Standards
& Testing
Agency

English grammar, punctuation and spelling

2013 technical report

Contents

Table of figures

1 Introduction

In July 2012, in response to Lord Bew's independent review of Key Stage 2 testing, assessment and accountability, the Government announced that a new statutory English grammar, punctuation and spelling test[1] (hereafter known as 'the test') would form part of the statutory assessment arrangements for children at the end of Key Stage 2 from the 2012-13 academic year.

The test contributes to the assessment of a child in English and is based on the relevant sections of the 1999 National Curriculum statutory programme of study for English at Key Stage 2 and Key Stage 3 and related attainment targets. The domain will include questions that measure:

- sentence grammar (through identification and grammatical accuracy);
- punctuation (through identification and grammatical accuracy);
- vocabulary (through grammatical accuracy); and
- spelling.

The test will be administered during the Key Stage 2 test week that commences 13 May 2013.

1.1 Purpose of this document

This document provides an initial technical evaluation of the test, including information relating to Ofqual's common assessment criteria of validity, reliability, minimising bias, comparability and manageability as set out in its *Regulatory Framework for National Assessment arrangements* (Ofqual, 2011). This document is primarily aimed at a technical audience, but contains information that will be of interest to all stakeholders involved in the test, including schools. This technical report will detail how the test and its framework was developed and demonstrate how well the test meets the purposes set out below.

This document does not contain specific information about test questions. The evidence found in this report is primarily from a large scale technical pre-test that took place in June 2012. This has informed the 2013 test cycle and will inform all future test cycles.

1.2 Purpose of the test

As outlined in the review of Key Stage 2 assessment by Lord Bew[2], the main purpose of statutory assessment is to:

- Ascertain what pupils have achieved in relation to the attainment targets outlined in the National Curriculum.

[1] http://www.education.gov.uk/a00192285/government-response-to-bew-key-stage-2-review-published
[2] https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-00068-2011

In addition, a number of principal uses were also identified:

- to hold schools accountable for the attainment and progress made by their pupils and groups of pupils;
- to inform parents and secondary schools about the performance of individual pupils; and
- to enable benchmarking between schools, as well as monitor performance locally and nationally.

1.3 Test format

There are two components of the test at levels 3-5 and three at level 6. Both levels consist of:

- a section of short answer questions assessing grammar, punctuation and vocabulary; and
- a spelling section.

The level 6 test also includes an extended task, which assesses the technical aspects of writing.

The test will be administered on paper with the spelling component administered aurally by a test administrator. The total testing time for each of the levels 3-5 and level 6 tests will be approximately 1 hour.

2 Executive Summary

The English grammar, punctuation and spelling test has been developed by STA in line with its usual test development procedure for National Curriculum tests. Although the timeline for development has meant that the time available for some of the activities has been reduced, this report demonstrates that a sufficient process has been followed to ensure high quality test materials.

A number of independent experts, including teachers, academics and other education professionals have been involved throughout the development process. Evidence from this expert review has been used alongside evidence from trialling and a number of validity studies in order to produce the test framework and test questions.

The STA believes that the processes used to develop tests are demonstrably robust and in line with international best practice such that there can be confidence in the outcomes of this process.

2.1 Common assessment criteria

2.1.1 Validity

The development of a validity argument must start with an understanding of the purpose of the assessment. The statutory purpose of National Curriculum tests is to assess 'the level of attainment which [pupils] have achieved in any core subject'. In addition, the Bew review set out three additional principal uses for National Curriculum tests:

- holding schools accountable for the attainment and progress made by their pupils and groups of pupils;
- informing parents and secondary schools about the performance of individual pupils; and
- enabling benchmarking between schools; as well as monitoring performance locally and nationally.

Since these three uses relate to how the data is used following live administration, it is not possible to provide a full validity argument for them at this time. The evidence in this report, however, does provide evidence relating to the statutory purpose.

To determine whether the test is a sufficiently valid assessment of the level of attainment which children have achieved in English grammar, punctuation and spelling there are a number of questions that need to be answered:

- Is the test framework an appropriate assessment of the relevant sections of the National Curriculum programme of study in English?
- Is the test an appropriate assessment of English grammar, punctuation and spelling?
- Are the reported outcomes of the test appropriate with respect to National Curriculum levels?

In relation to the first question, the test framework was developed to closely align to the relevant elements of the National Curriculum programme of study for English and the reference codes assigned to the assessable elements of the test are explicitly linked to the relevant section of the programme of study. This ensures that all of the questions in the test can be directly linked to aspects of the National Curriculum. The development of the test framework has involved a number of experts in the field and has been supported by evidence from trialling. Therefore, STA believes that the test is reflective of the relevant sections of the National Curriculum programme of study for English and that the framework is appropriate.

In relation to the second question, the test development process has collected a great deal of evidence relating to the content of the test and whether the questions appropriately assess the relevant skills, in particular the work on construct irrelevant variance that showed very few questions assessing something other than the construct. The experts involved in the development of the test have a wealth of expertise and experience. Trialling has provided sufficient data on the questions to enable STA to construct a test to meet the specification in the test framework.

Although the independent experts who reviewed the materials raised some concerns about the nature of the test, they appreciated that this specification was a product of Lord Bew’s recommendations. On balance, the evidence from the independent experts gives STA sufficient confidence that the test is assessing English grammar, punctuation and spelling appropriately. STA therefore believes that the test is an appropriate assessment of English grammar, punctuation and spelling, within the parameters defined by Lord Bew’s recommendations.

The answer to the final question cannot be provided until standards have been set on the live 2013 test. However, STA is confident that the process that it will follow, which is widely used internationally, will ensure that reported outcomes are appropriate.

The development of a validity argument is an on-going process. STA will continue to collect evidence to demonstrate that the test is sufficiently valid for the purpose for which it is intended.

2.1.2 Reliability

To demonstrate sufficient reliability for the test, the following aspects must be considered:

- The internal consistency;
- The classification consistency;
- The classification accuracy; and
- The consistency of scoring.

The analysis of the evidence from the pre-test has demonstrated generally high levels of internal consistency for the test and reasonable standard errors of measurement for each component.

Classification consistency refers to the extent to which children are classified in the same way in repeated applications of a procedure. Although limited evidence is available at this stage, evidence from the test re-test/alternate forms study shows that the basic descriptive statistics across the groups of children participating were very similar and the correlation of each set of scores is high enough to have confidence in the reliability of the alternative forms.

Classification accuracy refers to how precisely children have been classified. Reasonable estimates of classification accuracy will only be valid once the test has been administered in all schools. Therefore, further work on reliability will be analysed and reported in autumn 2013.

Consistency of scoring relates to the extent to which children are classified the same way when scored by different markers. Evidence from the double marking study indicates a high level of marker agreement for the test questions.

At present, STA is satisfied that the test is a sufficiently reliable assessment.

2.1.3 Comparability

When introducing a new test there are often no existing assessments with which to be comparable. However, the test development process has also produced an anchor test that will be used to link standards in future pre-tests to those that will be set on the live test this summer, therefore ensuring comparability.

2.1.4 Minimising bias

The evidence from the SEN studies shows that the most problematic questions for children with SEN were those with unfamiliar language, complex or unclear instructions, a high word count and high working memory requirement. Questions with an unfamiliar layout and questions being too close together on the page were also problematic. However, the number of questions highlighted as being problematic was generally low, and were either able to amended or were excluded as far as possible. This is in part due to the work already done to make the questions clear, concise and with simple language.

2.1.5 Manageability

The test replaces the English writing test in the National Curriculum test timetable and has similar administration requirements in terms of time length and administration (a mixture of written test and aural test).This means that the test is not placing an additional burden on schools and should therefore be manageable. Evidence about the usefulness of the outcomes cannot be provided until results are available.

2.2 Overall statement in relation to common criteria

Having examined all of the evidence gathered so far through the test development process, STA is satisfied that the test is sufficiently a valid assessment of the domain, has acceptable levels of reliability and is fair for children and manageable for schools. However, as stated previously, the development of a validity argument is an on-going process and additional analysis will be carried out following the first live administration of the test to ensure that STA can continue to be confident in this assertion.

3 Test development process and expert review

The development process for the test began in July 2011 with a comprehensive analysis of the current National Curriculum programmes of study at key stages 1, 2 and 3 in order to ascertain the assessable domain for the test. At the same time, research was undertaken to review similar tests in other jurisdictions, non-statutory guidance and support material available to schools in the past ten years.

In summary, this initial research outlined:

- the areas of the National Curriculum that were in scope for testing;
- areas that were potentially in scope but in need of further review; and
- elements of grammar, punctuation, spelling, handwriting[3] and vocabulary that were assessed in other jurisdictions but which were outside the scope of the current National Curriculum.

3.1 Initial development and expert review 1

In September 2011, following the initial definition of the domain, an expert group was recruited (see Annex 1). The group was led by the Senior Test Development Researcher for English at the Standards and Testing Agency (STA). It considered the work to date and made a more detailed examination of the curricula and assessments of other high-performing jurisdictions. This included an examination of tests from states in New England and New York in the United States, Australia, New Zealand, China and South Korea, as well as national tests available in Scotland and a variety of 11+ test formats in Northern Ireland and England.

As a result of this work, the assessable domain was refined and a number of question formats were identified to guide question-writing, including proposed formats for the assessment of handwriting, extended writing at level 6 and spelling. A series of reference codes was also developed to help categorise short-answer questions[4].

The outputs from this group were further reviewed from an academic perspective by Professor David Crystal, Bangor University. Professor Crystal scrutinised the technical information and definitions of terms for the proposed tests and alerted the STA's Test development team to likely challenge from people from different academic perspectives within the fields of grammar and sociolinguistics. Professor Crystal's input was considered in detail and incorporated into the next phase of development.

[3] At this stage in the test development process it had not been decided whether or not to include the assessment of handwriting in the test. Further information on the decision to remove the assessment of handwriting from the test will be discussed later in this report.

[4] The reference codes for the test are available in the Test Framework available at http://www.education.gov.uk/schools/teachingandlearning/assessment/keystage2/b00218030/gps-sample-materials

Following a procurement exercise, the National Foundation for Educational Research (NFER) were contracted to develop questions on behalf of STA and to undertake an informal trial of these questions with groups of between 30-100 children.

In January 2012, STA convened a number of panels to review the initial questions that had been developed including a teacher panel, a test review group and an inclusion panel (expert review 1[5]). As a result of the panels, questions were amended in preparation for the informal trial and the content domain was refined further.

3.2 Informal trial

In February 2012, an informal trial of test questions in development was undertaken. While the small number of children involved meant that any quantitative data for the questions had to be treated with caution, all children involved in the trial were interviewed and provided a rich source of qualitative data to inform development. Of particular importance was to find out why children omitted a response, for example because they had not been taught the curriculum content explicitly, or they had difficulty understanding the requirements of the question.

Reports were written on all test questions that were taken to trial. This allowed the questions to be categorised into the following groups:

- Questions that required further amendments;
- Questions that were to be removed from further consideration due to poor technical functioning; and
- Questions that were ready for the next stage in the process.

Questions that required further amendment were modified in line with evidence from trialling and expert review.

3.3 Expert review 2

A second round of expert review was conducted on the questions. In addition a number of experts who had not previously been involved in development were asked to review the materials.

Professor Debra Myhill, University of Exeter, and Ruth Miskin and Janet Brennan, both members of the English National Curriculum review team, were invited to review and comment on all materials in the light of both the current Key Stage 2 context and the new curriculum currently in development. A detailed report was produced by Professor Myhill. Separate meetings took place between STA and Professor Myhill, and STA with Ruth Miskin and Janet Brennan which informed trial booklet construction. All reviews raised some concerns with the nature of testing grammar out of context as well as the

[5] The expert review meetings are part of the STA's test development process. They involve teachers, headteachers, SEN Coordinators and other education professionals who review and provide feedback on test materials.

identification of some technical issues with the content being assessed. Many of the technical issues were addressed by STA's test development research team, although some of the issues were outside the scope of the review. However, the independent reviewers recognised that the requirements of the current National Curriculum, in addition to some of Lord Bew's recommendations, limited the test development team's ability to respond to all of the concerns expressed.

Following this final round of review, the questions were finalised for the technical pre-test.

3.4 Technical pre-test

The technical pre-test took place in June 2012. Due to the confidential nature of the materials, administration of this trial was carried out by visiting administrators appointed by the trialling contractor.

3.4.1 Levels 3-5

Twelve test versions were trialled in the technical pre-test. Each booklet comprised a section of short answer questions, a spelling task, a handwriting task and one other component (either handwriting or spelling).

The short answer questions (SAQ) for each booklet were constructed in blocks of 25 marks. Each block appeared in two of the 12 test booklets. The questions were divided between the 12 blocks. It was intended that the answer booklets would be roughly equivalent in terms of length and demand, as far as could be judged without any statistical data to inform the allocation process, and in terms of coverage of the different short answer questions. Each booklet had between 44 and 48 short answer marks, worth one or two marks. V01, V02, V03, V04 had 45 questions, V05 had 43 questions, V06 had 42 questions, and the remaining six booklets had 44 questions.

There were four different spelling tasks. Two of the tasks (S1 and S2) were constructed from independent sentences and two of the tasks (S3 and S4) were passages. S1 and S2 both contained 20 spellings, S3 contained 18 spellings and S4 contained 19 spellings. Additionally, the spelling tasks were trialled as a CD recording or read aloud by a test administrator.

Three handwriting passages were trialled, each with 6 sentences. Additionally, three different administration methods were used for handwriting: CD recordings for dictation; read aloud by an administrator for dictation; or a copied passage. For the CD recording and administrator-read methods the children were assessed on their spelling and punctuation as well as handwriting. Figure 1 shows each of the test combinations. All children in participating schools were allocated to one of the twelve test booklets.

