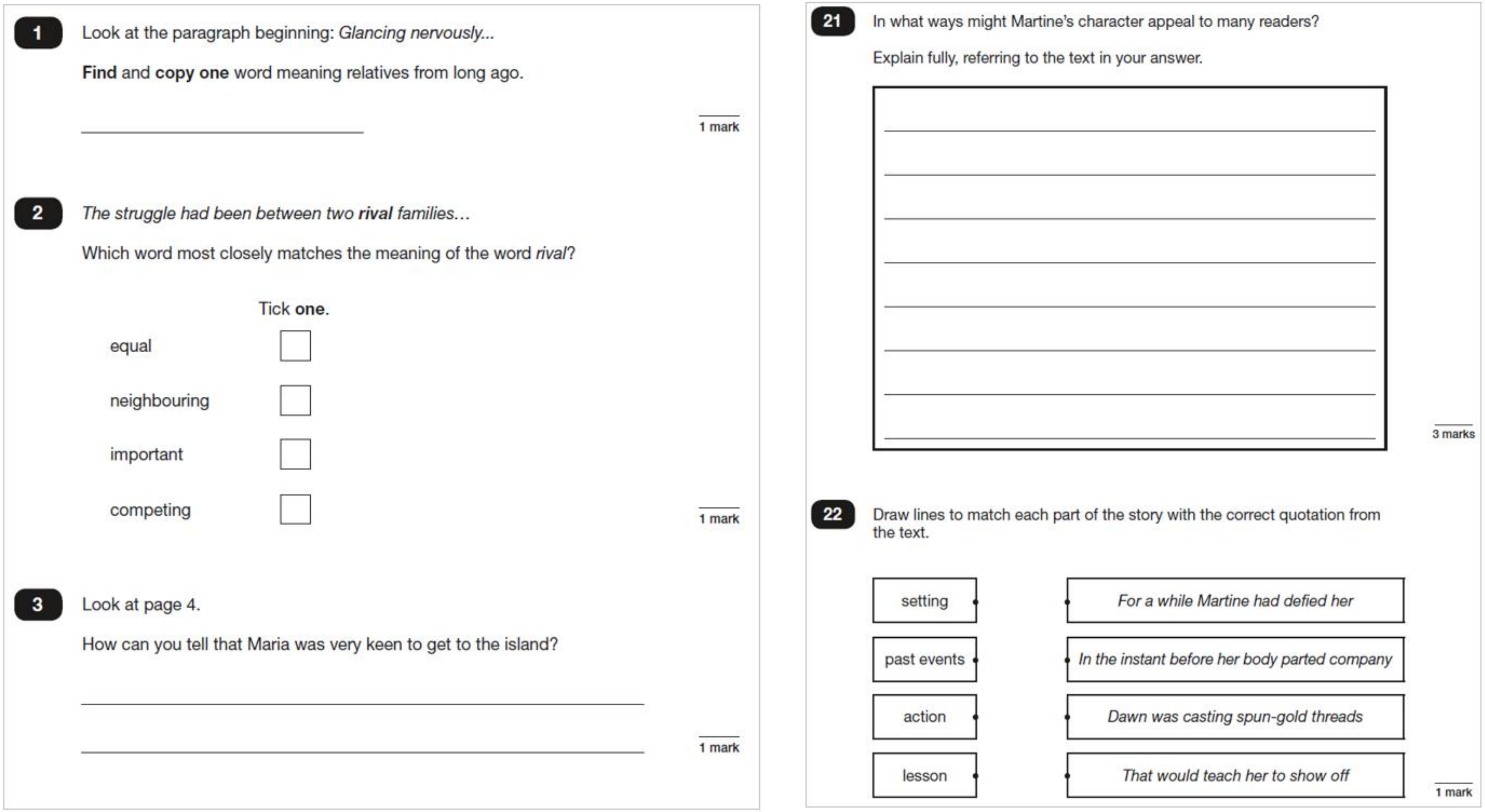
Then, in 1598, humans descended on this paradise, accompanied by their own animals – dogs, goats, cats (and a fair number of rats!). Curious and unafraid, the animals of Mauritius offered themselves up for slaughter and, within just a few decades, much of the island's unique wildlife had been wiped out forever.

One of the victims was a large, flightless relative of the pigeon. The island invaders started to call the bird a 'dodo', which meant 'silly bird'.

Although the dodo was hunted for food, this was not the main reason it died out. It is more likely that having never faced predators before, and unable to fly away, the adult birds fell prey to dogs and cats. Meanwhile, their eggs and chicks, defenceless in their nests on the ground, were easy pickings for rats.

Less than 100 years after man's arrival, the dodo, which had once numbered in the hundreds of thousands, slipped into the pages of folklore.

Figure 2. Extracts from the 2016 test – question and answer booklet.



1 Look at the paragraph beginning: *Glancing nervously...*

Find and **copy one** word meaning relatives from long ago.

____

1 mark

2 *The struggle had been between two* ***rival*** *families…*

Which word most closely matches the meaning of the word *rival*?

Tick **one**.

equal □

neighbouring □

important □

competing □

1 mark

3 Look at page 4.

How can you tell that Maria was very keen to get to the island?

____

____

1 mark

21 In what ways might Martine's character appeal to many readers?

Explain fully, referring to the text in your answer.

3 marks

22 Draw lines to match each part of the story with the correct quotation from the text.

setting	*For a while Martine had defied her*
past events	*In the instant before her body parted company*
action	*Dawn was casting spun-gold threads*
lesson	*That would teach her to show off*

1 mark

3 Reactions to the 2016 test

The 2016 reading test was taken by pupils in May 2016. It was already widely known that expected standards on both the mathematics test and the reading test would be set higher than in previous years. The level of attainment associated with the ‘expected standard’ threshold on the new tests would be higher than the level of attainment associated with the ‘Level 4’ threshold on the old tests. Additionally, the new reading test would have to target its cohort differently as there would no longer be a separate test for pupils with higher levels of attainment (the ‘Level 6’ test). However, the STA explained that this should not impact unduly upon pupils with lower levels of attainment, as the easiest questions would remain of a similar level of difficulty, when compared with previous years. Finally, teachers had been provided with sample materials for both mathematics and reading, to illustrate the format and likely content of the new tests.

Despite this preparation for the new suite of tests, including advanced warning of increased levels of challenge, many teachers were surprised at how hard to access the 2016 reading test seemed to be.

The following sections represent the views of teachers and stakeholders, helping us to understand the 2016 test experience as fully as possible. As noted earlier, we do not presume that all of these views are necessarily entirely accurate

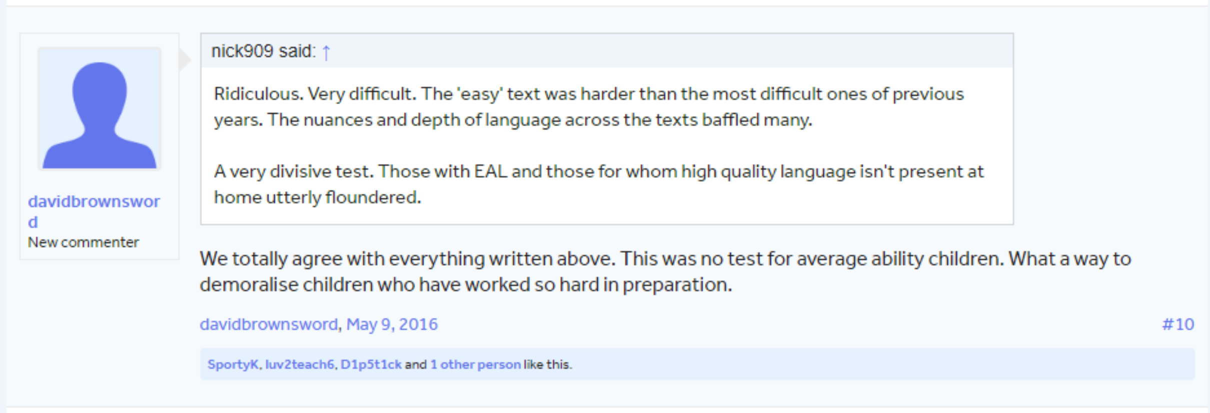
3.1 Reception by teachers

As soon as teachers had finished administering the 2016 reading test, comments began to appear on social media sites complaining about the level of difficulty of the test. Towards the end of the test day, Helen Ward (2016) summarised sentiments expressed in the TES Forum:

> Teachers report that even able children could not finish the test, and claim it was a 'demoralising' experience for some pupils
>
> Pupils have been left in tears by the first of this year's Sats, according to teachers who have branded the reading test “incredibly difficult”, “ridiculous” and “bloody tough”.
>
> The paper was taken by almost 600,000 pupils this morning. Teachers on the TES forums have reported that even able children were unable to finish the test.
>
> “The texts weren't so bad but the questions and the wording of them (vocabulary etc) was like something I have never seen before. I'm staggered,” said one teacher.
>
> “The questions were ridiculously hard from the start and I had a child in tears within five minutes, because in her words, 'I don't understand the questions'. This wasn't even a less able child,” another teacher commented.

Commentators on this forum expressed differing views on what might have made the test so challenging. Some focused more on the questions, while others focused more on the texts within the reading booklet, eg Figure 3.

Figure 3. A response to a post on the TES Forum, 9 May 2016



davidbrownsword
New commenter

nick909 said: ↑

Ridiculous. Very difficult. The 'easy' text was harder than the most difficult ones of previous years. The nuances and depth of language across the texts baffled many.

A very divisive test. Those with EAL and those for whom high quality language isn't present at home utterly floundered.

We totally agree with everything written above. This was no test for average ability children. What a way to demoralise children who have worked so hard in preparation.

davidbrownsword, May 9, 2016 #10

SportyK, luv2teach6, D1p5t1ck and 1 other person like this.

Although a number of teachers expressed similar concerns over the accessibility of the mathematics test, they were in a small minority. Many more teachers expressed relief that the mathematics test was not as challenging as the reading one.