Test booklet	Short answer question blocks	Handwriting task	Spelling task	Additional task	Number of children
TE3V01	SAQ1&2	HW1t	S1cd	S4cd	489
TE3V02	SAQ2&3	HW2t	S2t	S1cd	512
TE3V03	SAQ3&4	HW3t	S3cd	S4t	489
TE3V04	SAQ4&5	HW1cd	S4t	S2cd	514
TE3V05	SAQ5&6	HW2cd	S1t	S3t	508
TE3V06A&B	SAQ6&7	HW3cd	S2cd	HW3cy	502
TE3V07	SAQ7&8	HW1cy	S3t	HW3cd	515
TE3V08	SAQ8&9	HW2cy	S4cd	S1t	501
TE3V09	SAQ9&10	HW3cy	S3cd	HW1t	487
TE3V10A&B	SAQ10&11	HW1cd	S2t	HW1cy	513
TE3V11	SAQ11&12	HW2cd	S1cd	HW3t	498
TE3V12A&B	SAQ12&1	HW2t	S3cd	HW2cy	489

Note: t=test administrator read; cd=cd; cy=copy

Figure 1: Levels 3-5 test combinations

3.4.1.1 Levels 3-5 short answer question structure

Figure 2 shows the number of marks attributed to each short answer question reference code in each of the levels 3-5 test versions.

Reference code	SAQ 1	SAQ 2	SAQ 3	SAQ 4	SAQ 5	SAQ 6	SAQ 7	SAQ 8	SAQ 9	SAQ 10	SAQ 11	SAQ 12	All
Name and Identify	**8**	**8**	**8**	**8**	**8**	**9**	**9**	**7**	**8**	**9**	**7**	**7**	**96**
SG1 Grammatical terms/word classes	2	1	2	2	1	1	2	1	1	2	2	2	19
SG2 Features of sentences	1	1	1	0	1	0	1	0	1	0	0	0	6
SG3 Complex sentences	1	1	1	1	1	2	1	1	1	2	1	1	14
SG4 Standard English	1	2	1	2	2	2	2	2	2	1	1	1	19
P1 Punctuation	3	3	3	3	3	4	3	3	3	4	3	3	38
Grammatical Accuracy	**17**	**17**	**17**	**17**	**17**	**16**	**16**	**18**	**17**	**16**	**18**	**18**	**204**
GA1 Word classes	3	4	3	3	4	4	3	4	4	3	3	3	41
GA2 Features of sentences	1	1	1	2	1	2	1	2	1	2	2	2	18
GA3 Complex sentences	2	2	2	1	2	0	2	1	2	0	1	1	16
GA4 Standard English	4	3	4	3	3	3	3	3	3	4	4	4	41
GA5 Formal /Informal contexts	1	1	1	1	1	1	1	1	1	1	1	1	12
GA6 Punctuation	2	2	2	3	2	2	2	3	2	2	3	3	28
GA7 Vocabulary	4	4	4	4	4	4	4	4	4	4	4	4	48

Figure 2: Number of marks by short answer question reference code

Subsequent to the technical pre-test, it was decided that some questions were not appropriate to be taken forward and these were not included in the analysis. Four marks were lost from GA1, one mark from SG4 and 12 marks from GA4.

The number of marks available for each reference code above does not necessarily reflect the proportions required to construct an individual test in any one year of testing. Questions not used for the 2013 tests will remain in STA's question-bank and will be considered for future test cycles. Similarly, any areas of the curriculum not selected in this test due to limited question availability will be sampled in future tests following further rounds of question development.

3.4.2 Level 6

Three test versions were trialled. Each booklet contained an extended task , a section of short answer questions and a spelling task.

The extended tasks were selected to be similar in format to the former shorter writing tasks, although the mark allocation was adjusted to give prominence to the grammar, punctuation, and vocabulary elements. Tasks were also chosen to be functional and transactional in nature. Each task was marked out of a total of 16 marks. This was broken down as follows:

- sentence structure and punctuation (SSP) – six marks;
- text structure and organisation (TSO) - four marks;
- appropriacy and vocabulary (AV)– four marks; and
- handwriting (HW) - two marks.

It should be noted that, subsequent to the trial, a policy decision was taken to remove handwriting from the test domain.

The short answer questions contained at least 33 marks with no overlapping questions between booklets. The questions were divided between the three versions. It was intended to make the answer booklets roughly equivalent in terms of length and demand, as far as could be judged without any statistical data to inform the allocation process, and in terms of short answer question coverage. V1 had 24 questions and generated 34 marks. V2 and V3 each had 26 questions and 33 marks.

Three spelling tasks were trialled (all were teacher read) and each spelling task had 20 words to be spelled. In V1 and V2 the words were in themed passages, whereas, in V3 the words were in independent sentences.

Figure 3 shows each of the test combinations. All children in the participating schools were allocated two of the three booklets.

Test booklet	Extended task	Short answer question block	Spelling task	Number of children
TE6V1	W1	SAQ1	S1	1018
TE6V2	W2	SAQ2	S2	1082
TE6V3	W3	SAQ3	S3	1018

Figure 3: Level 6 test combinations

3.4.2.1 Level 6 short answer question structure

Figure 4 shows the number of marks attributed to each reference code in each of the Level 6 test versions.

Reference code	SAQ1	SAQ2	SAQ3	All
Name and Identify	**6**	**4**	**5**	**15**
SG1 Grammatical terms/word classes	3	3	2	8
SG5 Level of formality/informality	2	0	1	3
P1 Punctuation	1	1	2	4
Grammatical Accuracy	**29**	**31**	**30**	**90**
GA1 Word classes	4	4	4	12
GA3 Complex sentences	2	2	2	6
GA4 Standard English	3	2	2	7
GA5 Formal/Informal contexts	4	6	5	15
GA6 Punctuation	5	5	5	15
GA7 Vocabulary	11	12	12	35

Figure 4: Number of marks by SAQ

Subsequent to technical pre-test, it was decided that some questions were not appropriate to be taken forward and these were not included in the analysis. Two marks were lost from GA4, four marks from GA5, two marks from GA6 and three marks from GA7. Some questions also had marks reassigned so that two marks were added to SG5, one mark to GA1, one mark to GA3 and three marks to GA5.

Not all of the reference codes covered at levels 3-5 are included in the level 6 test as some subject content is no more difficult at this level. Similarly, there are some reference codes that relate to the more difficult content at Key Stage 2, or are sampled from the Key Stage 3 programme of study that appear at level 6 only. When trial booklets were constructed for the technical pre-test, the proportions of questions to be included in a live test were not yet confirmed. The reference code proportions for the trial booklets outlined above are therefore not necessarily reflective of a final live test booklet.

4 Analysis

4.1 Levels 3-5

4.1.1 Sample statistics

A sample of maintained and independent primary schools in England was drawn from the trialling agency's database of schools. It was stratified by school type, region and Key Stage 2 attainment (2011).

The achieved sample included 6017 children from 253 schools. Figure 5 shows the representativeness of the achieved sample. The North is over-represented in the sample and the Midlands under-represented compared to the population as a whole but the chi-square tests did not reveal a statistically significant difference. Junior schools are over-represented and independent schools are under-represented in the sample; the chi-square tests revealed statistically significant differences between the sample and the population. There were also statistically significant differences between the sample and population in terms of Key Stage 2 attainment.

This apparent lack of representation of the population is due to the self-selecting nature of the sample, in that it is difficult to control which schools agree to participate. Whilst the profile of the sample is not ideal in terms of representation of the stratification variables, there is no reason to suggest that the validity of the data on question performance has been compromised.

		Population		Sample	
		Count	%	Count	%
Key Stage 2 overall performance band 2011	Lowest 20%	2759	20.1	46	18.2
	2nd lowest 20%	2559	18.6	48	19.0
	Middle 20%	2279	16.6	47	18.6
	2nd highest 20%	2734	19.9	67	26.5
	Highest 20%	3011	21.9	44	17.4
	Assessing independents	360	2.6	1	0.4
	Not available	30	0.2	-	-
Primary school type	Primary/Combined	11579	84.3	211	83.4
	Junior	1228	8.9	31	12.3
	Middle	55	0.4	2	0.8
	Independent schools	360	2.6	1	0.4
	Other type	510	3.7	8	3.2
Region	North	4407	32.1	90	35.6
	Midlands	4153	30.2	71	28.1
	South	5172	37.7	92	36.4
Total		13732	100.0	253	100.0

Figure 5: Levels 3-5 sample representation at school level

Overall there was a slight under-representation of girls in the sample compared to the national population. This is reflected across the booklets, with seven of the 12 having a lower proportion of girls than boys. All but one of the booklets had lower proportions of children with English as an additional language (EAL) than is present in the national population. Overall the proportion of children with special educational needs (SEN) in the sample was very similar to the national population, however SEN children were over-represented in V04 (school action) and under-represented in V03, V05 and V07.

4.1.2 Short answer questions

The measure of internal consistency, coefficient alpha, was estimated to be between 0.86 and 0.91 on each of the booklets. The standard error of measurement across the 12 booklets was between 2.5 and 3 marks. This is in line with what would be expected from a test of this nature and indicates a reasonable level of reliability.

4.1.2.1 Administration time and section completion

Children were allowed 40 minutes to complete the short answer questions section for each of the 12 trial booklets. If children were still working after 40 minutes the administrator could allow children an extra 10 minutes to complete it, if this fitted in with the school time-table. All children were asked to fill in a start time box and then a finish time box when they had completed the test, or after the 40/50 minutes at the end of the test time. About 9 per cent of children did not complete the start and finish time boxes however the data gives a good indication of how satisfactory the administration time for the tests was. Figure 6 below shows the proportions of children completing the booklets after 35 minutes, the allowed 40 minutes, 45 minutes, 50 minutes, and after 55 minutes. In some cases the children appeared to have taken more than 50 minutes, or their time taken showed as a negative number. This may have been because they filled in the start and finish times incorrectly, or the data was captured incorrectly. These are not included in the table.

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	All
After 35 minutes	48	76	55	52	44	47	54	34	29	58	78	49	52
After 40 minutes	73	93	91	78	65	77	82	55	60	82	93	76	77
After 45 minutes	90	95	97	90	86	91	94	70	90	94	97	87	90
After 50 minutes	95	96	100	98	98	98	99	86	97	99	99	97	97
After 55 minutes	96	96	100	99	99	99	99	98	98	99	99	97	98

Figure 6: Short answer question section completion times (% of children)

It can be seen that 40 minutes did not appear to be sufficient time for most children to complete the short answer questions section. However most children who completed start and finish boxes were able to finish with an additional five minutes and almost all after an additional ten minutes. V08 had a lower proportion of children completing the section in the expected time. The booklets either side have common short answer questions: V09 also had lower proportions completing the section within 40 minutes but is similar to other booklets for 45 minutes and 50 minutes.

The booklets which had most children finishing after 40 minutes (V02, V03 and V11) were among the booklets with the most questions, and did not have the highest mean scores. This suggests that these booklets were not completed more quickly because the questions were easier.

The descriptive statistics on questions that were not reached suggests that in some instances children wrote the finish time at 50 minutes although they had not finished the test. This is possibly an artefact of the instructions given by the administrators. Looking at the not reached data for the penultimate question (since the not reached for the final question is also the omitted data), the booklets range from three per cent (V03) to 16 per cent (V04 and V09).

4.1.3 Spelling

The measure of internal consistency, coefficient alpha, was estimated to be between 0.89 and 0.92 on each of the tasks. The standard error of measurement was estimated to be between 1 and 2 marks.

4.1.4 Handwriting

The tasks were marked out of four for handwriting and other errors were identified using coding frames. All handwriting tasks that were administered by CD exhibited an increased number of children omitting sentences, compared to copying or teacher read administrations. Further exploration of dictation found that while there were 5358 credible dictation instances over all of the children in the technical pre-test, only 2272 were instances where all of the sentences were present. Of those only four per cent were without error. An analysis of background characteristics found that children with SEN performed significantly worse in handwriting than their non-SEN counterparts.

4.2 Level 6

4.2.1 Sample statistics

STA selected a sample of maintained and independent primary schools in England using a number of sources, including 2011 performance tables and level 6 test orders data, in order to include schools most likely to have children working at level 6. These schools were not intended to be representative of the national population.

The achieved sample included 1619 pupils from 174 schools. Figure 7 shows the representativeness of the achieved sample compared to the intended sample. Academy converters and independent schools are over-represented in the sample and community schools under-represented. Schools in the North West and West Midlands are over-represented; schools in London and the South East are under-represented. However, overall the sample can be seen to be broadly representative of the intended sample and chi-square tests revealed no significant differences between the sample and the population.

		Intended		Achieved	
		Count	%	Count	%
Primary school type	Academy converters	17	2.8	8	4.6
	Academy sponsor led	1	0.2	0	0.0
	Community school	219	39.0	62	35.6
	Foundation school	15	1.7	4	2.3
	Independent school	75	9.5	25	14.4
	Voluntary aided school	209	31.9	52	29.9
	Voluntary controlled school	70	14.7	23	13.2
Region	North East	21	3.5	6	3.4
	North West	96	15.8	34	19.5
	Yorkshire and The Humber	54	8.9	16	9.2
	East Midlands	59	9.7	18	10.3
	West Midlands	57	9.4	22	12.6
	East of England	49	8.1	14	8.0
	London	108	17.8	21	12.1
	South East	106	17.5	27	15.5
	South West	56	9.2	16	9.2
Total		605	100	174	100

Figure 7: Level 6 sample representation at school level

The gender profile of children across the booklets is very similar, with V1 having a slightly higher proportion of girls than the other booklets. The samples are very similar in terms of EAL and SEN, with few SEN children as would be expected in a sample of children working at this level.

4.2.2 Extended task

The measure of internal consistency, coefficient alpha, was estimated to be above 0.80 on each of the booklets. The standard error of measurement across the three tasks was approximately one mark.

4.2.3 Short answer questions

The measure of internal consistency, coefficient alpha, was estimated to be between 0.64 and 0.67 on each of the booklets. There are a number of reasons why the internal consistency for level 6 might be lower here related to whether the pupils in the trials were actually working at level 6 and the fact that the test covers a relatively small difficulty range. These will be investigated further when data from the live administration is available. The standard error of measurement across the three tests was between 2 and 3 marks.

4.2.3.1 Administration time and short answer questions completion

Children were allowed 20 minutes to complete the short answer questions section for each of the three trial booklets. If children were still working after 20 minutes the administrator could allow them an extra 10 minutes to complete it, if this fitted in with the school time-table. All children were asked to fill in a start time box and then a finish time box when they had completed the section, or after the 20/30 minutes at the end of the time given for the SAQ section. About 11 per cent of children did not complete the start and finish time boxes; however the data gives a good indication of how appropriate the administration time for the tests was. Figure 8 shows the proportions of children completing the booklets after 20 minutes, 25 minutes, 30 minutes, and after 35 minutes. In some cases the time taken was shown as a negative number. This may have been because children filled in the start and finish times incorrectly, or the data was captured incorrectly. These are not included in the table.

	V01	V02	V03
After 20 minutes	11	19	14
After 25 minutes	40	66	54
After 30 minutes	94	97	99
After 35 minutes	98	99	100

Figure 8: Short answer questions section completion times (% of children)

It can be seen that 20 minutes did not appear to be sufficient time for most children to complete the short answer questions section, and only around half the children finished with an additional five minutes. However, almost all children who completed start and finish boxes were able to finish with an additional ten minutes for V2 and V3. For V1, 35 minutes were needed for almost all children to finish.

The descriptive statistics on questions that were not reached suggests that some children wrote the finish time at 35 minutes although they had not finished the test. This is possibly an artefact of the instructions given by the administrators. Looking at the not reached data for the penultimate question (since the not reached for the final question is also the omitted data), V1 is 16 per cent, V2 is seven per cent and V3 is 12 per cent.

4.2.4 Spelling

The measure of internal consistency, coefficient alpha, was estimated to be between 0.78 and 0.84 on each of the booklets. The standard error of measurement across the three versions was between 1.5 and 2 marks.

4.3 Decisions following on from the analysis of trialling data

As a result of the analysis of the trialling data, aspects of the test have been amended to ensure valid and reliable outcomes.

4.3.1 Short answer questions

In the live levels 3-5 test, children will be given 45 minutes to complete 50 marks worth of short answer questions. For level 6, children will be given 20 minutes to complete 21 marks of short answer questions.

4.3.2 Spelling

The live administration of spelling will be teacher read sentences. This is a familiar Key Stage 2 administration model as it was formerly a part of the writing test. Both teachers and administrators raised concerns about the accessibility of CDs, in particular a concern that they were ‘too fast’. Additionally, the teacher read sentences represent better value for money than CD production. Finally, some children reported that the passages’ contexts could be distracting. Sentences are more efficient to construct, and words can be placed in order of difficulty more easily, in line with Lord Bew's recommendations.

4.3.3 Handwriting

Assessment experts raised a number of concerns about the proposed assessment of handwriting related to the authenticity of the tasks. A copying task enables the child to produce their best handwriting but this is not necessarily representative of their usual handwriting. A dictation exercise, on the other hand, often produces the child's worst handwriting because of the high cognitive demand involved in the short-term memory requirements as well as ensuring accurate spelling and punctuation. Neither task therefore gives a true representation of the child's general handwriting over a number of different tasks for a number of different purposes.

In the technical pre-test, teachers and administrators felt that the CD dictation was too fast and some children struggled to keep up. Evidence from trialling showed that, dictation disproportionately disadvantaged children with special educational needs, which is probably because of the short term memory requirements of the task. Handwriting was originally designed to provide a very small number of marks to the overall score and did not provide enough additional information on children's ability to warrant the expenditure of time and effort.

As a result, the policy recommendation was to assess handwriting by teacher assessment over a large sample of children's work and therefore it was removed from the test.

5 Test framework[6]

As a result of the evidence from the development process, the test framework was written. A test framework contains details of:

- the content domain that will be covered in the test;
- the cognitive processes associated with the measurement of the construct of grammar, punctuation, vocabulary and spelling;
- the test specification by which valid, reliable and comparable tests can be constructed year on year; and
- the measures that have been put in place in order to improve accessibility for all children and to minimise bias affecting particular groups.

5.1 Content domain

As detailed in the test development section of this report, the content domain was adapted and refined throughout the development process. All elements of the content domain are directly linked to statements in the Key Stage 2 National Curriculum (1999) at levels 3-5. Consistent with the level 6 tests available in mathematics and English reading, some content is drawn from the Key Stage 3 National Curriculum for English (2007) in the level 6 test only.

5.2 Cognitive domain

A review of relevant literature was undertaken by assessment researchers from STA, investigating the possible approaches to defining a cognitive domain for the test. Following this initial research, two specific models for the cognitive domain were investigated further for application to the tests:

- Single dimension hierarchies, e.g. Bloom’s taxonomy[7]; and
- Multi-dimensional models, e.g. Scale of Cognitive Demands[8].

The suitability of each approach was considered, with the following principal conclusions:

[6] http://www.education.gov.uk/schools/teachingandlearning/assessment/keystage2/b00218030/gps-sample-materials

[7] Bloom, B. S. (Ed.). Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. (1956). Taxonomy of educational objectives: Handbook I: The cognitive domain. New York: David McKay.

[8] Hughes S., Pollitt A. and Ahmed A. (1998) “The development of a tool for gauging the demands of GCSE and A-level questions” presented and published at BERA meeting August 27-30 1998.

	Advantages	Disadvantages
Single-dimensional hierarchies	▪ Representative of the key cognitive processes children go through in demonstrating skills in grammar, punctuation and spelling. ▪ Would allow clear specification of requirements for question writing and test construction. ▪ Only one classification required per question, therefore a quick and straightforward way to classify questions. ▪ Widely used in classroom practice and therefore familiar to the wider education community.	▪ Allows consideration only of a single dimension, without consideration of other cognitive demands and so cannot reasonably provide a valuable representation of overall demand in a test question.
Multi-dimensional models	▪ Widely used elsewhere in assessment, particularly at GCSE. ▪ By considering 'demand' more comprehensively (i.e. using multiple dimensions), the ratings can provide a very useful measure of predicted difficulty. ▪ A more nuanced approach to predicting overall demand (and therefore difficulty) will be particularly valuable in light of: ▪ the future removal of National Curriculum levels, requiring STA to develop an alternative, consistent way to predict and manage question and test difficulty; ▪ lessons learnt about the management and quality control of questions from external question writing agencies; and ▪ the need to ensure appropriate increases in demand between levels 3-5 and level 6.	▪ Classification of questions is more time consuming. ▪ The ratings for each dimension are less immediately meaningful as they are just numbers that represent extent of demand rather than word labels such as 'recall' or 'evaluation'. ▪ It may be more complex to use scales such as this to state desirable cognitive demand properties/proportions in tests or for question writers.

Figure 9 Cognitive domain models

After consideration of the above strengths and weaknesses, it was decided that a multi-dimensional model was most appropriate for the tests. The existing Scale of Cognitive

Demands[9] was examined in detail, together with associated literature and alternative multi-dimensional models.

The following process was used in order to develop the scale.

Developing a Cognitive Domain for an Assessment

A multi-dimension scale

- Determine the factors that make one question more or less demanding than another. This requires thought and discussion, as well as investigation of subject-specific literature.
- Determine the number of dimensions required. Too few may not adequately encapsulate the demands in the assessment, but too many becomes unmanageable.
- Write descriptions for each dimension explaining what constitutes low demand and what constitutes high demand in each dimension.
- Attempt to rate questions. It is crucial that there is a high level of agreement between different experts when rating questions.

Having followed the process outlined above and rated a number of test questions on the existing Scale of Cognitive Demands, it became apparent that two refinements to this model would be beneficial for this test:

- The strands of ‘strategy’ and ‘resources’ were so closely linked that, for questions considered for this test, the two strands could be combined.
- The ‘complexity’ strand should preserve a direct link to Bloom’s taxonomy, as it had already been judged a meaningful and useful scale, representative of the key cognitive processes children go through in demonstrating skills in grammar, punctuation and spelling.

As a result, the final cognitive domain agreed for the test classifies questions with ratings across four dimensions in order to arrive at an overall judgement of their cognitive demand:

- cognitive level;
- response complexity;
- abstraction rating; and
- strategy support rating.

These dimensions are presented in a tabular format in the test framework and examples are included. For the avoidance of issues with test security, illustrative question stems (rather than whole questions) were included. When the frameworks were shared with a

[9] Hughes S., Pollitt A. and Ahmed A. (1998) “The development of a tool for gauging the demands of GCSE and A-level questions” presented and published at BERA meeting August 27-30 1998.

review group of teachers and other stakeholders, these illustrative examples were particularly well-received.

5.3 Minimising bias

A number of processes are undertaken during test development in order to minimise bias in the questions. However, it is not possible to exclude all forms of bias from within questions without compromising the validity of the test and excluding elements of the content domain.

As a result, access arrangements should be used to further minimise bias for appropriate groups of children. Given the nature of the test and the domain being assessed, it was agreed that the full range of access arrangements applicable to Key Stage 2 assessments will be available to eligible children. This includes the use of a reader, which is not allowed in the English reading test, because the test is not an assessment of reading. It also includes the use of a scribe; although where spelling is assessed in the test, the spelling must be the child’s own.

Some children with profound hearing impairment who do not use sign supported communication will still be unable to access the spelling test. For these children a compensatory mark will be awarded based on the mean average scores on the live test. This is consistent with the previous spelling assessment in the English writing test

5.4 Test specification

A full test specification for both the levels 3-5 and level 6 tests can be found within the test framework document. The test specifications have been used to develop the 2013 tests and will continue to be used for as long as we assess the 1999 Key Stage 2 National Curriculum for English at levels 3-5 and level 6. The level 6 test additionally samples content from the 2007 National Curriculum for English at Key Stage 3.

The test specification was developed to meet Lord Bew’s recommendations, to ensure a valid assessment of the National Curriculum and reflect all of the evidence from trialling. The following section of the report explains the rationale for the details of the specification.

5.4.1 Timings and total marks

As discussed in the trialling section of this report, an analysis of the time taken for children to complete the test was undertaken on trialling data. In addition, in order to maintain manageability for schools, it was decided that total testing time for the test should not exceed that for the English writing test that it replaced in the test timetable.

Whilst ensuring children have sufficient time to answer the questions in the test is important, it is also important that the test contains sufficient questions to produce a reliable result. The number of questions in the pre-test produced outcomes with sufficient levels of reliability for a test of this type and therefore the total number of marks in the live test needed to be similar.

As a result, both the levels 3-5 and level 6 tests have been designed to be administered in approximately one hour with the number of marks per component as shown in the table below.

Component	Description	Timing of component	Number of marks
Short Answer Questions	Short answer questions on grammar, punctuation and vocabulary (selected and constructed response) presented in order of difficulty	45 minutes	50 marks
Spelling	20 sentences from which targeted spelling words have been removed	Around 15 minutes (not strictly timed)	20 marks
Total		1 hour	70 marks

Figure 10 Levels 3-5 test format

Component	Description	Timing	Number of marks
Extended task	An extended response to a writing prompt through which children are able to demonstrate precision, choice and accuracy of punctuation, syntax and vocabulary when writing	30 minutes	SSP 6 marks TSO 4 marks AV 4 marks 14 marks
Short Answer Questions	Short answer questions on grammar, punctuation and vocabulary (selected and constructed response) presented in order of difficulty	20 minutes	21 marks
Spelling	15 sentences from which targeted spelling words have been removed	Around 10 minutes (not strictly timed)	15 marks
Total		1 hour	50 marks

Figure 11 Level 6 test format

5.4.2 Proportion of marks

The proportion of marks awarded for the different aspects of the tests reflect the relative status of those aspects within the programmes of study and the defined domain for this test. Grammar and punctuation have the largest proportion of the marks, followed by spelling and vocabulary respectively.

Spelling forms a greater proportion of the total mark (29 per cent) than was the case in the previous writing tasks (14 per cent). This is because the content domain of the test forms only a small part of that which was assessed in the legacy English writing tests and spelling makes up a greater part of the new domain.

Vocabulary is sampled from the Reading attainment target and covers word meaning only, so it has a proportionately low allocation in any one test.

The tables below show the proportion of marks assessing each element of each component.

Component	**Element**	**Number of marks**	**Proportion of total mark (%)**
Short Answer Questions	Grammar	25-35	36-50
	Punctuation	10-20	14-28
	Vocabulary	5-10	7-14
Spelling	Spelling	20	29
Total		70	

Figure 12 Proportion of marks for the levels 3-5 test

Component	**Element**	**Number of marks**	**Proportion of total mark (%)**
Extended task	Grammar	8	16
	Punctuation	2	4
	Appropriacy and Vocabulary	4	8
Short Answer Questions	Grammar	10-15	20-30
	Punctuation	5-10	10-20
	Vocabulary	1-5	2-10
Spelling	Spelling	15	30
Total		50	

Figure 13 Proportion of marks for the level 6 test

The reference codes given in the test framework detail the specific content to be tested within the short answer question section of the tests. The test will sample from this content in any given year. Although each element may not be included within each test, the full range of content detailed in this document will be assessed over time.

5.4.3 Spelling

As mentioned in the trialling section, following the technical pre-test it was decided that the spelling task should consist of a number of separate sentences read aloud by the test administrator, in order to align with Lord Bew's recommendations related to the placement of questions in order of difficulty in the test.

At levels 3-5, the words assessed are selected to take account of children's developing ability to accurately spell a wide range of words and include a balance of common, polysyllabic words, polysyllabic words that conform to regular patterns and words with complex, regular patterns.