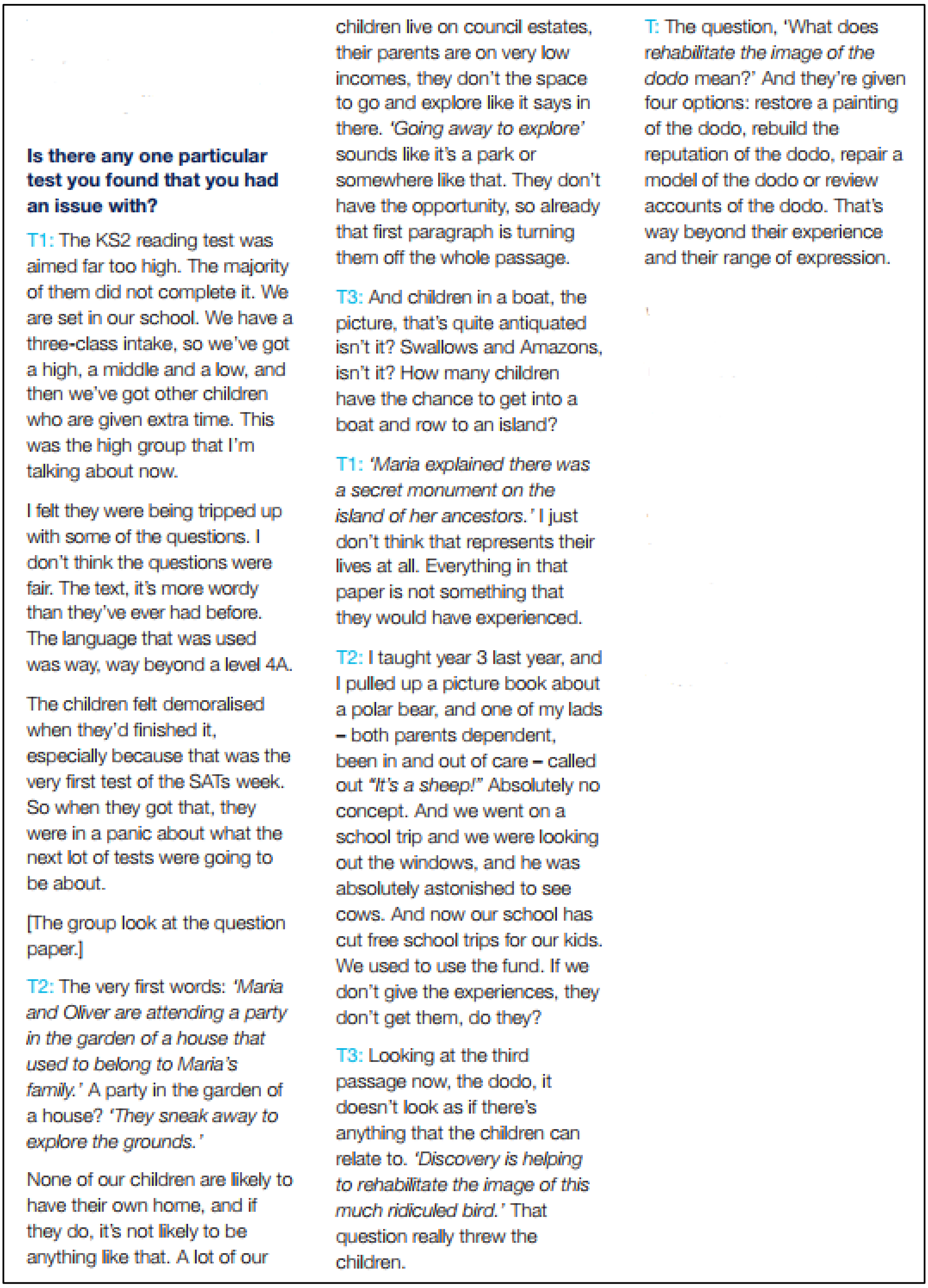
It is important to note that concerns over the accessibility of the reading test related to the test experience; not to the new, more challenging, test standard, which had not yet been set.[5] Indeed, although a more challenging standard was also to be set on the mathematics test, teachers did not express similar concerns over its accessibility

[5] The new standard was (subsequently) set by deciding whereabouts on the mark scale to locate the ‘expected standard’ threshold. A threshold mark is the lowest mark (total) at which pupils can reasonably be said to have met a particular standard. Pupils who score at or above the threshold mark on the new test are classified as having met the expected standard; whereas pupils who score below it are classified as not having met it. If the questions that comprised a particular test turned out to be more difficult than anticipated, then even those who would have attained the threshold standard would (on average) score lower than anticipated, and the threshold mark would need to be set lower than anticipated to accommodate this fact

3.2 Insights from an NUT report

The National Union of Teachers published a wide-ranging critique of testing, in November 2016, entitled *The Mismeasure of Learning* (NUT, 2016).

Figure 4. Transcript from a focus group discussion (extract from NUT, 2016, p.23)



Is there any one particular test you found that you had an issue with?

T1: The KS2 reading test was aimed far too high. The majority of them did not complete it. We are set in our school. We have a three-class intake, so we've got a high, a middle and a low, and then we've got other children who are given extra time. This was the high group that I'm talking about now.

I felt they were being tripped up with some of the questions. I don't think the questions were fair. The text, it's more wordy than they've ever had before. The language that was used was way, way beyond a level 4A.

The children felt demoralised when they'd finished it, especially because that was the very first test of the SATs week. So when they got that, they were in a panic about what the next lot of tests were going to be about.

[The group look at the question paper.]

T2: The very first words: *'Maria and Oliver are attending a party in the garden of a house that used to belong to Maria's family.'* A party in the garden of a house? *'They sneak away to explore the grounds.'*

None of our children are likely to have their own home, and if they do, it's not likely to be anything like that. A lot of our children live on council estates, their parents are on very low incomes, they don't the space to go and explore like it says in there. *'Going away to explore'* sounds like it's a park or somewhere like that. They don't have the opportunity, so already that first paragraph is turning them off the whole passage.

T3: And children in a boat, the picture, that's quite antiquated isn't it? Swallows and Amazons, isn't it? How many children have the chance to get into a boat and row to an island?

T1: *'Maria explained there was a secret monument on the island of her ancestors.'* I just don't think that represents their lives at all. Everything in that paper is not something that they would have experienced.

T2: I taught year 3 last year, and I pulled up a picture book about a polar bear, and one of my lads – both parents dependent, been in and out of care – called out *"It's a sheep!"* Absolutely no concept. And we went on a school trip and we were looking out the windows, and he was absolutely astonished to see cows. And now our school has cut free school trips for our kids. We used to use the fund. If we don't give the experiences, they don't get them, do they?

T3: Looking at the third passage now, the dodo, it doesn't look as if there's anything that the children can relate to. *'Discovery is helping to rehabilitate the image of this much ridiculed bird.'* That question really threw the children.

T: The question, 'What does *rehabilitate the image of the dodo* mean?' And they're given four options: restore a painting of the dodo, rebuild the reputation of the dodo, repair a model of the dodo or review accounts of the dodo. That's way beyond their experience and their range of expression.

Figure 4 reproduces an extract from this report. This is a transcript from a focus group discussion with Teesside primary teachers, which provides insights into their experiences of the 2016 reading test.

3.3 Inquiry into primary assessment

On 23 September 2016, the House of Commons Committee on Education launched an inquiry into primary assessment.[6] This was not a direct response to concerns over the reading test, as it had a far broader remit, to scrutinise reforms to primary assessment and their impact on teaching and learning in primary schools. Various submissions to the Inquiry did, however, specifically reference concerns over the accessibility of the reading test.

From the Association of School and College Leaders (ASCL) submission:

> Children found the new reading test this year particularly challenging. Many found the content difficult to access, and the introduction of 'harder' questions early in the test led many of them to become demotivated and unable to fully demonstrate their ability.

From the Association of Teachers and Lecturers (ATL) submission:

> Our members have expressed serious concerns about the design of the new key stage 2 reading test, and given that these tests undergo 3 years of development and trials we are concerned that this test slipped through the net. The low pass mark and the unusual distribution of marks suggests that this test was significantly harder than the other assessments. Furthermore, we question whether it is necessary to have a test that many children are not expected to complete in the set time. There is a lack of confidence in this specific test and we have yet to receive adequate assurance that the 2017 test will be more accessible.

From the National Association of Head Teachers (NAHT) submission:

> A survey was sent to NAHT members working as school leaders in primary schools in June 2016 and this received 2,628 responses. […]
>
> At KS2, 98% of respondents felt that the KS2 English paper was more difficult than expected to some degree […]
>
> An overwhelming majority of respondents (98%) reported that tests at KS2 were not appropriate for children with SEN […]

[6] https://www.parliament.uk/business/committees/committees-a-z/commons-select/education-committee/inquiries/parliament-2015/primary-assessment-16-17/

> The reading test needs to be redesigned. The accessibility of the tests for all pupils, including those with SEND, must be improved in order that they can effectively demonstrate their progress.

From the United Kingdom Literacy Association (UKLA) submission:

> The reading SAT was deliberately planned to be too long for many children to finish in order to distinguish between faster and slower readers. There is a misconception here, since effective reading is not a matter of speed but of understanding and response. A ‘slower’ reader may be a more effective reader [than] a ‘faster’ reader. The reading SAT test should be reframed so that it genuinely tests reading capability rather than speed.

Concerning the reading test, the Select Committee report (HCEC, 2017) concluded:

> 23. The level of difficulty was discussed through written evidence many times, with some teachers commenting on its inaccessibility to pupils with special educational needs and disability (SEND) or who are working at a lower ability.[21] Michael Tidd, a deputy headteacher, told us:
>
> > The reading test particularly, this year was virtually inaccessible for a good chunk of children who are not perhaps designated as having special needs but who are also not yet at the new expected standard.[22]
>
> 24. It was felt that the test had not been thoroughly tested with pupils and teachers. However, when we raised this issue with Claire Burton, STA, she assured us that the test went through a thorough development process:
>
> > It was scrutinised by teachers, inclusion experts, it had been sat and trialled in schools beforehand, and broadly the test did perform as we expected it to. It had sufficient marks at the lower end of the scale that we were able differentiate pupils there. It also included that higher-level content, so we were able to look at the pupils who had previously perhaps been performing at that level 6 test that had been removed. It did all of those things.[23]

4 Additional evidence

The following sections present additional evidence:

1. on test and item functioning data (provided by STA)
2. on the readability of the reading booklet
3. from our content validation study
4. from SEND representatives' reflections on the test
5. on pupils' enjoyment of the texts (provided by STA).

4.1 Test and item functioning data

As noted earlier, STA shared with us data on the technical ‘functioning’ of the test and of the individual questions and sub-questions (‘items’) that comprised it. Figure 5 presents data on the facility (ie easiness) of each item within the reading test, based on data from the full cohort of pupils who attempted the test in May 2016. The vertical blue lines separate items relating to each of the three texts. The facility index is the average mark awarded to responses to a particular item, expressed as a percentage of that item’s maximum mark. For one-mark items, this is equivalent to the percentage of pupils who answered the question correctly. The higher the facility, the easier pupils found the question.

Figure 5. Item facility indices (whole cohort data)

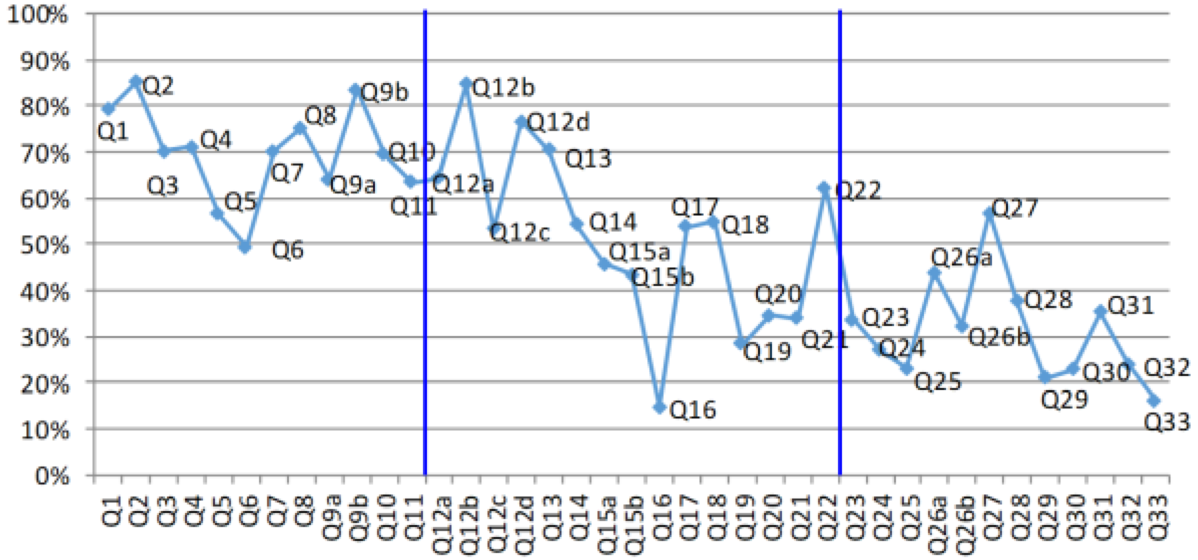


Figure 5 indicates that the test had a balance of easier and harder items. Its mean facility index of 51%, and median index of 54%, suggested that questions of middling difficulty were answered correctly by around half of all pupils. This kind of pattern is consistent with an effectively functioning test.