At level 6, marks for spelling were previously awarded within the extended writing task in the legacy Writing tests. In the English grammar, punctuation and spelling test, children's ability to use ambitious and precise vocabulary is tested in the extended task. Test development research showed it was fairer to test children's spelling in a separate task, similar to that at levels 3-5, so that they could focus on choosing appropriate and precise vocabulary when writing.

As the level 6 test discriminates at a single-level only, it was felt that 15 words were adequate to test children rather than the 20 required across levels 3-5 in the levels 3-5 test. Words are chosen that demonstrate the spelling strategies required for lower-frequency, less familiar words.

5.4.4 Proportion of question types and test format

The Bew review stated that the test should include questions '… where there are clear 'right' and 'wrong' answers, which lend themselves to externally-marked testing'. In the government's response, STA was asked to research international tests of 'language arts' and to introduce a test of this nature. This set the parameters for the nature of the test and informed the style of questions test development researchers were able to consider.

The short answer question section of the levels 3-5 and level 6 tests are categorised into two broad formats:

- Selected response, requiring selection of the correct answer.
- Constructed response, requiring the child to write a short answer of their own within a specified format.

There are a number of areas of the domain where a constructed rather than selected response was necessary in order to fully test the content domain. While it is still possible to have right and wrong answers in a constructed response question in line with the requirements of the test, a decision was taken in both the levels 3-5 and level 6 tests to set a limit on the number of constructed response questions that can appear in any one test, although more constructed response questions can appear at level 6.

In the short answer question section, therefore, most responses will require only a tick, circle, line or very short written response. Some test questions do require a full sentence to be written but these will usually be placed towards the end of the paper in order to allow children every opportunity to gain more straightforward marks quickly. Consistent with Lord Bew's recommendations for test organisation in relation to the English reading test, questions in this component will, as far as possible, be placed in '…clear order of difficulty'[10]. The difficulty of individual questions is determined quantitatively through pre-testing.

The proportions of each question type are described below.

Question type (short answer questions)	Approximate proportion in component (%)
Selected response	70-80
Constructed response	20-30

Figure 14 Format of questions for the levels 3-5 test

Question type (short answer questions)	Approximate proportion in component (%)
Selected response	60-70
Constructed response	30-40

Figure 15 Format of the questions for the level 6 test

These formats are further categorised into the following sub-types and examples of each are provided in the test framework.

Question type	Rubric sub-type
Selected response	'Identify…'
	'Match…'
Constructed response	'Complete/correct/rewrite…'
	'Find and write…'
	'Explain…'

Figure 16 Sub-types of question

5.4.5 The extended task at level 6

While STA understood Lord Bew's recommendation that writing composition should be subject to teacher assessment only, the National Curriculum level descriptor at level 6 focuses on children's ability to make accurate and precise choices from a range of punctuation marks, sentence structures and vocabulary. A short extended writing task has been included at level 6 only in order to allow children to demonstrate these level 6 skills. The prompt will identify a clear purpose, audience and format for writing; it will allow children to demonstrate their ability to use a range of precise, accurate and appropriate punctuation, syntax and vocabulary.

The mark schemes for this task are different from those used in the legacy English writing tests with more marks awarded for sentence structure and punctuation (SSP – 6 marks),

[10] Bew et al., 2011, Independent review of Key Stage 2 testing, assessment and accountability

and fewer for text structure and organisation (TSO – 4 marks) and linguistic appropriacy and vocabulary (AV – 4 marks). Composition and effect (including viewpoint) is not assessed as this is recognised as being the domain of teacher assessment. The extended task is designed to elicit more functional and transactional writing, where more formal standard English is appropriate.

6 Test construction

In addition to the classical test theory analysis presented in the previous chapter, item response theory (IRT) was carried out on the short answer questions and spelling words. One, two and three parameter models were tested to determine best model fit using AIC and BIC[11] goodness of fit indices. The two parameter graded response model showed the best fit and will be used going forward as the IRT model for the test.

Accurate analysis using item response theory requires three assumptions to be tested and met: model fit (as tested above), local independence and unidimensionality. Local independence was assessed with the Q3[12] statistic, separately for the short answer questions and the spelling words. In both instances the Q3 statistic was very close to zero, indicating the assumption of local independence was upheld. Unidimensionality was tested by examining the scree plots from exploratory factor analysis. In all cases the scree plots indicated the presence of a single factor. Hence, the assumption of unidimensionality was also upheld. Taken together, these tests indicate the use of the two parameter graded response model is an appropriate model for the data. Therefore, it can be used to inform test construction, as well as the setting and maintaining of standards of the test. The Q3 values and scree plots can be found in Annex 2.

Based on all of the evidence presented from expert review and technical pre-test, STA's test development researchers examined the questions to determine which were appropriate to continue to question selection.

Once the questions were selected, a test construction algorithm was run to maximise the information function across the ability range, while adhering to the test specifications. For each of levels 3-5 and level 6, it was necessary to produce two equivalent tests from the data, one for the live test in 2013, the other to act as an anchor for maintaining standards going forward. Particular constraints were taken into account such as avoiding questions that would clash due to content overlap (called 'enemies') whilst producing tests that adhere to percentages of marks for particular content and question type, as noted in the test specifications for levels 3-5 and level 6.

Once four tests were built to the above specifications, STA's test development researchers, assessment researchers and psychometricians met to agree the final content of the tests. This provided an opportunity to interrogate the algorithm and determine if there were subject-specific issues that needed to be addressed that the algorithm could not account for (for example, previously uncategorised enemies, the presentation of questions on the page and their relative positions in the test).

[11] AIC stands for Akaike information criteria, BIC stands for Bayesian information criteria. Both are measures of the relative goodness of fit of a statistical model.

[12] Yen, W.M. (1984) 'Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model' in Applied Psychological Measurement, 8, 125-145.

6.1 Setting and maintaining standards

6.1.1 Standard setting

As with any new assessment, a full standard setting exercise will be carried out following the first administration of the test as it requires data from children who have taken part in the live test. A standard setting procedure is required to facilitate decisions regarding the placement of thresholds for the new tests that translate a description of expected performance for children working at each National Curriculum level into the score range required to achieve that level. Once the standard has been set in the first year of the tests different processes will be used to maintain that standard for future years.

The performance descriptors will be developed by curriculum and assessment experts and will be validated by groups of teachers prior to the administration of the first test. Once developed, these performance descriptors will be made available to schools in March 2013. The performance descriptor for each level will describe what is expected of a minimally competent child at that level in order for participants to make judgements during the standard setting exercise.

For the levels 3-5 test, a bookmark procedure will be used to set the standard. For the level 6 test a combination of the bookmark procedure and the body of work procedure will be used due the slightly different nature of the test. Further detail of these two methods can be found in Annex 3. The standard setting exercises will be carried out twice with two independent groups of teachers in order to validate the approach and outcomes. If the outcomes from the two groups are similar, an average of the outcomes will be taken in order to derive the final thresholds. If the outcomes are not similar, further work will be undertaken to determine the final thresholds.

6.1.2 Maintaining standards

In subsequent years, standards will be maintained through the process of equating. The anchor test questions will be used alongside the live analysis and the technical pre-test analysis to ensure standards are maintained accurately and in line with best practice.

7 Reporting

In all previous National Curriculum tests, the reported outcome for each child has been a National Curriculum level as defined by the National Curriculum level descriptors for each attainment target. However, this test does not align fully with any single attainment target and therefore it is not strictly possible to determine a National Curriculum level as a result.

Figure 9 highlights the key features from the National Curriculum level descriptors that are relevant to the tests and that have been taken into consideration (alongside the associated sections of the programmes of study) when defining test content relevant to the levels 3-5 and level 6 tests. Although drawn largely from the attainment target En3: writing, the test does not fully cover those aspects of compositional writing that are subject to teacher assessment and also includes elements of En1: speaking and listening and En2: reading.

NC level	Extracts from level descriptors
3	**Vocabulary**: sequences of sentences extend ideas logically and words are chosen for variety and interest. **Sentence grammar**: the basic grammatical structure of sentences is usually correct. **Spelling**: spelling is usually accurate, including that of common, polysyllabic words. **Punctuation**: punctuation to mark sentences - full stops, capital letters and question marks - is used accurately.
4	**Vocabulary**: vocabulary choices are often adventurous and words are used for effect. **Sentence grammar**: pupils are beginning to use grammatically complex sentences, extending meaning. **Spelling**: spelling, including that of polysyllabic words that conform to regular patterns, is generally accurate. **Punctuation**: full stops, capital letters and question marks are used correctly, and pupils are beginning to use punctuation within sentences.
5	**Vocabulary**: vocabulary choices are imaginative and words are used precisely. **Sentence grammar**: sentences, including complex ones, and paragraphs are coherent, clear and well developed. **Spelling**: words with complex regular patterns are usually spelt correctly. **Punctuation**: a range of punctuation, including commas, apostrophes and inverted commas, is usually used accurately.
6	**Vocabulary and sentence grammar**: pupils experiment with a range of sentence structures and varied vocabulary to create effects. **Spelling**: spelling, including that of irregular words, is generally accurate. **Punctuation and structure**: a range of punctuation is usually used correctly to clarify meaning, and ideas are organised into well-developed, linked paragraphs.

Figure 17: Extracts from English level descriptors

As a consequence, the information provided by the test on children’s performance will be used to report a result that is indicative of a child working at a particular level. Although this may be incorrectly interpreted as a level in English grammar, punctuation and spelling, it is not believed that this technical difference will cause any particular problems. It is anticipated that the shorthand for describing the outcome from the test as a level will be widely used.

8 Validity studies

There were five research studies conducted to examine the validity of the test. Ofqual's regulatory framework for national assessments (2011[13]) states that an assessment should 'generate outcomes that provide a valid measure of the knowledge, skills and understanding that the learner is required to demonstrate as specified by the assessment objectives'. Therefore, the studies were designed to provide evidence to support the validity argument about the inferences that can be made about the outcomes of the test.

Two studies examined construct irrelevant variance and the factors that appear to modify the difficulty of the questions. A further study investigated the factors that might affect children with SEN as they interact with the test. The final two studies examined the marker agreement of the questions and the stability of the outcomes in a test-retest/alternate forms context.

Each study is examined in the sections below through background information, the methodology used in the investigation, the high-level outcomes of the study and a summary of the steps that were taken in the test development process to address the outcomes. Full reports for the studies on construct irrelevance, modifiers of difficulty and children with SEN can be found in the Annexes.

8.1 Construct irrelevant variance

Construct irrelevant variance (CIV) exists when a test contains excess, reliable variance that is irrelevant of the interpreted construct (Messick, 1989[14]). This means that the test is measuring something beyond what was intended; for example, if the construct being assessed was mathematical ability but some of the questions unintentionally measured reading ability in addition to mathematical ability.

8.1.1 Methodology

Researchers from the STA visited ten schools to observe the administration of trial questions. During these visits the researchers observed the administration to determine if any elements might have caused construct irrelevant variance and interviewed children and teachers who had been involved in the trial. The interviews with children took place immediately following the test, and were designed to highlight any questions that were easier or more difficult than intended for construct irrelevant reasons.

After the trial, STA researchers examined responses to each of the levels 3-5 short answer questions. Selected children's responses were examined, but more emphasis was placed on less structured question types where there was more information on children's thinking.

[13] http://www2.ofqual.gov.uk/files/2011-regulatory-framework-for-national-assessments.pdf
[14] Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.). Washington, D.C.: American Council on Education.

STA researchers were able to make use of the qualitative evidence available from the question writing agency's report on question performance in a small scale trial prior to undertaking the research.

8.1.2 Outcomes

There are a number of potential sources of CIV that need to be considered when developing and constructing tests. Only a small number of questions with potential CIV were uncovered in this study, suggesting that the vast majority of the questions in the trial were effective in assessing the intended construct with little or no 'noise'. Further detail can be found in Annex 4.

In short answer questions, STA's test development researchers need to make sure that:

- examples do not provide undue support in answering questions;
- questions do not include unfamiliar language or context that is not critical to what the question is trying to measure;
- questions are clear about what is being asked of the child; and
- layout and format of questions does not impede children in providing their responses.

The mode of administration in the spelling task uncovered some potential sources of CIV:

- clear and explicit guidance should be provided to administrators of the spelling task to ensure administration is as standard as possible for all children. This should include:

    - information about the amount of time required between each word in the sentence(s);
    - whether the administrator should wait until all children finish writing;
    - guidance about how to avoid unintentionally providing clues to the spelling when reading the word; and
    - guidance about how to avoid missing a spelling or mispronouncing words.

As mentioned previously, the handwriting component has been excluded from the levels 3-5 test, however Annex 4 provides information on the issues that were found with the different modes of administration and format. There does not appear to be any suggestion that the issues were sources of potential CIV.

8.1.3 Summary

Only a small number of questions with potential CIV were uncovered in this study, suggesting that the vast majority of the questions in the trial were effective in assessing the intended construct with little or no 'noise'. Key findings from this study were used to inform question selection for the 2013 live test and will inform question writing for future test cycles. Judgements regarding CIV inevitably include some level of subjectivity. It is for subject specialists to determine whether the issues raised lead to questions that are flawed and inappropriate for use, or whether, for example, the extent of a CIV issue is minimal and not of significant concern. Test development researchers and question writers will concentrate on identifying these potential sources and proactively make these decisions.

8.2 Modifiers of difficulty

This section provides a brief outline of the qualitative and quantitative analyses that have has been undertaken to identify the main factors that moderate the difficulty of test questions in the short answer questions section of the test. Full detail can be found in Annex 5.

There are three main purposes for this work:

1. Understanding the factors that moderate the difficulty of test questions provides important validity evidence. In particular, it may help to expose construct irrelevance issues. This is because once the factors that appear to affect question difficulty have been identified the construct relevance of those factors can be adequately considered.
2. To provide question writers and test developers with useful information regarding the construct relevant (and irrelevant) means by which question difficulty can be intentionally manipulated.
3. To aid question writers in making reasonable predictions about question difficulty when writing test questions in the future.

8.2.1 Methodology

Three sources were used in the process of identifying the factors appearing to affect the difficulty of questions. These are listed below.

- Responses to each of the short answer questions in the levels 3-5 technical pre-test were analysed, including at least 50 responses to each of the questions in levels 3-5 version 1. While analysis of children’s responses can be a valuable tool in understanding the sources (both invalid and valid) of difficulty of test questions, the method is of somewhat limited value in selected-response questions. This is because children’s responses to these questions tend to be limited to ticks, lines or other non-textual indications, which usually tell us little about the way in which children interpret test questions. In some of the less structured question types, more responses were analysed, as more could be gauged about children’s thinking in these questions.
- Approximately 40 children were interviewed immediately after taking the tests. A number of themes, relating to both difficulty and construct relevance, emerged from the analysis of these interviews. Evidence from interviews was also used to identify factors affecting question difficulty.

- The agency responsible for the administration of the Item Validation Trial (which took place in January 2012) produced a detailed report[15] on the performance of questions in the trial. While sample sizes were small, the authors were able to make use of the qualitative evidence available.

8.2.2 Outcomes

Eighteen factors were found to affect the difficulty of the trial test questions. These were further distilled into three higher-order factors: knowledge of technical language and punctuation rules, response strategy, and sentence complexity. While ‘Response strategy’ can also be considered a cognitive demand (with a higher rating on the factor implying increasing cognitive demand), the same cannot be said for KTL and KPR, as increasing the difficulty of this factor does not imply an increase in cognitive demand. Rather, it means that it is anticipated that pupils are less likely to possess the knowledge required to answer. For example, pupils may be less likely to be able to identify a preposition than a noun (perhaps because the latter tends to be taught earlier and more frequently), but this difference in difficulty would not reflect a difference in cognitive demand.

- *Knowledge of technical language (KTL) and punctuation rules (KPR)* refer to the extent to which knowledge of the meaning of technical terms and knowledge of relevant punctuation rules are required for children to answer questions correctly. The difference between performance on questions requiring KTL and those not requiring KTL, when focusing on the same word class, was often marked. More specifically, questions assessing use of grammar, but requiring no KTL, tended to be considerably easier than those that did not assess use of grammar and did require KTL. This issue does not appear to apply to KPR.
- *Response strategy* is affected by the complexity, explicitness and familiarity of the response requirements. The few questions that were classified as ‘high’ on this factor tended to be constructed response questions requiring children to generate their own language, or less familiar question types with a less explicit response strategy.
- *Sentence complexity* refers to the difficulty of the target phrases or sentences. It can be manipulated in a number of ways, such as increasing the:
    - length of sentences;
    - number of related sentences (i.e. those that constitute a ‘passage’ rather than discrete sentences);
    - abstractness of the ideas within phrases/ sentences; or
    - difficulty of vocabulary within sentences.

Forty-five trial questions were rated on a three-point scale against these factors to form a question difficulty scale. The relationship between the question facility values (based on a

[15] This report cannot be released at this time as it contains confidential information about questions that will be used in future tests.

sample of 489 children) and the ratings was analysed. In each case, there was a strong relationship evident. KTL showed the highest correlation (-0.66), with correlations of -0.40 and -0.32 for response strategy and sentence complexity respectively. That is, the less technical knowledge, response strategy or sentence complexity that a question exhibits, the more likely that it will have a higher facility.

A regression analysis was also undertaken. The adjusted R Square for the regression is 0.61, meaning that the 3-factor model of KTL, response strategy and sentence complexity explains about 61 per cent of the variability in the difficulty of questions. The correlations, along with the R square value, suggest that the three factors are able to explain a reasonable proportion of the difficulty of the questions.

8.2.3 Summary

Two key implications of this study were identified.

8.2.3.1 Construct relevance of the three factors

While the construct relevance of KTL and KPR cannot be questioned (assuming that the knowledge being assessed is included within the content domain), the construct relevance of the other two factors, which are not directly related to the test content, were deemed to be worthy of further consideration.

In the case of response strategy, a fundamental question to consider is the extent to which children’s performance for some question types is affected by difficulties in understanding the task requirements, rather than because they lack the subject knowledge to answer. Only two of the 45 trial questions were given the highest ratings for this factor, suggesting that only a small number of question types may carry this concern. It may be of value in the near future to conduct some further small-scale trialling, incorporating pupil interviews, of any question types where there is uncertainty regarding children’s understanding of task requirements. Including such questions in sample materials for schools may be effective in ensuring children have an understanding of requirements in those questions.

As with response strategy, only two of the trial questions examined (different questions from those flagged in response strategy) were given the highest rating for sentence complexity. The key question is whether this increased demand is construct-relevant. In other words, under which conditions, if any, is it valid to manipulate question difficulty by increasing sentence complexity? Fundamentally, this is a decision that must be made by subject experts. It may be worth providing specific guidance to question writers regarding what level of sentence complexity is appropriate.

8.2.3.2 The effect of practice on test performance

The factor that was found to correlate best with (and explain the variation in) question facility is KTL/KPR. The body of knowledge within the subject is not vast, and it is anticipated that once children are specifically taught this content, performance on questions assessing KTL in particular will increase, perhaps markedly. Questions assessing *use* of grammar tended to be easy, so it may be that children's performance will increase markedly in the tests (even relative to the 'standard' improvements in test performance seen after the introduction of a new assessment) in the next few years, with most questions displaying very high facilities. When the assessment becomes a competency based test, with one threshold which the majority of children are expected to achieve, then this is not a concern. However, while the purpose of the assessment remains to discriminate adequately between a broad range of ability (such as the case in the current levels 3-5 tests in mathematics and English reading), then consideration may need to be given for how this can be achieved in future test cycles.

8.3 Identified issues for children with special educational needs

This section reports on the potential accessibility issues of questions, tests and administration methods for children with SEN that were found in the test questions that were trialled in the summer of 2012. For the purposes of this research the special educational needs identified and investigated were visual impairment, hearing impairment and dyslexia. Considering the requirements of these groups was felt to ensure coverage the majority of issues for the test that are different from other National Curriculum tests.

A brief overview of the methodologies used and a summary of the outcomes precede a section on how this information was used in the test construction process and will be used in future test development cycles. The full report can be found in Annex 6.

8.3.1 Methodology

A high level literature review was conducted to identify research on technical English accessibility, skills and testing for children with visual impairments, hearing impairments and dyslexia. The literature review can be found in Annex 6.

Researchers from STA interviewed experts in visual impairment, hearing impairment and dyslexia. Interview questions focused on identifying general issues with the concept of testing grammar, punctuation and spelling with children with different educational needs.

Researchers from STA conducted small scale trialling of some of the test questions in specialist secondary schools[16]. After the children attempted questions from each section they were interviewed by the researchers. Booklets from the trial that took place in June 2012 were used with the children with hearing impairments and the children with

[16] As the research was undertaken in the Autumn term it was necessary to involve children in secondary schools in order that they were as close in age as possible to those who would take the tests at the end of year 6.

dyslexia. A selection of questions was converted into braille and modified large print for the children who were blind or partially sighted.

8.3.2 Outcomes

A number of considerations need to be undertaken to ensure that children with special educational needs can demonstrate their knowledge and understanding without compromising the accessibility of the questions.

In particular:

- Children in all three SEN groups had difficulty with the page layout and design of questions. Largely this seemed to be related to the amount of information they had to process and the unfamiliarity of some of the question types.
- Instructions on the test and in the questions need to be clear and concise so that children understand what is expected of them.
- Language should be simplified as far as possible without changing the construct being assessed.
- The examples, which were meant to scaffold children into questions, seemed to be ignored by or cause confusion in the children who took part in this study.
- Further work will be done with these groups to ensure that modifications or changes to the questions will not disadvantage the children
- Slow production of children’s responses, in part due to the density and complexity of information they are trying to process, needs to be considered when thinking about section timings.
- It is important to bear in mind that children with SEN will not always be presented with their preferred response types when taking statutory tests. For example, different print sizes and font may suit children with different special educational needs, and may be different to what the children are used to in the classroom.

As mentioned previously, the handwriting component has been excluded from the levels 3-5 test. However, Annex 5 provides information on the issues that were found with the different modes of administration and format, as well as full detail on these outcomes.

8.3.3 Summary

The most problematic questions for children with SEN were those with unfamiliar language, complex or unclear instructions, a high word count and high working memory requirement. Questions with an unfamiliar layout and questions being too close together on the page were also problematic. However, the number of questions highlighted as being challenging in this way was generally low. This was in part due to the work already done to make the questions clear, concise and with simple language.

Many of the issues raised are also contributors to CIV. The work done on reducing CIV and producing questions with simple language, clear instructions and in an accessible layout will help to make questions more accessible. More details on research looking at CIV in the test questions can be found in Annex 4.

The findings from this research were used during question selection and test construction for the 2013 test. Based on issues found during the study some questions have been changed and were available for selection. Other questions were not considered for the 2013 test, and will be reviewed and potentially modified before they are included in future years.

8.4 Double marking study

This section reports on the double marking study that was undertaken on the test questions. The purpose of the double marking study was to ensure that the test questions and mark schemes are clear and unambiguous so that marking will not contribute to CIV. It should be noted that the test was designed with the expectation that it would be marked on-screen.

A brief overview of the procedure used and a summary of the outcomes precede a section on how this information was used in the test construction process and will be used in future test development cycles.

8.4.1 Procedure

In order to ensure that the marking of the test questions is reliable, a double marking study was conducted using processes which mirror those that will be employed in the live marking of the tests as far as possible. Some examples of why it was not possible to follow the live marking procedure include:

- this study took place in October 2012, after the initial marking/coding exercise of the trial.
- the markers were aware that the purpose of the exercise was to examine the reliability of the marking of the questions and that no assessment of their marking ability was being made as a result of the outcomes of this study.

As noted previously, there were 12 levels 3-5 versions and three level 6 versions in the trial. The questions included in the double marking study were from six of the levels 3-5 trial versions and all of the level 6 versions. This ensured that all of the trial questions were included in the double marking study.

As with live marking, lead senior markers were assigned to a test version. They created marker training materials and all markers involved in the double marking study participated in a marker training event. Markers were assigned to a single test version. After the marker training event, all markers were standardised to ensure they could mark the questions to the mark scheme, before being permitted to mark children’s responses.

There were two marking windows. The on-screen marking system was set up such that no marker marked the same child's response across marking windows. A separate data file for each marking window was provided by the on-screen marking supplier.

8.4.2 Outcomes

Initial analysis was conducted to ensure that the double marking data did not deviate strongly from the original trial. If it had, it would have been an indication that there was a problem with the set-up inputs; however, there was no strong deviation from the scoring from the original trial.

Next, the short answer questions and spelling words selected for inclusion in the 2013 live test were examined to determine the level of marker agreement across the two marking windows. A minimum number of observations took place for each marking window:

- levels 3-5 questions - 486 observations.
- level 6 questions - 1017 observations.

Cohen's Kappa was used as an index of marker agreement. Cohen's Kappa ranges from 0 (indicating agreement only by chance) to 1.00, with larger values indicating better agreement. Per cent marker agreement was also calculated. Questions were flagged if Cohen's Kappa was less than 0.80[17]. This flagged two of 66 questions in the live levels 3-5 test and one of 33 questions in the live level 6 test. It is worth noting that the levels of marker agreement in level 6 were generally lower (as measured by Cohen's Kappa and per cent agreement). These three questions were examined again by STA's Test development researchers and it was determined that the mark schemes needed to be amended to ensure they are clear. Additionally, one question has also undergone formatting changes which will help with the marking. Another of the questions had the name in the target sentence changed to ease marking (the name began with 'm' so it was difficult to determine if had been capitalised as required by the question instructions). All three questions have had intensive work on the marker training material which should improve the marking reliability of these questions.

A comparison was also made in the marker agreement for the extended task in level 6. As expected when marking a piece of writing, the absolute marker agreement across the three strands was in the region of 50 per cent. However, when markers could be within one mark of each other (referred to as adjacency), the marker agreement was over 95

[17] Although there are no agreed magnitude guidelines for values of Kappa, given that factors other than agreement can influence magnitude, Fleiss (1981) Statistical methods for rates and proportions (2nd ed.). New York: John Wiley, states that values over 0.75 are excellent, values between 0.40 and 0.75 are fair to good, and below values below 0.40 are poor. As a result, 0.8 was selected as a suitable value for consideration.

per cent for TSO and AV, and SSP was 85 per cent, which is satisfactory for this type of assessment.

8.4.3 Summary

The double marking study provided an opportunity to determine the expected level of marker agreement for the test questions. Only three questions across levels 3-5 and level 6 were flagged with Cohen's Kappa values less than 0.80. These questions and their mark schemes were examined further and resulted in changes to the questions and mark schemes that will be used in the 2013 live tests. All other questions not included in the 2013 levels 3-5 or level 6 live tests were also analysed. The outcomes of the double marking study will be taken into consideration when questions are reviewed for inclusion in a future live test, and in future question writing.

8.5 Test re-test / Alternate forms study

To contribute to evidence of reliability, this section reports on an investigation into the relationship of test outcomes when two versions of the trial questions are administered to a child within the same test administration period.

All children who took the level 6 trial questions took two versions. This study examines the inferences that can be made about the test-retest/alternate forms reliability of the test in these circumstances.

This evidence is somewhat artificial because the purpose of the trial was to be able to construct a live test from questions across the three trial versions. Nevertheless, because the versions have been developed to the same specification, some inferences about reliability can be made.

8.5.1 Methodology

There were three versions of the level 6 trial questions. Each version contained an extended task, short answer questions and spelling questions. Each question only appeared in one version. The versions were designed to the same specification and expected to perform similarly. However, it should be noted that there was no quantitative data to aid in the booklet construction of the trial questions; this could explain some differences in outcomes of the three versions.

Three groups of children took different combinations of level 6 trial questions:

- Group 1 took versions 1 and 2.
- Group 2 took versions 2 and 3.
- Group 3 took versions 1 and 3.