The test design model specified that the least-demanding text ought to come first, with the following texts increasing in level of difficulty. The observed facility indices broadly supported this claim; tending to be higher for the first text than for the second, and tending to be higher for the second text than for the third. In theory, it would be desirable also to see a degree of difficulty ‘ramping’ across the items that relate to each text, indicating that the easier questions on each text preceded the harder ones. In practice, this can be challenging to achieve; particularly when test developers are also attempting to order the questions at least roughly in relation to the chronology of the text, to make the test more accessible from a different perspective. STA explained that test construction is particularly challenging for the reading test, and provided us with a detailed explanation of why, which is reproduced below in Box 1 (presented after Figure 7).

From Figure 5, it seems that STA had some success with intra-text difficulty ramping, although there were notable exceptions. For instance, the hardest questions on both the first and second texts occurred mid-way through (Q6 – 49% and Q16 – 15% respectively). Similarly, the easiest question on the third text occurred mid-way through (Q27 – 57%). Again, though, intra-text ramping is hard to achieve for a reading test, and the observed patterns are not unreasonable.

Figure 6. Item omission rates (whole cohort data)[7]

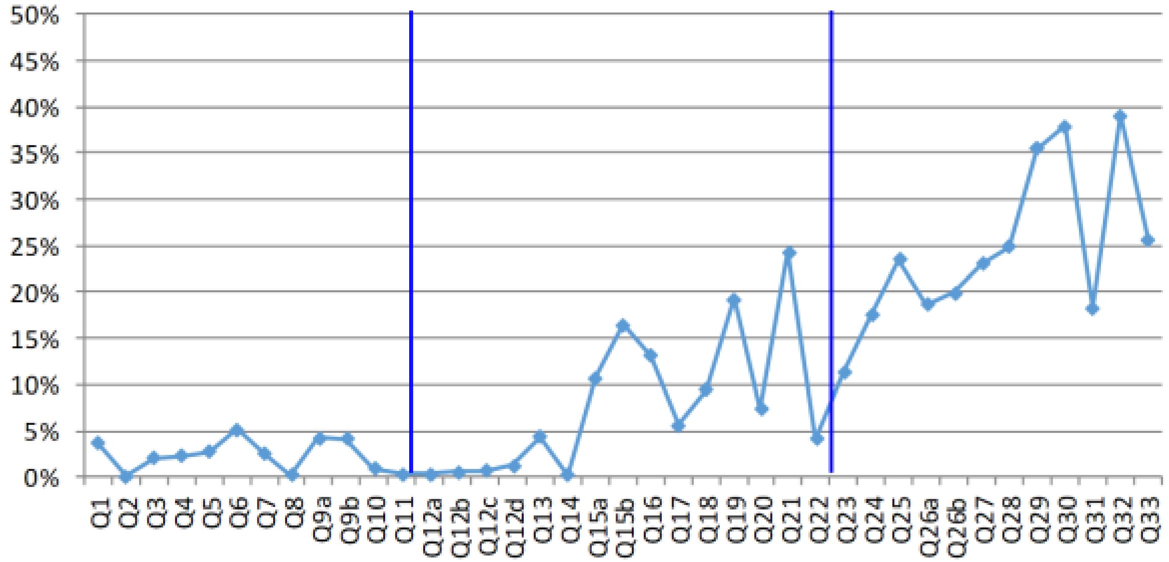


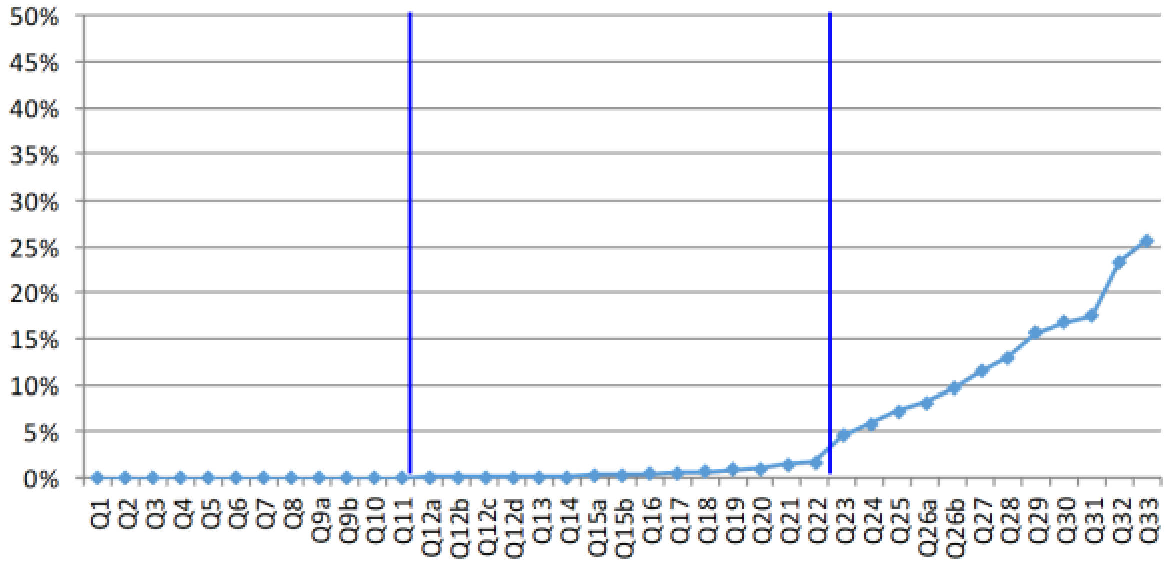
Figure 6 presents a different kind of data, based on item omission rates, ie the percentage of pupils who failed to record any response for an item. Omit rates are particularly useful for identifying inaccessible questions, ie questions that pupils felt incapable of even attempting to respond to. For instance, while questions 12c and 18, from the second text, had roughly equivalent facility indices – having been answered correctly by 53% and 55% of all pupils, respectively – nearly 10% of pupils failed to attempt question 18 in comparison with fewer than 1% of pupils who failed to attempt question 12c. Sometimes, differences like these are likely to be a function of the type of response required. For instance, question 12c required pupils to circle the

[7] Note that Figures 6 and 7 are presented differently from Figure 5, as their scales have been truncated to a maximum of 50%.

correct response (ie to select it from a list); whereas question 18 required pupils to produce a correct response (ie to construct it from scratch).

Figure 7 presents related data concerning the percentage of pupils who made it to a certain point in the test before giving up or running out of time. ‘Not reached’ rates relate to the question that follows the last question in the answer booklet for which a pupil provided any kind of response, ie to the first question that is deemed not to have been reached. Figure 7 suggests that the first noteworthy jump in pupils giving up or running out of time coincided with the first question of the third text, which had a not reached rate of 4.56% (compared with 1.68% for the last question of the second text). In other words, nearly one in twenty pupils were deemed not to have reached the first question relating to the third text. The penultimate and final questions of the third text had not reached rates of 23.52% and 25.64%, respectively; meaning that around a quarter of all pupils were deemed not to have reached the end of the test.

Figure 7. Item not reached rates (whole cohort data)



Box 1. Constraints on constructing the reading test (provided by STA)

When constructing any test, there are a number of constraints that affect the items that are selected. These are set out in the test framework. For example, there is often a limit on the number of 3-mark items or on the proportions of open versus selected response items. However, with reading tests, there are additional constraints because of the use of reading texts and the interaction between texts and items.

Texts

In order for a test to be balanced, there needs to be consideration of a number of factors:

- the different text types that could be included;
- the length of the texts and therefore the test overall;
- the reading demand of the texts;
- the accessibility of the texts for particular sub-groups of the population; and
- the content of the text in terms of its suitability for the age range and any sensitivities for any cultural or religious beliefs.

Texts are selected and placed into packs to balance all of these issues. These packs are then trialled together and only if all texts within a pack function appropriately does the pack move through the test development process. If a text fails, then a new pack is formed and the pack is re-trialled before moving to the next stage of the process.

Interaction of texts and items

As items are linked to a text, they are not fully independent. As a result, it is necessary to ensure that items do not attempt to assess the same element of the text, give away answers to other items in the test or enable the pupil to give the same response to different items within the test.

It is best practice to provide items on as much of the text as possible, so that children are not unnecessarily reading whole paragraphs without any associated items. It is also always important to ensure that the items assess the key parts of the text (eg the turning point of the story, essential information) rather than trivial details.

Test construction

All of these factors together increase the challenge in constructing a reading test over those for other tests. In order to address this, STA commissions more texts and items at the initial item writing stage to enable a range of items to be available for test construction. Texts and items are refined through the three-year development process, and items and texts that do not function appropriately or are identified as not being appropriate for pupils at the end of key stage 2 are archived through this process.

The archiving of items will reduce the choice of items available during test construction and limit the options to ensure an appropriate test that meets the test framework and addresses the constraints above.

Finally, STA supplied us with data from the Technical Pre-Test (TPT) of the 2016 reading test, which is essentially a ‘pilot’ administration, conducted a year in advance of the live administration with a sizeable sample of pupils. The main function of the TPT is to help link standards on one version of a test to another, via equating techniques. However, it also provides additional insights into test and item functioning, eg whether particular items might be unusually easy or hard for particular groups of pupils. This analysis is known as Differential Item Functioning (DIF) and is typically conducted for fairly straightforward comparisons, for which sufficient numbers of pupils are likely to fall into each of the comparison groups; for instance, boys versus girls, or pupils with English as an Additional Language (EAL) versus non-EAL pupils. When DIF is detected for a particular item, it means that pupils from one of the groups performed substantially better or worse than would be expected on that item, in comparison with how they performed on the rest of the test. This could happen if a question included a word that happened to be unduly hard to access for pupils from one of the groups; or if, for instance, the context in which a question was set was unduly easy to access for pupils from one of the groups.