The administration was counter-balanced. This means that within each group some children took one version, then the other, while the rest of the children took the second version first. Any differences in the results of the two versions cannot therefore be attributed to order effects such as fatigue, because they have both been administered first and second within the group.

8.5.2 Outcomes

As can be seen in Figure 18 the means, standard deviations and standard errors of measurement are quite similar across the three groups. This provides evidence that the group correlation coefficients can be interpreted as reliability estimates. The correlations are reasonably high which suggests score equivalence between versions. Equipercentile equating of the versions provides further evidence of the linear association between forms, with equated scores being within the standard error of measurement.

	Group 1		**Group 2**		**Group 3**	
	Version 1	Version 2	Version 2	Version 3	Version 1	Version 3
n	532		536		467	
mean	37.07	39.30	39.43	34.44	35.77	33.22
SD	8.49	9.15	8.99	8.32	8.72	8.24
SEM	3.50	3.66	3.71	3.43	3.49	3.40
Correlation	0.77		0.79		0.77	

Figure 18: Total score means, standard deviations, standard errors of measurement and correlations

8.5.3 Summary

Children who participated in the level 6 trial took two versions of the level 6 test. These versions were constructed to the same specification, though no quantitative data existed at the time of booklet construction to verify their equivalence. Basic descriptive statistics across the groups of children participating were very similar and the correlation of each set of scores is reasonably high so that we can have confidence in the reliability of the alternate forms.

8.6 Overall summary

Taken together these five studies provide evidence for the validity of the outcomes of the test. Only a small number of questions with potential CIV were uncovered, suggesting that the vast majority of the questions in the trial were effective in assessing the intended construct with little or no ‘noise’. KTL/KPR explained a large proportion of the variation in question difficulty. Other modifiers of difficulty, response strategy and sentence complexity, also explain some variation in question difficulty. Question writers can influence these three factors in order to affect question difficulty, but care must be taken so that manipulation does not become construct irrelevant. Many of the issues raised in the SEN study are also contributors to CIV. The work done on reducing it and producing questions with simple language, clear instructions and in an accessible layout will help to make questions more accessible. The double marking study provided an opportunity to determine the expected level of marker agreement for the test questions. It also showed where more clarity in the mark schemes was required. The knowledge that has been gained from the outcomes of these four studies also provides an opportunity to positively influence future question writing. Finally, the alternate forms reliability estimate for the level 6 versions suggests that the versions were reliable, even though they were constructed without the benefit of quantitative data.

9 Conclusion

This section of the report will focus on Ofqual's common assessment criteria (Ofqual, 2012) and will attempt to demonstrate the quality of the test using the evidence from the test development process.

9.1 Validity

Ofqual's regulatory framework for national assessments (2011) states that an assessment should 'generate outcomes that provide a valid measure of the knowledge, skills and understanding that the learner is required to demonstrate as specified by the assessment objectives'. It states that 'Validity is the central concept in the evaluation of the quality of assessments' such that 'processes and procedures [are] expected to [ensure and generate] evidence to support the way in which the assessment outcomes are interpreted and used'. The document also states that:

> The validity of an assessment refers to the extent to which evidence and theory support the interpretation that the assessment outcomes meet their intended uses.
>
> The evaluation of validity involves the development of a clear argument to support the proposed interpretation of the outcomes and as a consequence the intended uses of the assessment. The validity argument should be built on statements of the proposed interpretation and supporting evidence collected from all stages of the assessment process.

Therefore, the development of a validity argument must start with an understanding of the purpose of the assessment. The statutory purpose of National Curriculum tests is to assess 'the level of attainment which [pupils] have achieved in any core subject'. In addition, the Bew review set out three additional principal uses for National Curriculum tests:

- holding schools accountable for the attainment and progress made by their pupils and groups of pupils;
- informing parents and secondary schools about the performance of individual pupils; and
- enabling benchmarking between schools; as well as monitoring performance locally and nationally.

Since these three uses relate to how the data is used following live administration, it is not possible to provide a full validity argument for them at this time. The evidence in this report, however, does provide evidence relating to the statutory purpose.

To determine whether the test is a sufficiently valid assessment of the level of attainment which children have achieved in English grammar, punctuation and spelling there are a number of questions that need to be answered:

- Is the test framework an appropriate assessment of the relevant sections of the National Curriculum programme of study in English?
- Is the test an appropriate assessment of English grammar, punctuation and spelling?
- Are the reported outcomes of the test appropriate with respect to National Curriculum levels?

In relation to the first question, the test framework was developed to closely align to the relevant elements of the National Curriculum programme of study for English and the reference codes assigned to the assessable elements of the test are explicitly linked to the relevant section of the programme of study. This ensures that all of the questions in the test can be directly linked to aspects of the National Curriculum. The development of the test framework has involved a number of experts in the field and has been supported by evidence from trialling. Therefore, STA believes that the test is reflective of the relevant sections of the National Curriculum programme of study for English and that the framework is appropriate.

In relation to the second question, the test development process has collected a great deal of evidence relating to the content of the test and whether the questions appropriately assess the relevant skills, in particular the work on construct irrelevant variance that showed very few questions assessing something other than the construct. The experts involved in the development of the test have a wealth of expertise and experience. Trialling has provided sufficient data on the questions to enable STA to construct a test to meet the specification in the test framework.

Although the independent experts who reviewed the materials raised some concerns about the nature of the test, they appreciated that this specification was a product of Lord Bew's recommendations. On balance, the evidence from the independent experts gives STA sufficient confidence that the test is assessing English grammar, punctuation and spelling appropriately. STA therefore believes that the test is an appropriate assessment of English grammar, punctuation and spelling, within the parameters defined by Lord Bew's recommendations.

The answer to the final question cannot be provided until standards have been set on the live 2013 test. However, STA is confident that the process that it will follow, which is widely used internationally, will ensure that reported outcomes are appropriate.

The development of a validity argument is an on-going process. STA will continue to collect evidence to demonstrate that the test is sufficiently valid for the purpose for which it is intended.

9.2 Reliability

Ofqual's regulatory framework for national assessments (2011) states that an assessment should 'generate outcomes that provide a reliable measure of a learner's performance'. The document also states that:

> Reliability is about consistency and so concerns the extent to which the various stages in the assessment process generate outcomes which would be replicated were the assessment repeated. Reliability is a necessary condition of validity, as it is not possible to demonstrate the validity of an assessment process which is not reliable. The reliability of an assessment is affected by a range of factors such as the sampling of assessment tasks and inconsistency in marking by human markers.

To demonstrate sufficient reliability for the test, the following aspects must be considered:

- The internal consistency;
- The classification consistency;
- The classification accuracy; and
- The consistency of scoring.

The analysis of the evidence from the technical pre-test has demonstrated generally high levels of internal consistency for the test and reasonable standard errors of measurement for each component.

Classification consistency refers to the extent to which children are classified in the same way in repeated applications of a procedure. Although limited evidence is available at this stage, evidence from the test re-test/alternate forms study shows that the basic descriptive statistics across the groups of children participating were very similar and the correlation of each set of scores is high enough to have confidence in the reliability of the alternative forms.

Classification accuracy refers to how precisely children have been classified. Reasonable estimates of classification accuracy will only be valid once the test has been administered in all schools. Therefore, further work on reliability will be analysed and reported in autumn 2013.

Consistency of scoring relates to the extent to which children are classified the same way when scored by different markers. Evidence from the double marking study indicates a high level of marker agreement for the test questions.

At present, STA is satisfied that the test is a sufficiently reliable assessment.

9.3 Comparability

Ofqual’s regulatory framework for national assessments (2011) states that an assessment should ’generate outcomes that are comparable in standards over time’. The document also states that:

> Comparability is about generating assessment outcomes that are comparable in standards over time and between assessment cycles. Where a test has equivalent forms – as is the case with National Curriculum assessments, where, for example,

> the Key Stage 2 mathematics test in each year comprises different questions, but is still treated as the same test over time – then it is important to ensure comparability of outcomes.

When introducing a new test there are often no existing assessments with which to be comparable. However, the test development process has also produced an anchor test that will be used to link standards in future pre-tests to those that will be set on the live test this summer, therefore ensuring comparability.

9.4 Minimising bias

Ofqual's regulatory framework for national assessments (2011) states that an assessment should 'minimise bias, differentiating only on the basis of each learner's ability to meet National Curriculum requirements and early learning goals'. The document also states that:

> Minimising bias is about ensuring that an assessment does not produce unreasonably adverse outcomes for particular groups of learners. The minimisation of bias is related to fairness to all children and is also closely related to statutory equality duties.

The evidence from the SEN studies shows that the most problematic questions for children with SEN were those with unfamiliar language, complex or unclear instructions, a high word count and high working memory requirement. Questions with an unfamiliar layout and questions being too close together on the page were also problematic. However, the number of questions highlighted as being problematic was generally low, and were either able to amended or were excluded as far as possible. This is in part due to the work already done to make the questions clear, concise and with simple language.

9.5 Manageability

Ofqual's regulatory framework for national assessments (2011) states that an assessment should be 'manageable so that the scale of the assessment process is balanced by the usefulness of the outcome'. The document also states that:

> Manageability relates to the feasibility of carrying out particular assessment processes. A manageable assessment process is one which places reasonable demands on schools and children. The evaluation of the reasonableness of the demands will be based on the scale of the assessment process on the participants, balanced by the usefulness of the outcomes. As with the other common criteria (validity, reliability, comparability and minimising bias), judgements about manageability must be balanced with considerations around the other common criteria.
>
> The responsible body or bodies are expected to demonstrate that there are appropriate documented procedures in place to meet the criteria.

The test replaces the English writing test in the National Curriculum test timetable and has similar administration requirements in terms of time length and administration (a mixture of written test and aural test).This means that the test is not placing an additional burden on schools and should therefore be manageable. As stated previously, evidence about the usefulness of the outcomes cannot be provided until results are available.

9.6 Overall statement in relation to common criteria

Having examined all of the evidence gathered so far through the test development process, STA is satisfied that the test is a sufficiently valid assessment of the domain, has acceptable levels of reliability and is fair for children and manageable for schools. However, as stated previously, the development of a validity argument is an on-going process and additional analysis will be carried out following the first live administration of the test to ensure that STA can continue to be confident in this assertion.

9.7 Future work

A number of activities are planned as part of the continuing development of the tests. These include:

- The standard setting exercise;
- The development of a more complete validity argument;
- Additional work to determine the reliability of the test with the data from the live administration, in particular the estimation of classification accuracy; and
- Analysis of live test data, in particular through an additional anchor test study to investigate whether item parameters for questions trialled in the technical pre-test are stable over the introduction of the test.

Annex 1: External experts

Test development expert group:

Dr Frances Brill

Research Manager - National Foundation for Educational Research (NFER)

Frances Brill is a literacy assessment specialist. She has contributed to a wide range of literacy and assessment development projects at the NFER. Frances has a background in linguistics and education, holding a PGCE, an MA in general linguistics and a PhD in educational linguistics. Her teaching experience includes key stage 2 and lecturing in syntax to undergraduates. Frances is the co-editor of NFER's peer-review international journal, *Educational Research*. She is a Fellow of The Chartered Institute of Educational Assessors and a Practitioner of the Association for Educational Assessment – Europe.

Paul Wright

External consultant – Curriculum

Paul Wright is an independent Curriculum Expert contracted to the STA to scrutinise test development processes for Key Stage 2 tests. Paul worked as an English Consultant, Programme Manager and Senior Curriculum Adviser for QCA before becoming an independent consultant. Paul was part of the team responsible for writing the Key Stage 3 National Curriculum programmes of study. For many years, Paul was a secondary teacher of English, drama and media studies.

Hester Glass

Independent consultant - English, EAL, Test Development

Hester Glass is a former independent educational researcher, specialising in curriculum and test development. She has undertaken work at a national and international level, specialising in English and English as an additional language. Hester has significant experience as a Key Stage 3 English Test Developer at Cambridge Assessment and recent teaching experience in school, including as Head of English. Hester was also a Lecturer at the Open University for three years.

Anne Basden

Experienced senior marker and primary head teacher

Anne Basden is an experienced marker and has been on the senior marking team since the test model changed in 2003. She has marked the Key Stage 2 English tests since their conception. Anne has also contributed to test development work in terms of mark scheme refinement and took a lead role in the marking of the 'single level tests'. Anne is a head teacher in Rutland.

Independent academic review:

Prof David Crystal

Independent consultant and author – grammar and linguistics

David Crystal works as a writer, editor, lecturer, and broadcaster. David is an experienced lecturer in linguistics, first at Bangor, then at Reading. He published the first of his 100 books in 1964, and became known chiefly for his research work in English language studies, in such fields as intonation and stylistics, and in the application of linguistics to religious, educational and clinical contexts, notably in the development of a range of linguistic profiling techniques for diagnostic and therapeutic purposes. He has written prolifically ever since and is now Honorary Professor of Linguistics at the University of Wales, Bangor.