These analyses can be quite sensitive, such that differences may be observed that are ‘significant’ in a strict statistical sense, but that are small enough to be considered negligible. However, when a large difference occurs, this indicates a potentially problematic (ie biased) item. A high-DIF item might need to be excluded from the test, unless it is possible to argue persuasively that an ‘unexpected’ difference in performance between the groups would actually be expected on an item like this.[8]

Appendix 1 illustrates that no high-DIF items were identified for the 2016 reading test, during the TPT, when the comparison groups were: boys versus girls; and EAL versus non-EAL. Having said that, DIF works by comparing how groups perform on particular items with how they perform on the test as a whole. In other words, if the test as a whole happened to be biased against a certain group, then this would not be identified via the DIF statistic. Similarly, if a particular text happened to be particularly hard to access for a certain group of pupils, then it is quite possible that this would have a performance-inhibiting impact across all of the items relating to that text. And if that happened to be the first text, then its impact might ripple out to performance on the remaining texts/items, too. This lack of independence between item responses means that the DIF statistic is potentially less capable of identifying bias within the key stage 2 reading test than within other tests.

[8] In that case, it might also be appropriate to revisit the prior assumption that a question like this taps into an important element of the curriculum and, therefore, that it ought to be included in the test.

4.2 Readability of the reading booklet

Teachers raised concerns with the readability of the texts within the 2016 reading booklet. It was suggested that the 2016 texts were less readable than the sample texts and, subsequently, the 2017 texts. A few teachers examined this empirically, by feeding these texts into readability engines (Parker, 2017).

We replicated these informal analyses, copying texts from the three versions of the reading booklets into MS Office Word, and using this software to produce readability statistics.[9]

Table 3. Readability statistics for the three versions of the reading booklet

	Sample	2016	2017
Sentences per paragraph	2.3	2.7	3.2
Words per sentence	13.5	15.9	12.8
Characters per word	4.3	4.4	4.3
Flesch Reading Ease	77.5	68.0	78.7
Flesch-Kincaid Grade Level	5.7	7.7	5.3
Passive Sentences	7.1%	14.2%	7.4%

The intention underlying this analysis was simply to replicate, in a consistent manner, the kinds of analyses that have been reported via social media and in the educational press. We do not have an opinion on whether the particular statistics reported in Table 3 are necessarily the *most* suitable for judging the readability of a reading test booklet, but they should be sufficiently reliable to at least give an indication of readability levels.[10] These informal analyses suggested that the 2016 booklet was significantly harder to read than either the sample booklet or the 2017 booklet.

[9] This required a few fairly arbitrary, but also fairly trivial, decisions concerning whether to include absolutely all of the words in the booklets. For instance, it was decided to include all words that a pupil would be likely to read; including headings on the front page and instructions concerning blank pages. Similarly, it was decided to exclude any word that a pupil would be unlikely to read; including credits on the back page.

[10] Flesch statistics take into account word and sentence length, and the types of words being used (see Flesch, 1948); they appear to be less able to take into account the comprehensibility of the text (see Ofqual, 2012, p.19). Although statistics, such as these, do not tell the whole story about readability – see Janan and Wray (2012) for a useful overview – comparisons across texts, using the

Further insight into the readability issue comes from a study conducted by MetaMetrics® (Sandford-Moore, Koons and Bush, 2016), which compared the complexity of the 2016 reading booklet with the complexity of the sample booklet and with the complexity of previous booklets. In previous years, there were two test booklets: one for the Level 3-5 test; and one for the Level 6 test. MetaMetrics® measured complexity using The Lexile® Framework for Reading and the Lexile Analyzer®. The Lexile scale ranges from 0L and below for 'early reader' passages to above 1600L for 'advanced reader' passages. Table 4 reproduces data from this study, relating to key stage 2 reading booklets from 2012 to 2016. For each of the reading booklets, two outputs are presented:

1. a range of Lexile measures, representing each of the texts separately (3 for the new test and for the previous Level 3-5 test, 2 for the previous Level 6 test)
2. a single Lexile measure, representing the booklet overall.

Table 4. Reading complexity (in Lexiles) for booklets ❖ and texts ◆

	2012	2013	2014	2015	2016
L 3-5	930-1080◆	780-1060◆	820-1090◆	870-1120◆	
	990❖	990❖	1020❖	1040❖	
L 6	1060-1230◆	770-990◆	1040-1080◆	1060-1190◆	
	1100❖	840❖	1060❖	1110❖	
All pupils Sample					860-990◆
					910❖
All pupils Live test					880-1160◆
					1110❖

In terms of the overall complexity measure, Table 4 indicates that the sample test booklet (910L) was less complex than the Level 3-5 booklets from each of the previous four years (2012 to 2015); whereas the 2016 live test booklet (1110L) was more complex than each of those previous four booklets. In fact, the 2016 test

same statistic, can still provide useful insights into differences between those texts, related to readability.

booklet was similar in complexity to the Level 6 booklets over the same period (which were targeted at pupils with higher levels of attainment than the Level 3-5 booklets). Having said that, the complexity range across the 3 texts of the 2016 test booklet (880L to 1160L) was very similar to the complexity range across the 5 texts for each of the previous four years, when the Level 3-5 and Level 6 booklets are combined (e.g. 870L to 1190L for 2015). In other words, although the most complex text in 2016 test was substantially more complex than the most complex text in the sample test, it was not substantially more complex than the most complex text in the 2015 test (even when considering only the Level 3-5 booklet). Furthermore, the range of text complexity levels within the 2016 test (which had to accommodate all test-taking pupils from the cohort) was similar to the range of text complexity levels observed from 2012 to 2015 (when combining all Level 3-5 and Level 6 texts within each year).

4.3 Content validation study

In February, 2017, we conducted a content validation study, in order 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 STA's *Test Framework* documents and, in turn, to the national curriculum framework document; 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 Frameworks* and in the national curriculum.

STA's approach to sampling thinking skills seemed to be relevant to concerns that had been raised with the 2016 reading test, because it focuses on the demands that questions make of pupils. STA specified this approach in the *Test Framework* for reading, within a blueprint for sampling from the cognitive domain.[11] This blueprint was specified both in terms of: (i) types of thinking skills necessary for responding to the test questions – known as domain strands; and (ii) levels of each of these different types of thinking skills that would be required to respond correctly to the test questions – known as the strand ratings. STA identified five cognitive domain strands:

1. Accessibility of the target information. This means: the number and proximity of features that need to be located in the text; the extent to which the location of the information within the text is identified in the question; the extent to which competing information in the text and / or distractors may mistakenly be selected. It can be thought of as, 'Where can the information be found?'

2. Complexity of the target information. This means: the lexico-grammatical density of the stimulus; the level of concreteness / abstractness of the target information; the level of familiarity of the information needed to answer the question. It can be thought of as, 'What is the language of the text like?'

3. Task-specific complexity. This means: the degree of cognitive complexity involved in answering the question, from retrieval through to inference and higher-level skills. It can be thought of as, 'How much work is needed to answer the question?'

[11] The purpose of a test blueprint is to ensure that each new version of a test contains the right balance of topics and thinking skills. The balance of topics assessed in each test is specified by the 'content domain' blueprint; while the balance of thinking skills assessed in each test is specified by the 'cognitive domain' blueprint.

4. Response strategy. This means: the complexity of the written response required; the extent to which pupils need to organise / structure their response. It can be thought of as, ‘How easy is it to organise and present the answer?’

5. Technical knowledge required. This means: the extent of knowledge of vocabulary required by the question and the text; the subject-specific technical language, and knowledge required that is not given in text. It can be thought of as, ‘How complex is the language of the question and / or the knowledge needed to answer it?’

STA provided a level rating for each reading test item from the 2016 test, for each of the five domain strands, on a scale from 1 to 4 (informed by their own experts’ judgements). This was intended to indicate the level of the thinking skill that would be required to provide a (correct) response to the item; that is, the item’s level of demand on the relevant domain strand. A rating of 1 indicated a low level of demand, whereas a rating of 4 indicated a high level of demand. The sampling blueprint identified the proportion of the test that would be allocated to items of differing levels of demand across the strands.

In addition to investigating whether the 2016 reading test sampled from the content and cognitive domain in the manner specified by STA’s *Test Framework*, we also investigated the extent to which an independent group of experts agreed with STA’s classification of each item from the reading test, both in terms of the content domain and in terms of the cognitive domain. Results from our study are presented and discussed in a separate report (Newton & Cuff, 2017).

Presented below are a number of findings from the content validation study that seemed to be of particular relevance to the issue of accessibility. Indeed, we explicitly raised the accessibility issue with our independent experts, during focus group discussions, and they offered their own reflections on this issue in the light of their experiences rating test items.

4.3.1 Sampling the cognitive domain

For the first three of the five cognitive domain strands, the *Test Framework* for reading specified that questions ought to target thinking skills across the full range of levels (ie 1 to 4), although predominantly at levels 2 to 4. In accordance with the cognitive domain blueprint, all of the items in the 2016 reading test were predominantly rated (by STA) as 2-4, for these three strands. While ratings for accessibility of target information and task-specific complexity spanned the full range of levels, in accordance with the blueprint, the range for complexity of target information only spanned levels 1-3 (again, as rated by STA), which represents a slight departure from the blueprint. For the last two of the five cognitive domain strands – response strategy and technical knowledge required – the ranges of levels identified were within those specified in the *Test Framework*. Overall, then, the items in the 2016 test reflected appropriate levels of demand, according to STA ratings, as judged in relation to the *Test Framework* blueprint.

A key component of our content validation study involved a group of subject experts independently rating each item from the 2016 test in terms of the same cognitive domain strands. Our experts were asked to provide five ratings for each item, ie for each of the five strands, to represent their own intuitive impression of whether the level of demand posed by the item seemed to be high (3 or 4) or low (1 or 2). So, alongside STA's ratings for each item, we were able to consider our experts' ratings (using the average, across all experts, for each rating). So, to what extent did the items in the 2016 test reflect appropriate levels of demand (ie ranges corresponding to the test blueprint) when judged in terms of the mean item-level ratings provided by our experts?

Across the items that comprised the 2016 test, mean expert ratings of accessibility of target information demands spanned the full range of rating levels, 1-4, in accordance with the blueprint; although mean expert ratings of complexity of target information and task-specific complexity demands only spanned rating levels 2-4. As for the STA ratings, all of the items in the 2016 reading test were predominantly rated (by our experts) as 2-4, for these three strands, in accordance with the cognitive domain blueprint. Mean expert ratings for response strategy were also within the guidelines specified by the *Test Framework*.

Table 5. Numbers of items/marks rated at each level by our experts (TKR strand)

	Number of items		**Number of marks**	
	STA	**experts**	**STA**	**experts**
Scale point 1	14	0	16	0
Scale point 2	16	21	22	24
Scale point 3	9	17	12	25
Scale point 4	0	1	0	1

For the technical knowledge required (TKR) strand, levels of demand, according to our experts' ratings, also remained within the parameters established by the *Test Framework*, albeit only just so. According to the *Test Framework*, the majority of TKR demands are meant to be at level 1 or level 2. According to our experts' mean ratings, none of the items were classified at the lowest level, and considerably more were classified at level 3 (see Table 5). The differences between STA's ratings and those of our experts were exacerbated when considered in terms of available marks.

Although these findings related to the TKR strand are interesting, and potentially relevant to the issue of accessibility, it is important to remember that we did not ask our experts to attempt to replicate the STA's approach to rating items, as specified in the *Test Framework*. Our experts' judgements reflected a more intuitive impression of the level of demand associated with each item, with no attempt to standardise those judgements across experts.

During the focus group discussions, some experts also commented on the high demands for the technical knowledge required strand:

> "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- to 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"

4.3.2 Item ordering

Several experts raised concerns about the order in which items were presented within the test. Specifically, there was a 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, we created a crude proxy indicator of overall item demand, by averaging demand ratings across the five cognitive domain strands; first using STA ratings and then using mean ratings from our experts.

The idea of specifying the cognitive domain in terms of different domain strands is to identify the different kinds of demands that questions make of pupils; that is, the different thinking skills that pupils need to apply in order to answer each question correctly. The greater the demands made by a particular question, the more difficult it will be to answer that question correctly. The overall level of demand, for any particular question, ought therefore to be related to its difficulty, ie to the proportion of pupils who answer the question correctly.[12]

It is important to note that our computation of mean demand for each item provides only a limited indication of its overall level of demand, because there are no guarantees concerning how different demands will interact. But it seems quite plausible that there should be some relationship between mean and overall demand;

[12] This relationship will not be perfect, however, because difficulty can also be influenced by unintended demands, ie demands that are not actually relevant to the attainment that is supposedly being measured. This can happen when, for instance, a question is badly worded, or appears alongside a misleading picture, or is set in a confusing context, or suchlike.

and therefore that there will also be some relationship between mean demand and the difficulty of each item.

Figures 8 and 9 present mean demand ratings alongside difficulty index values for each item, for STA ratings and our experts' ratings, respectively. Difficulty index values were computed by subtracting from 100% our facility index values. This allowed us see the relationship between mean (judged) demand and (observed) difficulty directly.

Although there are notable points of divergence between STA ratings and those of our experts, concerning particular items, it is interesting to note that the item-level mean demand ratings parallel quite closely the empirical index of difficulty, determined from data from the full cohort of candidates in May 2016. This suggests: that these demand ratings do, indeed, provide important information concerning intended demands; and, more generally, that the demand strands do, indeed, identify the different kinds of demands that questions make of pupils.

It is also interesting to see, from Figures 8 and 9, how items 15 to 21 – which comprise the second half of the set of questions on the second text – are predominantly either high in mean demand, or high in difficulty, or both. There seems to be a particularly abrupt transition in the middle of this text from low demand/difficulty items to high demand/difficulty items. A number of our experts picked up on this transition, 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!"

4.3.3 Overall impressions

When asked to comment on the appropriateness of the demands of the paper for specific groups of pupils, some experts argued that the paper may have been too demanding for EAL pupils. This may have been partly due to the order of the texts.

> "The thing with EAL children [is]… as much as you try and provide those children with a language rich environment, if your children are coming into your school with no English whatsoever,… there's some language in there that they just… may not come across… Which is why I agree that maybe if the [text about the] dodo had gone in first, they may have encountered language that they would have encountered before."

Figure 8. Mean STA demand rating (bars) versus difficulty index (line) for each item

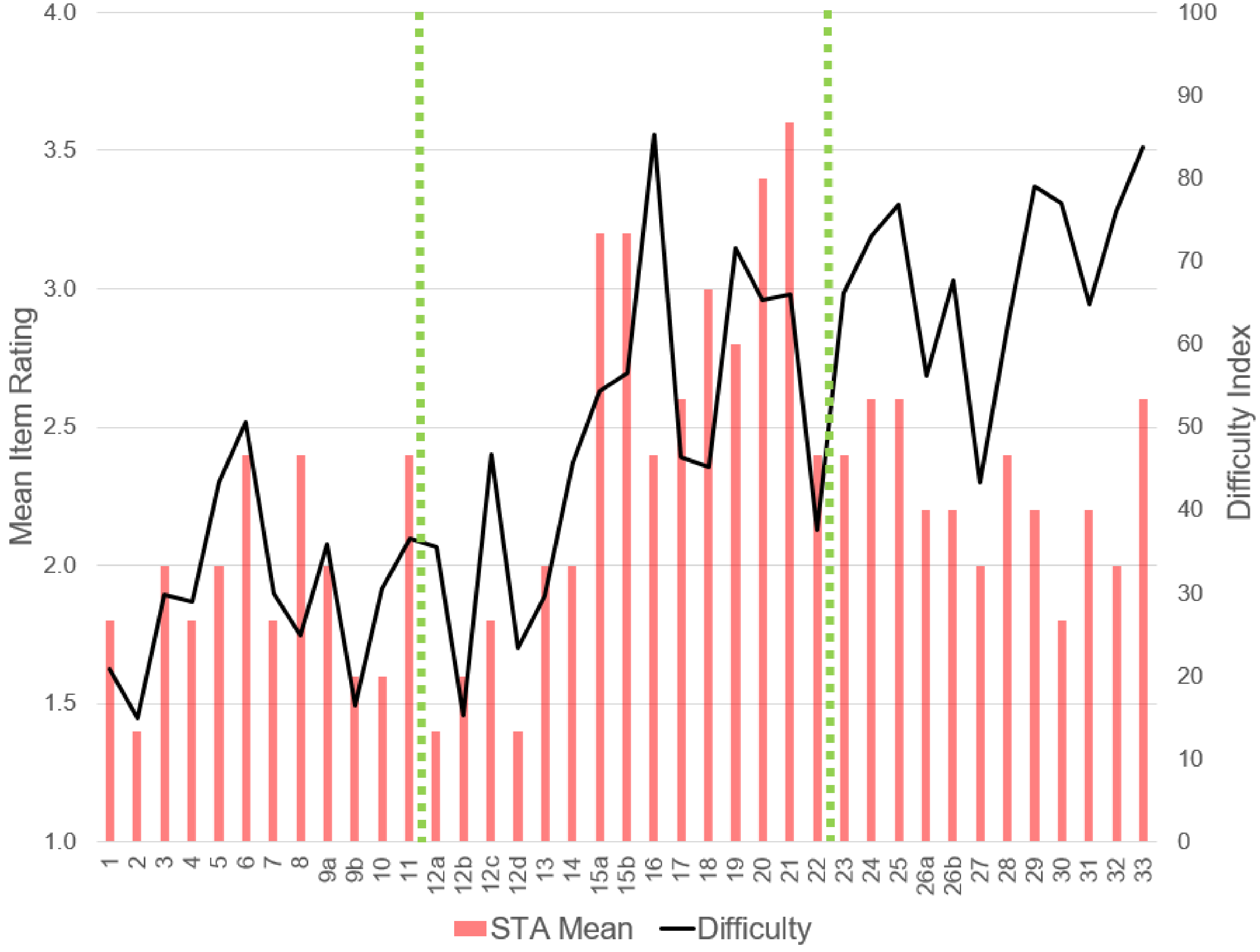
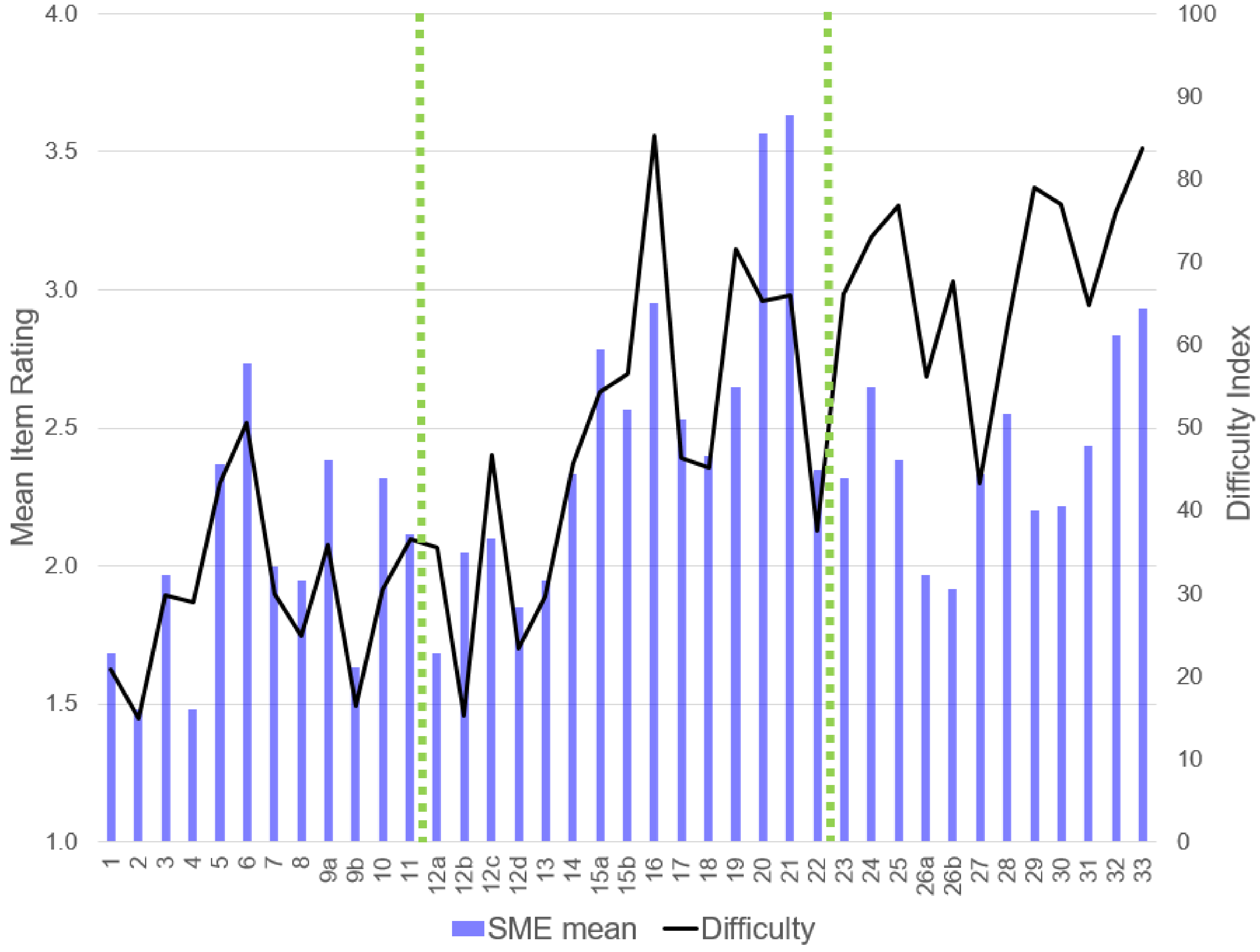


Figure 9. Mean expert demand rating (bars) versus difficulty index (line) for each item



4.4 Insights from SEND representatives

Once we had completed the analysis of results from our content validation study, we held discussions with representatives from various SEND groups, to explore their views on the accessibility of the 2016 reading test, in the light of results from our study. Our discussions included representatives from the National Association of Special Educational Needs (NASEN), the British Association of Teachers of the Deaf (BATOD), and the Royal National Institute of Blind People (RNIB). Those who attended our focus groups were also able to provide insights into challenges for pupils with other kinds of needs, eg pupils with autism.

4.4.1 Methodology

Five representatives took part in one of two focus group discussions; three in the first, two in the second. The discussions within these focus groups were largely unstructured, ie no specific interview schedules were produced, although some of the concerns and the figures presented in the preceding sections were used as prompts for discussion.

At the start of each session, extracts from forum posts and from the Select Committee report were shown to participants, to set the context for the research. They were then posed questions for consideration, with the primary questions being:

- Was the 2016 reading test hard for pupils to access?
- If so, then what caused it to be hard for pupils to access?