Prof Debra Myhill

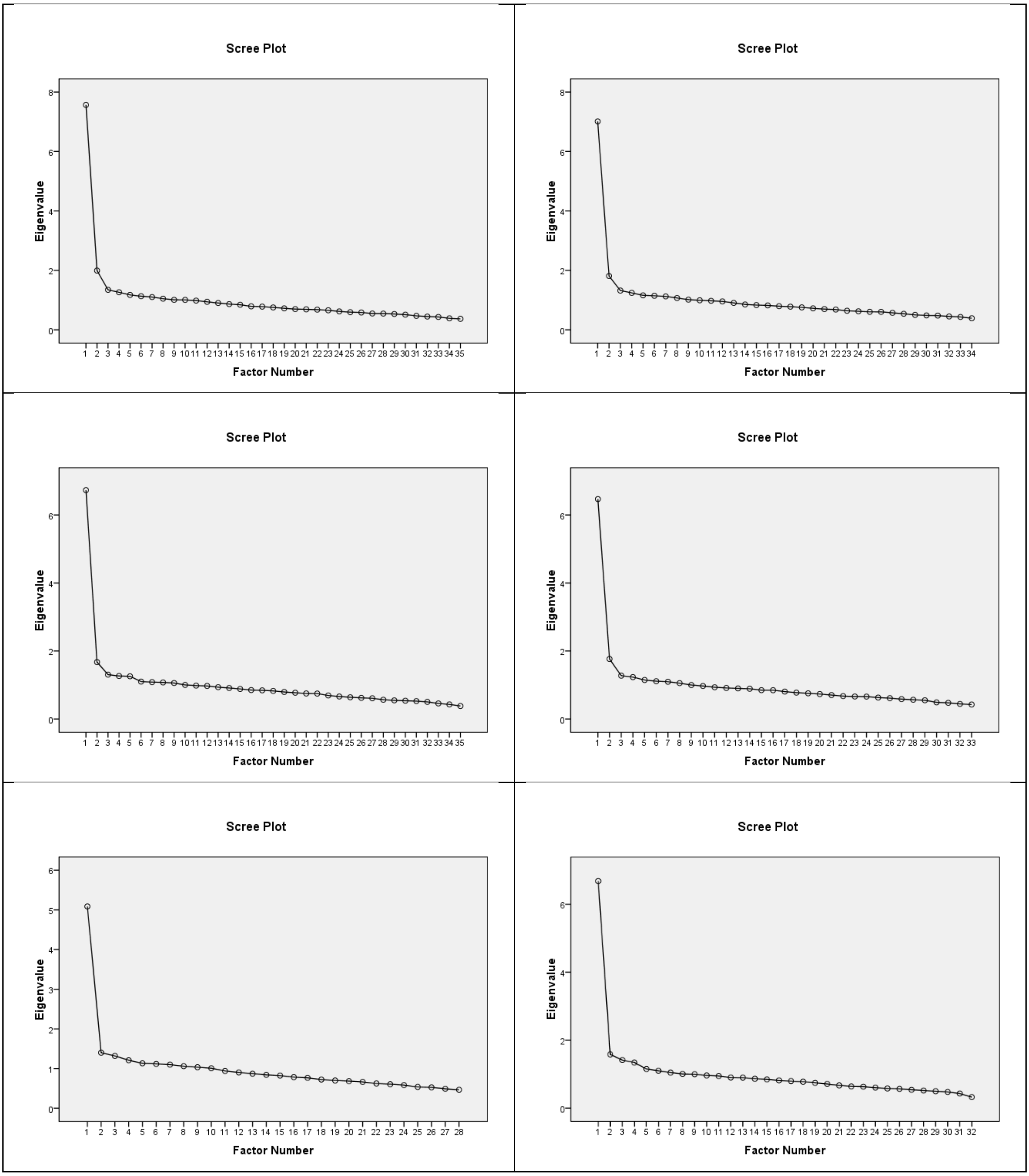
Independent consultant and author – English grammar and writing in education

Debra Myhill is Associate dean for Research and knowledge transfer of the College of Social Sciences and International Studies, and subject leader for English with media at the University of Exeter. She is an experienced author, researcher and consultant, who is working with the Department on the new programme of study for English at primary level.

Annex 2: Assumption checking

	Average Q3	Standard deviation	Maximum absolute value of Q3
Level 3-5 SAQ	-0.03	0.05	0.36
Level 3- 5 spelling	-0.04	0.05	0.20
Level 6 SAQ	-0.02	0.06	0.81
Level 6 spelling	-0.02	0.05	0.21

Figure 19: Q3 values



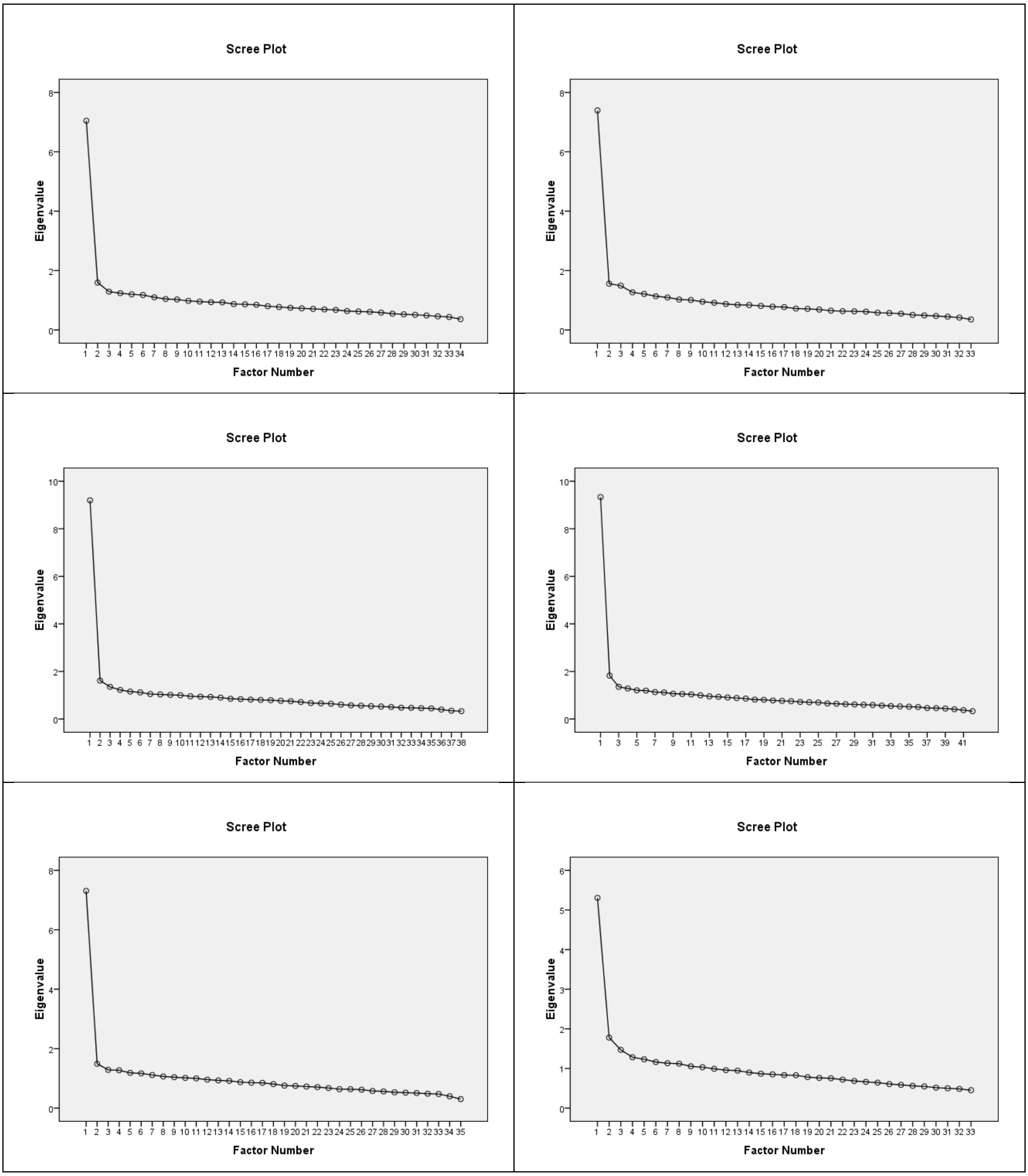
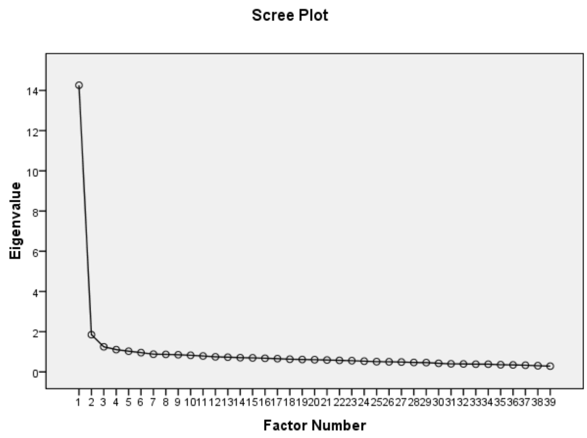
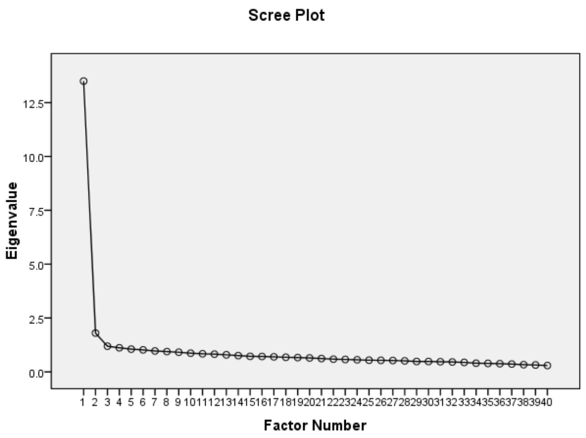
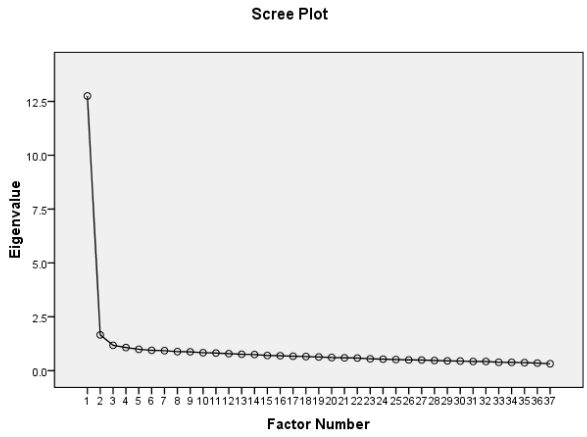
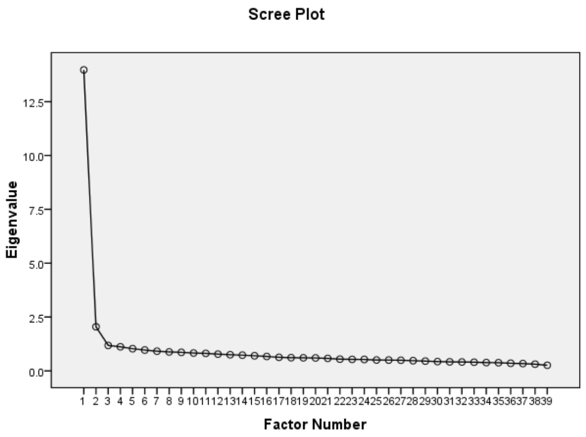


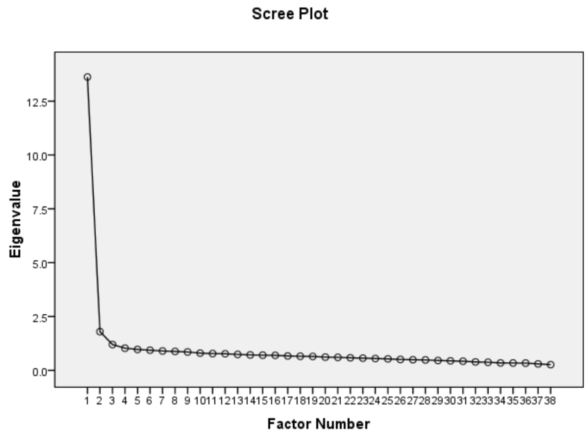
Figure 20: Level 3-5 SAQ scree plots

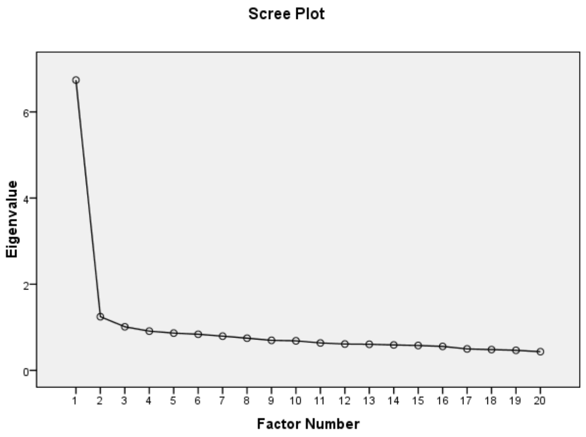

Scree Plot
Eigenvalue
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Factor Number


Scree Plot
Eigenvalue
12.5
10.0
7.5
5.0
2.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Factor Number


Scree Plot
Eigenvalue
12.5
10.0
7.5
5.0
2.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
Factor Number


Scree Plot
Eigenvalue
12.5
10.0
7.5
5.0
2.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Factor Number


Scree Plot
Eigenvalue
12.5
10.0
7.5
5.0
2.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Factor Number


Scree Plot
Eigenvalue
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Factor Number