Participants were then encouraged to freely discuss their views/experiences, with follow-up questions being asked by the researchers. Subsequently, they were given an overview of the content validation study and its findings, as well as the aforementioned prompts (including a figure representing mean cognitive domain demand ratings and figures representing item functioning data). Following this presentation, a further discussion was held, which was again largely unstructured.

Discussions were audio recorded, and these recordings were transcribed by an external transcription company. The findings reported in the following section reflect the commonly discussed themes, with quotes extracted from transcripts to support conclusions made.

As our participants were specifically invited to represent SEND concerns related to accessibility, their comments should be interpreted in this light. Issues raised for discussion were identified as being of particular significance for pupils with one kind of SEND or another, even though many of the issues discussed are likely to be of wider significance, eg to pupils with EAL, or to low-attaining pupils. In addition, the following sections identify accessibility challenges that may or may not represent validity threats; that is, some of them may represent legitimate reading demands that certain pupils will understandably have difficulty with. From a validity perspective, what we are most interested in, here, are demands that might render the test unduly hard to access, ie demands that prevent the test from measuring reading comprehension accurately for certain groups of pupils.

4.4.2 Findings

When asked for their overall thoughts on the test, all participants agreed that the test was quite hard to access for many pupils, but that it was unduly hard to access for some pupils. They attributed this perception to factors relating to the questions in the test, to factors relating to the reading booklet (ie the reading texts), and to the combination of question and text factors. Each shall now be discussed in turn.

4.4.2.1 Question factors

4.4.2.1.1 Opening questions

Echoing some of the concerns raised in the content validation study, the most common comments made by participants concerned the order in which demands were presented within the test paper. Specifically, participants felt that the first few questions were insufficiently accessible, which they believed may have affected pupils' motivation for the remainder of the test paper. While the group acknowledged that this may have affected a range of pupils, they suggested that it may be particularly true for pupils with certain kinds of SEND.

> "What it felt like was it went straight into the harder stuff without any kind of lead-in to the harder [questions]… I think if you've got special needs… and you fail on those first two questions, that then, I think, has an impact on the whole of the test."
>
> "There always used to be that first page which had some very simple circular answer-the-questions-type things, which are very accessible for children with special needs… I can see that they might have been seen as a little bit of a wasted page, because in a way a lot of children would have got those right, but actually I do think it would have evened the playing field a bit for those that struggle."
>
> "If you've got a one-to-one session with a special needs child, and you start off on a back foot, forget the whole session… So you've got to start by winning… and then they'll engage… I think children with any kind of disadvantage actually will have that same reaction to a really hard struggle at the beginning… Even though it's perfectly within what they've normally done on other tests."

4.4.2.1.2 Question format

Participants also proposed that pupils with certain kinds of SEND may have found the beginning of the 2016 test paper to be particularly difficult because each of the first few questions required a different set of cognitive strategies to answer: Question 1 required a one-word answer, Question 2 was multiple-choice, Question 3 required a constructed response, Question 4 required a one-word answer, and Questions 5-7 required lists of responses. In contrast, the 2017 paper, which did not seem to attract the same concerns regarding accessibility, grouped the first few questions according to the type of response: Questions 1 and 2 were multiple-choice, and then Questions 3-9 were all short constructed response items.[13] The 2017 test may therefore have proved more accessible to pupils with SEND than the 2016 test, as there was less of a requirement to switch cognitive strategies.

> "If you've got special needs, once you've processed how to answer the first one, you then understand how you've got to answer the others. Whereas if you look at [Questions] one, two, three and four, they're all more or less asking you to do a different thing, so you've got to reprocess each time."
>
> "A lot of children with SEN will have processing difficulties… If they've got to re-process how they attack each question… then that's going to slow them down… And of course the children on the autistic spectrum who are quite rigidly focused… to then change that… is hard."

Certain types of questions found within the 2016 test may have posed particular issues for pupils with certain types of SEND. For example, one question required the identification of a picture (Q8 – "Which of these drawings best represents the monument?"). Participants felt that this was unduly hard to access for blind pupils, who were required to identify the correct monument by feeling a 3-dimensional representation of several objects. Multi-part questions were also commented upon, particularly those where information needed to be retained from the first part in order to answer subsequent parts.

> "Anything with pictures and diagrams [is] really, really difficult [for pupils with visual impairments]."
>
> "[Question 9 – 2016 test] had one after the other, and the second question used the pronoun 'he'. So if you use 'he', it assumes that you're reading from one sentence to the next, but of course the kids stopped and wrote an answer in between… They may have forgotten who 'he' is by that

[13] According to STA, this grouping was incidental rather than intentional.

> point… [In the] 2017 paper… the sentences are all the same structure: 'Gaby did this, Gaby did that'. And they've repeated the name and not given a pronoun… That is just so much easier to manage"

Participants also commented that the 2016 test seemed to contain a relatively large number of written response type items, which may be less accessible for some pupils, which was seen to be less of an issue for the 2017 test.

> "Are we testing children's ability to write when it's a reading test?… Obviously you'd need to make a written response sometimes, but we need to be thinking about demonstrating that they've read it and understood it, not how well they can write a response. And I think the 2017 [test] enables them to do that more than these [2016] questions… A lot of [SEND] children will have particular difficulty with writing, spelling, handwriting and writing composition."

4.4.2.1.3 Degree of inference required

Participants noted that a relatively large number questions in the 2016 test required a high degree of inference. Again, while a range of pupils may have struggled with questions requiring a degree of inference, this was seen by the participants to present particular difficulties for pupils with SEND, many of whom may have delayed language acquisition or difficulties in interpreting abstract concepts.

> "There was a lot of inference [questions], which was very difficult for pupils with delayed language to access."
>
> "I struggled with some of the questions myself! ... Just in terms of what… they were looking for. Particularly with the inferences, and for deaf children it would just be just beyond them."
>
> "It's the same for any child on the autistic spectrum… Any child with a processing issue really will always find inference and deduction, that step a little bit too far… because they'll take language very literally… A question like [Q5 – 'they cross the glassy surface of the lake – give two impressions this gives you of the water'] - they'll be sitting there going, 'what glass, where's the glass come from?'"

4.4.2.2 Text factors

4.4.2.2.1 Amount of text to read

A number of comments were made about the amount of text that pupils needed to read for the test, which was again perceived to be a particular issue for the 2016 paper. This was largely attributed to the fact that the 2016 reading booklet contained three continuous passages, whereas the sample and 2017 booklets included shorter/more structured pieces (eg, poems/text boxes). It was suggested that this could have explained why some pupils may not have reached the final text.

> "The [2016] reading paper required far more reading stamina than in previous years."
>
> "In the past we've had poems… and obviously the quantity… of reading then is not as great as always reading a full text."
>
> "There's all those studies, isn't there, of the length of time you can properly concentrate for, which varies between 20 and 40 minutes, doesn't it? And I think if you've got special needs [it's] probably less than that."
>
> "It's a shame really because, looking at that, the questions from the third passage were a lot more accessible."

Participants suggested that while time pressure and reading stamina demands can affect all pupils, this may have been particularly challenging for some pupils with SEND. In particular, it was noted that many pupils with SEND may struggle to scan/skim-read in order to find relevant sections of text quickly.

> "There was a lot of reading that had to be done really quickly and scanned, which deaf pupils can't do very easily... Auditory processing is slower in deaf children than in hearing children… so reading is a slower process… And if they're skimming they tend to miss [information]."
>
> "That's the same for autists. Autistics have that auditory processing issue."
>
> "The quantity of the braille that [blind children] have to get through is absolutely enormous, and there isn't any skimming… They have to re-read everything over and over again… My children never ever get on to the third story, and that's having double time! They could do with somebody just pointing, going through it and saying 'this is where you start reading, now read here', but we're not allowed to do that."

4.4.2.2.2 Text format

As already touched upon, participants suggested that the way in which the reading booklet was presented could have made the 2016 test unduly hard to access for pupils with certain kinds of SEND. For example, they pointed out that the 'boxed' structure used for the first text in the sample booklet seemed much more accessible than the extended writing structure used for the first text of the 2016 booklet. In addition, the second text of the sample booklet was a poem, with just two distinct sections, whereas the second text of the 2016 paper was another extended piece of writing.

> “In the example materials, some of the texts had information presented in small chunks, small boxes or short paragraphs, making it appear more accessible… But then when they got the actual [2016] papers it was a solid block of text.”
>
> “There was so much text [in the 2016 paper]… there was nothing to draw the eye apart from straight text, and a lot of pupils would just go, ‘I can’t read that, it’s too hard’… It’s going to have more effect on the performance of less confident students and those performing at a lower level… Even though it may not have been actually hard to actually read, just the view of that to a non-confident reader is going to be ‘oh that’s hard’.”

Breaking texts into easily identifiable subsections was noted to be particularly helpful for blind pupils, as it makes it easier for them to quickly find the relevant section of the text needed to answer the question, without needing to read the text from the beginning each time to locate the relevant section.

> “If it’s got headings… They would know that that was question and answer. So if it’s broken up into headings and subheadings it makes it much [easier].”

Some participants also noted that preventing texts from crossing pages would also be helpful.

> I think that’s where [the 2017 reading booklet is] better, because I don’t think any of [the texts] were over more than two pages… I think it [matters] for those children who are at the lower end who are struggling, and who are reading perhaps more slowly… You think you’ve perhaps finished when you get to [the second page of text 2 of the 2016 reading booklet] and then you turn over and… your heart must just sink, as a child who struggles to read… if it could be just on [2 pages]… it definitely helps; you can see everything in front of you.”

It was suggested that the choice of font may have had some impact on accessibility for some pupils. Specifically, the first text of the 2016 reading booklet used a serif font,[14] while the second and third texts used a font without serifs, more commonly

[14] Serifs are small flourishes added to the tips of some letters.

known as ‘sans-serif’. It was felt that it may have been more appropriate to have used the more accessible (sans-serif) font in the first text.[15]

> “Everything else is sans-serif and this [first text] is a serif font… It’s harder to read because the eye travels more because you’ve got the flicks and the extra bits on the end of the letters.”

4.4.2.2.3 Context of the narrative

Also contributing to the perception of inaccessibility for some pupils was the observation that contexts within which texts were situated might have seemed relatively unfamiliar to them, echoing concerns raised in the main content validation study. Participants felt that this may have posed additional demands for pupils with SEND, as well as those from low socio-economic backgrounds.

> “The life experiences you needed to access the test were quite unusual. I mean, to put yourself in the place of somebody riding a giraffe, I found that quite hard to envisage.”
>
> “[The first text had] very, very difficult content, particularly with children with visual impairment. They would… probably never seen in their lives a monument; they wouldn’t know what it was. There’s no incidental learning… Asking them questions about a monument, and what did the monument look like, and has it got a crown on the top of it – absolutely meaningless to them. So the whole content was very, very difficult.”
>
> “I know it’s supposed to be harder, but… there’s got to be something within it that children with limited understanding can access… You’re not [testing their reading], because you get children who can actually read the words and sound them out, but if they don’t know what they mean, they can’t [access it], if they don’t have a mental image.”
>
> “I think it’s quite middle class. If you like The Secret Garden, Tom’s Midnight Garden, all those sorts of things then you’d be fine… but really if not – which most children aren’t, let’s be honest – there’s nothing much to grasp on to that you’re familiar with.”

It was noted that the sample and 2017 texts appeared to be better in this regard, as they had narratives containing more familiar contexts for most pupils.

[15] Sans-serif fonts are generally recommended for easier readability (eg Ofqual, 2012; CCEA & Welsh Government, 2015), but the academic literature on this matter does not appear to be conclusive (eg Arditi & Cho, 2005).

> “Rescuing a cat [2017 test – text 1] is a little bit more the norm than going to an island and finding a monument [2016 test – text 1]. The content of it was better.”
>
> “I shouldn’t generalise either, but when I think about some of the key interests of a lot of autistic people I mix with – dinosaurs and space [sample test] – you’re hitting an awful lot of interests there… in terms of themes around what children are interested in.”

4.4.2.2.4 Language/vocabulary

It was suggested that some of the vocabulary required to answer questions was difficult to access. Again, this may have imposed particular demands for some pupils with SEND whose vocabulary may not be as rich, particularly those with delayed language acquisition. This issue might have been compounded by the unfamiliar context in which the texts were presented, if pupils were less able to infer the meaning of unfamiliar words from the context of the story.

> “There were a lot of things where actual experience was required to understand the language and the vocabulary, which deaf children just might not have had. It was just very difficult for them to envisage and therefore to work out if they didn’t know the specific vocabulary… they couldn’t work it out from their own experience.”
>
> “The use of some of the vocabulary that was so specific like ‘milled’ and ‘parched’, it’s just not within the experience of deaf children, because it’s not in everyday use… They won’t have come across that, because they don’t learn by overhearing, so it’s very, very difficult to access.”
>
> “When you’re working with a child with hearing impairment with comprehension, you typically will take the passage and pre-tutor them, so that you prepare the vocabulary and you give them the vocabulary needed to be able to go away and work on the text. But in a test situation you can’t do that. So you’re completely at the mercy of the test paper.”

Participants also commented on the abstractness of some of the language used within the texts. For example, the first text was noted to contain a lot of imagery, which posed added demands on pupils. Again, being the first text, this could have had knock-on effects on some pupils’ motivation at this early stage.

> “The very first passage has a lot of imagery in it and is quite… high level in terms of the content… I think as a child who’s struggling with reading you can imagine that the stress levels are just rising like this and the feeling of panic. And you’ve only got an hour and you’ve got how many pages, quite

> a lot of pages, lots of text. And once that kicks in, that stress from straightaway finding the first page difficult, your brain stops functioning as effectively"

4.4.2.3 Combination of question and text factors

Several participants suggested that although each of the aforementioned factors may have contributed to the perception of inaccessibility for certain pupils, each individual factor would have been unlikely to make the test unduly hard on its own. Rather, the reason why the 2016 test may have seemed to be so hard to access was more likely due to a combination of multiple challenging demands related both to questions and to texts. One group member described the test as a 'perfect storm' of factors.

> "I wonder if it's a combination of the different issues, so the booklet being text heavy and difficult and… the difficulty of some of the questions within the booklet, and you get a situation where the children feel under too much [pressure]… It's the combination I think… 2016 was just a perfect storm: just the texts [and] the way the questions [were]."

Time pressures were again noted here. In particular, participants felt that the amount of text that needed to be read and the number of questions that were asked made it very difficult for some pupils to complete the test within an hour.

> "33 questions in an hour plus all your reading. It's hefty, isn't it?"
>
> "It's a lot to do in an hour. You'd have to go some, even as a bright child, to get through the whole lot in an hour, wouldn't you?"
>
> "I actually sat down and did both tests. And it took me 45 minutes to do this and I thought this is far too long. It shouldn't be taking me 45 minutes of actually proper concentration."

It should be noted that many pupils with SEND will have approved access arrangements, which may include extra time to complete the test. However, while potentially easing time pressures for some, this was not always seen to be necessarily a solution to the problem.

> "To be honest even if you have extra time you've got to know what to do with it anyway. If you're so stressed out by the content of what you're doing, that extra time is not going to help; it's just going to mean that you're sitting there for longer crying or… just feeling very tense. So it works for some children but it doesn't necessarily work for a lot of children."

4.4.3 Reflections on insights from SEND representatives

Our participants provided a variety of insights into why the 2016 test might have presented particular challenges for pupils with one kind of SEND or another. Some of these challenges seemed to be construct-irrelevant, involving demands that are not part of the reading comprehension construct. Being able to read much more easily in one font than another is presumably construct-irrelevant in this sense. Consequently, if a text's font happened to present particular challenges for certain pupils, then this would represent a construct-irrelevant demand, potentially rendering questions on that text unduly hard to access for those pupils (even if only by a relatively small amount). Other challenges seemed to be construct-relevant, involving demands that are clearly part of the reading comprehension construct. Being able to infer meaning from texts and accompanying questions is clearly construct-relevant in this sense. Our participants felt that the texts required a high degree of inference, and that this might have been particularly challenging for pupils with certain kinds of SEND, eg those with delayed language acquisition or difficulties in interpreting abstract concepts. However, although this may well have made the test harder to access for these pupils, the fact that inference is a construct-relevant demand suggests that this is not necessarily a threat to validity.

An important question concerns whether interactions of challenging but construct-relevant demands – perhaps in combination with additional construct-irrelevant demands that, in the normal run of events, might be considered sufficiently trivial to overlook – could render the test unduly challenging for particular pupils, or for particular groups of pupils. The idea of a 'tipping point' is a useful metaphor here; at which a certain accumulation of challenges begins to make the overall demand of the test more than the sum of its component demands, if only for some. The perceptions and reflections expressed by our participants suggested that some pupils may have experienced an effect like this. This kind of effect, were it to have occurred, would have been extremely difficult, if not impossible, to predict in advance, on the basis of test and item reviews. Indeed, STA explained to us that their own SEND experts did not identify the kind of issues presented above.

STA noted that during their Inclusion Panel meeting, which involves a range of SEND experts, participants are asked to comment on the suitability of all aspects of the text as well as the individual items. For the texts in the 2016 tests, the majority of comments received from the Inclusion Panel related to layout, and many of the suggestions, such as making the captions clearer in the Dodo text, were enacted. There were no comments from the panel that suggested the text content was inappropriate for pupils with SEND or was more likely to be difficult for SEND pupils. Comments on individual items were considered alongside feedback from teacher and other curriculum and assessment experts before the wording was finalised.

4.5 Test enjoyment ratings

A final piece of evidence comes from the test development cycle, and concerns pupils' experiences during the technical pre-test (TPT). After taking the test, pupils were asked whether they found each of the three texts enjoyable. Table 6 summarises responses from the TPT of the 2016 test (trialled in 2015) and the 2017 test (trialled in 2016). Overall, the 2017 texts were enjoyed more than the 2016 ones, although there was little difference between enjoyment ratings for the first text of each test, and differences between ratings for the second and third texts were not vast.

Table 6. Percentage of TPT sample that found each text enjoyable.

	% enjoyed
2016 test	
The Lost Queen	64.0%
Wild Ride	56.0%
The Way of the Dodo	52.0%
2017 test	
Gaby to the Rescue	64.9%
Swimming the English Channel	64.5%
An Encounter at Sea	61.0%

5 Conclusion

The two aims of our review were to:

- improve our understanding of teachers' concerns over the accessibility of the 2016 reading test, including the variety of potential causes of those concerns; and
- consider whether there may be any questions for the Standards and Testing Agency concerning the potential for enhancing test development procedures for future years.

Many teachers felt that the 2016 test was hard for pupils to access. Of course, a hard test is not necessarily a bad test, as long as it allows all pupils for whom it has been designed to demonstrate their true level of attainment in the subject area. However, it would be a problem if features of the test prevented some (or all) pupils from demonstrating their attainments. If so, then, for those pupils, the test would be unduly hard to access.

The 2016 reading test was designed differently from previous years. It was intended to differentiate between a wider range of pupils, and it was designed to accommodate a higher 'expected standard' threshold. It was therefore likely to result in a somewhat different kind of experience. The evidence reviewed in this report concerns how pupils, their teachers and other stakeholders experienced the 2016 test, and how they reflected on that experience. In considering the collated evidence, it is important to appreciate that we have made good use of perceptions of inaccessibility, including reflections on possible causes of inaccessibility. Not all of the conclusions reached by teachers and stakeholders will necessarily have been entirely accurate: some may have been partially accurate; others may have been inaccurate. Having said that, they all constitute evidence of teachers' and other stakeholders' concerns, and therefore help us to understand the 2016 test experience.

One of the most challenging issues to identify and address is the threat of differential validity; in particular, features of assessment tasks that inadvertently inflate or deflate results for particular subgroups of pupils. This includes the possibility that features of texts or questions render an assessment task somewhat inaccessible for certain pupil subgroups, but very much more inaccessible for others. Many of the perceptions and reflections collated above raised concerns like these; for example, a focus group of primary teachers from Teesside suggested that it might have been relatively harder for pupils from lower socio-economic backgrounds to engage with certain of the texts; similarly, our SEND representatives suggested that it might have been relatively harder for pupils with certain kinds of SEND to deal with abrupt transitions in response strategy demands across questions. Again, although these teachers and stakeholders have raised important concerns, the present report is unable to substantiate or refute them. So their views should be understood as sources of relevant evidence alongside other sources of relevant evidence; eg alongside views expressed by teachers and SEND experts during the test development process, and alongside statistical evidence on test and item functioning produced as part of the test review process. It is important to acknowledge that not all of these sources of evidence are entirely consistent. For instance, some of the most extreme views aired on social media suggested that even high-attaining students would have been flummoxed by the test; whereas evidence from the mark distribution demonstrated that many high-attaining students performed extremely well. In fact, the distribution of marks was very Normal (in a statistical sense) with the 'typical' pupil scoring around half marks and with similar numbers of pupils scoring very high marks as very low marks.

The following observations illustrate the issues that emerged from our review:

- although teachers were anticipating a higher 'expected standard' threshold, concerns began to be expressed before that standard had been set, rather than directly in response to it;
- concerns expressed on social media had different emphases. Some commentators were mainly critical of the texts; others were mainly critical of the questions. Some described the test as generally very hard; whilst others described the test as particularly hard for certain groups of pupils, including those with EAL and SEND.
- conclusions expressed more formally by teacher and subject associations echoed concerns expressed on social media, including the following perceptions and reflections:
    - certain contexts within certain texts may have been especially challenging for large numbers of pupils to engage with, eg the 'antiquated' feel of the first text;
    - particularly difficult questions that occurred earlier in the test may have demoralised lower-attaining pupils and prevented them from demonstrating their true level of attainment on later questions; and
    - too many pupils failed to complete the test in the allocated time.
- test and item functioning data seemed to support some of these concerns, for instance:
    - although the first few questions were answered correctly by the vast majority of pupils, the hardest questions on both the first and second texts occurred mid-way through their respective question sets;

  - perhaps relatedly, omit rates began to rise considerably towards the second half of the questions on the second text; and
  - nearly one in twenty pupils were deemed not to have reached the first question relating to the third text, and around a quarter of all pupils were deemed not to have reached the end of the test.

- readability analyses suggested that the 2016 reading booklet was harder to read than both the sample booklet and the 2017 booklet. However, the 2016 reading booklet was not necessarily more complex than booklets from previous years. Indeed, according to one analysis, the range of reading complexity levels across the (three) texts from 2016 was similar to the ranges observed across the (five) texts (split across the two tests) from the years 2012 to 2015.

- according to outcomes from our content validation study:

  - items in the 2016 test reflected appropriate levels of demand – both according to STA ratings and according to those of a group of independent experts – as judged in relation to the *Test Framework* blueprint;
  - however, in terms of the technical knowledge required strand, our independent experts rated items from the 2016 test somewhat higher than STA had; and
  - items 15 to 21 (the second half of the second question set) were predominantly either high in mean demand, or high in difficulty, or both; suggesting an abrupt transition in the middle of this question set from low demand/difficulty items to high demand/difficulty items.

- insights from SEND representatives highlighted a number of question, text, and question-text-interaction factors that might have proved particularly challenging for pupils with SEND (and perhaps also for other pupils too), for instance:

  - although the first few questions were answered correctly by the vast majority of pupils, they were still felt to be insufficiently accessible for pupils with SEND (and this impact might have been compounded by variation in cognitive strategies required to answer them);
  - the fact that all of the texts were presented as continuous passages may have been particularly challenging for certain groups, eg pupils with visual and auditory impairments and pupils with autism;

  - the language and vocabulary of the 2016 test was perceived to be technically very challenging, particularly so for pupils with SEND; and
  - time pressure was also considered to present an unnecessary barrier.

So, in the light of the evidence that we have collated and summarised above, is it possible to conclude that the 2016 reading test either was, or was not, unduly hard to access? A range of possible answers could be given to this question, for instance:

1. it was unduly hard to access for many, if not most, pupils
2. it was quite hard to access for many pupils, but it was unduly hard to access for certain groups of pupils, eg
    - lower-attaining pupils
    - pupils from certain socio-economic groups
    - pupils with EAL
    - pupils with SEND
3. it was quite hard to access for many pupils, but it was not unduly hard to access for any pupils (who would normally be entered for the test)
4. it was no more nor less accessible than the sample test or the 2017 test, it just seemed (to teachers) to be harder for pupils to access.

On the balance of evidence presented, it does seem reasonable to rule-out answer 4. On the basis of readability statistics alone, it seems fair to conclude that the 2016 test was less accessible than might have been expected, if those expectations had been set on the basis of the sample test.

Equally, it seems reasonable to rule-out answer 1. The vast majority of pupils did make it to the end of the test, many pupils performed well on the test, and the individual items seemed also to function well.

So was it just a hard test, ie answer 3? A hard test is not necessarily problematic from the perspective of delivering accurate results, as long as those results are still capable of distinguishing between pupils with different levels of reading comprehension – both at the top of the scale as well as at the bottom – and as long as the difficulty of the test impacts similarly upon all pupils. On the other hand, an unexpectedly hard test can be problematic: the harder the test turns out to be, and the less anticipated its level of difficulty, the more likely it is that this will unduly inhibit the performance of a certain number of pupils. So an unexpectedly hard test is not desirable, from a test development perspective.” Alternatively, was the test unduly hard for certain groups of pupils, ie answer 2? In other words, did accessibility issues prevent the test from measuring reading comprehension accurately for a considerable number of pupils from certain groups?

Again, on the balance of evidence presented, it seems plausible that the combined impact from multiple ostensibly negligible challenges – stemming from both question and text factors – may have rendered the 2016 reading test unduly hard to access for at least some pupils. Unfortunately, aggregate impacts of this sort can be difficult, if not impossible, to detect statistically, and can be equally hard to identify via test reviews. Although accessibility issues may have prevented the test from measuring reading comprehension accurately for particular pupils, and perhaps also for certain groups of pupils, we do not have sufficient evidence to be able to reach any definitive conclusion concerning which pupils might have been affected in this way, nor how many pupils, nor to what extent. Indeed, the evidence is mixed.

The data on test and item functioning, for instance, paint a fairly positive picture. The test and its items appeared to function well, with little evidence of bias related to boys/girls or EAL/non-EAL. On the other hand, the DIF statistic used to investigate bias is not sensitive to test-level bias, only to item-level bias, and the kind of inaccessibility that we have been considering might well impact as much at the test level as at the item level. Discussions with SEND representatives, as well as with representatives from English associations who reviewed drafts of our content validation study report, raised general concerns with the 2016 reading booklet, for instance:

- an old fashioned feel to the first text and a sense of it disadvantaging pupils from lower socio-economic backgrounds
- a prevalence of idiomatic usages presenting particular challenges for EAL and SEND pupils
- the amount of technical knowledge required to make sense of the final text.

Ratings from our independent experts also raised the question of whether the test made greater demands on pupils’ technical knowledge than STA had anticipated.

The business of test design/development is not a precise science, and it always involves trade-off and compromise. However, test designers/developers need to be able to provide reassurance that their processes are sufficiently rigorous, and that design/development decisions strike appropriate balances. The experience of 2016 raises a number of important questions for STA, in particular:

- **Are pupils given sufficient time to complete the reading test?** Although 'reading at speed' might be considered a legitimate part of the reading comprehension construct – an aspect of reading fluency – it is not immediately obvious what that ought to mean in terms of the percentage of pupils expected to reach the end of the test. Clarification on this issue would be useful.

- **Are there ways in which test and item review processes can be made more rigorous?** The fact that STA did not foresee the intensity of concerns that teachers expressed with the 2016 reading test, nor certain of the particular concerns identified within the present report – despite having followed well-established processes which included consulting a teacher panel, a test review group, and an inclusion panel during the test development phase – raises the question of why, and of whether such concerns might have come to light had these panels been run differently, or if a different kind of review process had been adopted, or suchlike.

- **Can alternative approaches be adopted to investigate potential biases more effectively for pupils from various groups (eg SEND, EAL, socio-economic)?** The potential for bias within reading tests can be harder to spot than for tests in many other subject areas. This is partly because demands related to reading load, task context, etc, are clearly construct-irrelevant in relation to these other subject areas, and their potential for causing construct-irrelevant score variance, ie bias, is fairly obvious. In other words, it is clear, in principle, how bias might occur; even though, in practice, it might be tricky to determine whether it actually has occurred. Conversely, for reading tests, demands related to reading load, task context, etc, are either construct-relevant, or less obviously construct-irrelevant, which makes their potential for causing bias far less obvious; even though bias is still a very real threat. Potential biases are also harder to spot for the key stage 2 reading test because it comprises sets of questions linked to common passages, rendering outcomes from DIF analyses harder to interpret. These issues raise the question of whether additional steps can be taken to address problems, like these, which are specific to the key stage 2 reading test.

6 References

Arditi, A., & Cho, J. (2005). Serifs and font legibility. *Vision Research*, *45*, 2926–2933.

Council for the Curriculum, Examinations and Assessment (CCEA) & Welsh Government (2015). *Fair Access by Design: Guidance for awarding organisations on designing high-quality and inclusive qualifications*. Retrieved August 24, 2017, from http://ccea.org.uk/sites/default/files/docs/accreditation/compliance/fair_access_by_design_june_2015.pdf

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221-233.

House of Commons Education Committee (HCEC) (2017). *Primary Assessment. Eleventh Report of Session 2016–17. HC 682*. London: House of Commons.

Janan, D. and Wray, D. (2012). *Readability: The limitations of an approach through formulae*. Paper presented at the British Educational Research Association Annual Conference, University of Manchester, 4-6 September. Retrieved September 13, 2017, from http://www.leeds.ac.uk/educol/documents/213296.pdf

National Union of Teachers (NUT) (2016). *The Mismeasurement of Learning: How tests are damaging children and primary education. Reclaiming Schools: The Evidence and the Arguments.* London, UK: National Union of Teachers.

Newton, P.E. & Cuff, B.M.P. (2017). *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.* Coventry, UK: Office of Qualifications and Examinations Regulation.

Ofqual (2012). *Guidance on the Principles of Language Accessibility in National Curriculum Assessments Research Background*. (Ofqual Report 12/5217). Coventry, UK: Office of Qualifications and Examinations Regulation.

Parker, K. (2017). This year's Sats paper 'more readable'. Retrieved July 17, 2017, from https://www.tes.com/news/school-news/breaking-news/years-sats-paper-more-readable

Sanford-Moore, E.E., Koons, H. and Bush, L. (2016). A*n Examination of the UK's Key Stage Tests' Reading Section Complexity*. Retrieved July 17, 2017, from https://lexile.co.uk/lexile-international-prod-media/pdfs/MM_UK_Key_Stage_Reading_Test_Digital.pdf

Standards and Testing Agency (STA) (2016). *English Reading: Administering the English reading test reading booklet and reading answer booklet*. Retrieved July 17, 2017 from https://www.gov.uk/government/publications/key-stage-2-tests-2016-english-reading-test-materials

Ward, H. (2016). Sats: pupils in tears after sitting “incredibly difficult” reading test. Retrieved July 17, 2017, from https://www.tes.com/news/school-news/breaking-

news/sats-pupils-tears-after-sitting-incredibly-difficult-reading-test

Appendix 1 Differential Item Functioning Data

Question	Tech. Pre-test DIF gender	Tech. Pre-test DIF EAL
Q1	No significant DIF	No significant DIF
Q2	Negligible favouring boys	No significant DIF
Q3	No significant DIF	No significant DIF
Q4	Negligible favouring girls	No significant DIF
Q5	No significant DIF	No significant DIF
Q6	No significant DIF	No significant DIF
Q7	No significant DIF	No significant DIF
Q8	Negligible favouring girls	No significant DIF
Q9a	No significant DIF	Negligible favouring Not EAL
Q9b	Negligible favouring girls	No significant DIF
Q10	Negligible favouring boys	No significant DIF
Q11	Negligible favouring boys	No significant DIF
Q12a	No significant DIF	No significant DIF
Q12b	Negligible favouring boys	No significant DIF
Q12c	No significant DIF	Negligible favouring Not EAL
Q12d	No significant DIF	No significant DIF
Q13	No significant DIF	No significant DIF
Q14	No significant DIF	No significant DIF
Q15a	Negligible favouring girls	No significant DIF
Q15b	No significant DIF	No significant DIF
Q16	No significant DIF	Negligible favouring Not EAL
Q17	No significant DIF	No significant DIF
Q18	No significant DIF	No significant DIF
Q19	Negligible favouring boys	Negligible favouring Not EAL
Q20	Negligible favouring girls	Negligible favouring EAL
Q21	No significant DIF	No significant DIF
Q22	No significant DIF	No significant DIF
Q23	No significant DIF	No significant DIF
Q24	No significant DIF	No significant DIF
Q25	No significant DIF	No significant DIF
Q26a	Negligible favouring boys	No significant DIF
Q26b	No significant DIF	No significant DIF
Q27	No significant DIF	No significant DIF
Q28	No significant DIF	No significant DIF
Q29	No significant DIF	No significant DIF
Q30	No significant DIF	No significant DIF
Q31	No significant DIF	No significant DIF
Q32	No significant DIF	Negligible favouring Not EAL

Q33	No significant DIF	No significant DIF

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