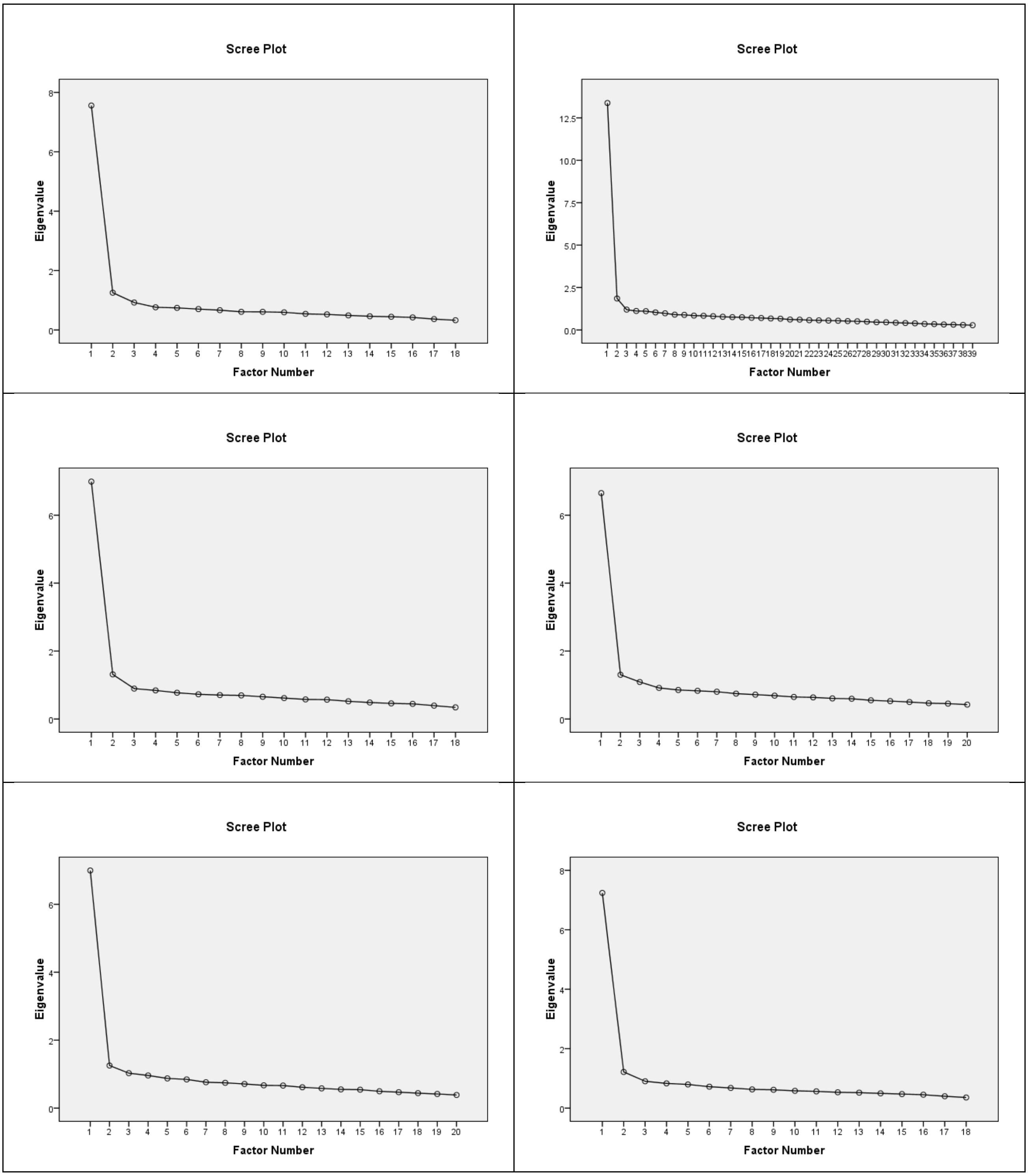


Figure 21: Level 3-5 spelling scree plots

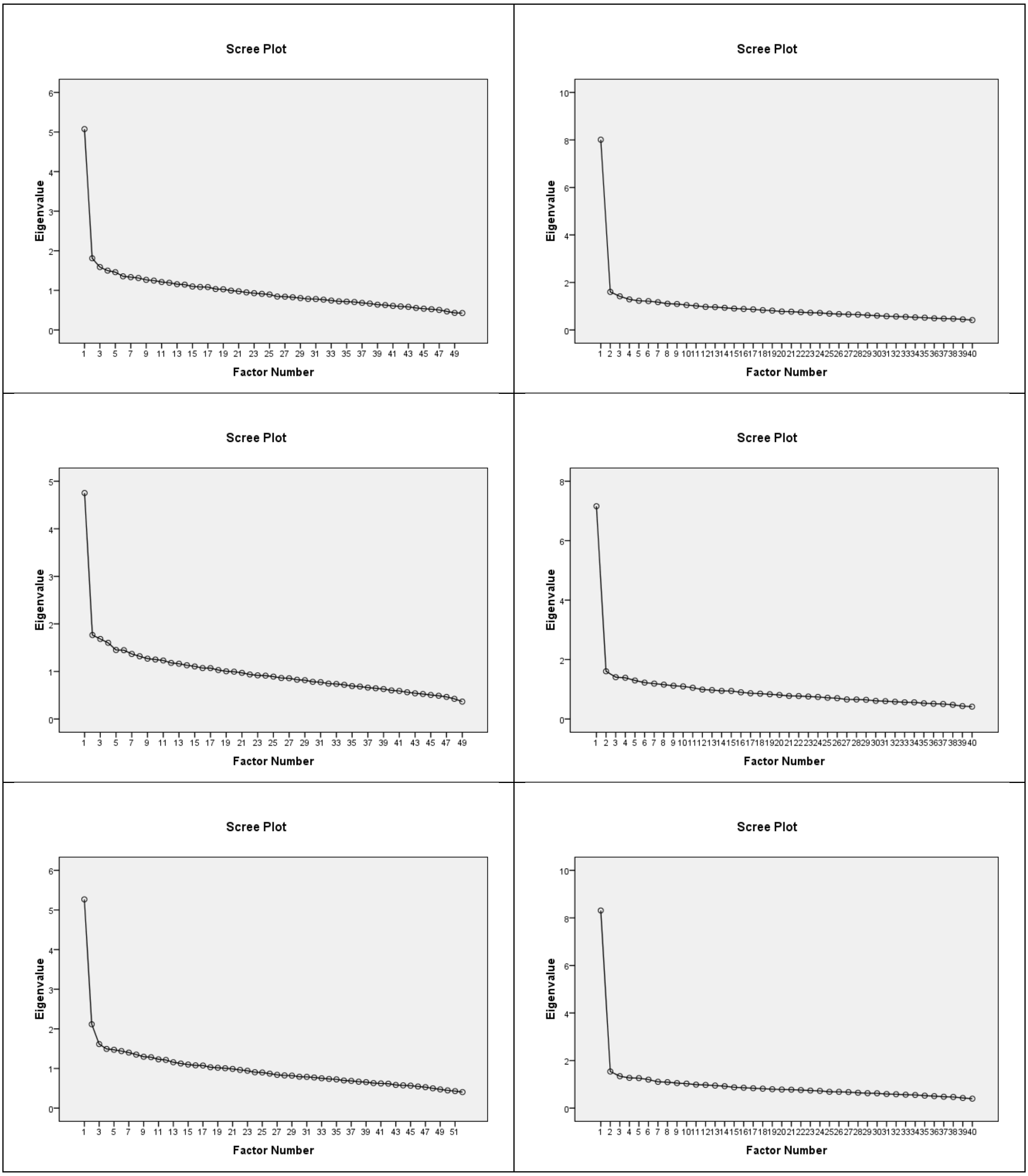


Figure 22: Level 6 scree plots

Annex 3: Standard setting procedures

For the Bookmark Procedure, question level data from the live test administration in May 2013 will be analysed using a two-parameter graded-response IRT model. Question mark points will then be placed in order of the level of ability required to have a probability of two-thirds of getting the question correct, in order to create an ordered item booklet. Participants will use the performance descriptor to imagine a child just at the level threshold (minimally competent) and go through the ordered item booklet, deciding for each question whether the minimally competent child, at each threshold, would be able to achieve the mark two-thirds of the time. This will happen over several rounds, with participants working individually, in small groups and as a whole group, to bring participants to consensus on the recommended threshold for each level.

In the Body of Work Procedure a sample of extended tasks will be ordered by total score. The participants do not know the exact score any test script received, but do know they are ordered by total score. The first round is referred to as ‘rangefinding’. Participants will examine the children’s scripts using the performance descriptors to categorise each script into a performance category (level 6 or below level 6). In this round participants may discuss their ratings with one another. Before the second round, called pinpointing, facilitators and analysts will discuss which scripts should be eliminated based on the participants’ ratings. Variability in judgements about scripts is a possible indication of the location of a cut score. Therefore in the pinpointing round, additional scripts will be included with those from the rangefinding for the remaining mark range. Again, the scripts will be reviewed and classified by the participants. A logistic regression exercise will be used to analyse the likelihood of specific scripts being classified to specific categories (for example below level 6 or level 6) and calculate the cut score.

For the level 6 test the cut scores from the Body of Work and Bookmark Procedures will be summed to derive the final cut score.

Annex 4: Study into potential construct irrelevant variance

Messick described two types of CIV: construct irrelevant difficulty and construct irrelevant easiness. The former occurs when ‘aspects of the task that are extraneous to the focal construct make the test irrelevantly more difficult for some individuals or groups’ and the latter ‘when extraneous clues in question or test formats permit some individuals to respond correctly in ways irrelevant to the construct being assessed’. The former leads to lower scores for some examinees and the latter to higher scores.

The purpose of the study is to uncover potential sources of CIV. Judgements regarding CIV inevitably include some level of subjectivity, and it is for subject specialists to determine whether the issues raised lead to questions that are flawed and inappropriate for use, or whether, for example, the extent of a CIV issue is minimal and not of significant concern.

Methodology

Three sources of evidence were used in the process of identifying potential construct irrelevant elements in the trialled test questions. These are listed below.

School visits

Four STA researchers visited ten schools to observe the administration of the test trials, and conduct interviews afterwards. The schools were selected so that five were trialling the levels 3-5 tests and five were trialling the level 6 tests. The schools were also split geographically between London (five schools) and the West Midlands (five schools). Administrators and schools were made aware of the visits, and researchers explained they were there to observe and not to help with the administration of the test.

The school visits incorporated three strands:

Observation of the administration - This was carried out to identify any construct irrelevant elements in the administration methods. The researcher observed the administration of the test and made detailed notes of any questions which were raised by children and administration issues. Where possible the researcher also spoke to the administrator to collect their feedback on the test.

Interviews with children - The researcher interviewed children following the test. Two interviews were conducted with two children at a time and the researcher used a semi structured interview schedule. In total 40 children were interviewed. Interviews were recorded (where permission was granted) and detailed notes were made by the researcher. The children’s test scripts were available to refer to. Interview questions focused on individual questions in the test and aimed to highlight any questions that were easier or more difficult than intended for construct irrelevant reasons. Interviews lasted between ten and 45 minutes.

Interviews with teachers - In some schools researchers also interviewed teachers to collect their thoughts on the tests. Questions covered the test instructions, difficulty of questions, whether topics had been covered in class and the administration of the test.

Children's responses to technical pre-test questions

Responses to each of the short answer questions in the levels 3-5 technical pre-test were analysed. While analysis of children's responses can be a valuable tool in understanding the construct relevant and irrelevant elements within test questions, the method is of somewhat limited value in selected response questions. This is because children's responses to these questions are limited to ticks, lines or other non-textual indications, which usually tell us little about the way in which children interpret test questions. In some of the less structured question types, more responses were analysed as more could be gauged about children's thinking in these questions.

Item Validation Trial (IVT) Report

The agency, NFER, responsible for the administration of the Item Validation Trial (which took place in January 2012) produced a detailed report[18] on the performance of questions in the pre-trial. While sample sizes were small, the authors were able to make use of qualitative evidence from interviews (with children) and questionnaires (completed by teachers and administrators).

Outcomes

Short answer questions

Only a small number of questions with potential CIV were uncovered in this study, suggesting that the vast majority of the 368 questions in the trial were effective in assessing the intended construct with little or no 'noise'. This section contains a discussion of possible sources of construct irrelevant easiness and difficulty found in the short answer questions section of the levels 3-5 and level 6 tests.

Construct irrelevant easiness

In some cases, examples were provided within the question as a means of ensuring that pupils understood the task requirements. While in many cases these examples were found to be effective, there was a small number of questions for which the examples had unintended consequences. In these instances, they provided pupils with an algorithm for responding, which, if followed, would lead to a correct answer without the need for any

[18] This report cannot be released at this time as it contains confidential information about questions that will be used in future tests.

subject knowledge. A significant amount of evidence for this was found in the pupil interviews.

Construct irrelevant difficulty

Construct irrelevant difficulty can occur when language in the question stem, question or options is unfamiliar to a child. Where the unfamiliar language is not part of the construct being assessed, the language constitutes a source of construct irrelevant difficulty.

Children interviewed as part of the research highlighted a number of unfamiliar non-technical words and some children asked for specific words to be read to them. It was not possible to see whether they answered these questions incorrectly because of unfamiliar language, but it was evident that some children took longer on questions because of this.

Children identified some questions where they felt there was more than one correct answer. For example, children pointed out cases where exclamation marks and full stops could be used interchangeably in questions which required them to add punctuation at the end of a sentence. While the questions clearly ask for the ‘most likely’ punctuation, the issue still appeared to cause confusion for some children.

Some questions, particularly open response questions, were intentionally designed to have more than one correct answer. In the interviews children reported that they spent a disproportionate amount of time on these questions. This was not because they found them difficult, but because they were trying to think of the ‘best’ word they could provide as an answer.

For a child to be able to demonstrate their knowledge it is important that they understand what is being asked within a question. Children were confused by what was required for some questions. This included not knowing where they should record their answer, unfamiliar question types and missing out parts of a multi-part question.

The layout of the question or the formatting of the page can affect how a child answers the question. For example, the number of lines available for an open response question could influence a child’s opinion on how long/detailed the response should be. Children noted a number of questions where the layout or formatting was not user-friendly and examples of where the layout affected how they answered the questions.

Unfamiliar contexts in target sentences or passages can increase difficulty, particularly when children are required to comprehend the sentence/passage before they can answer the question. If knowledge of the context is not being tested the context should be accessible to all children to reduce construct irrelevance.

Spelling

No construct irrelevant factors were uncovered within the *content* of the spelling test. However, there were a number of potentially construct-related differences in the task that were related to the mode of administration (CD and administrator-read) being tested in the trialling.

Mode of administration

While on the CD version there are predetermined ‘gaps’ between each word/sentence, in the administrator-read (AR) version this is not the case. While the administrator will have been provided with guidance, there was evidence that administrators were not consistent in the time they left between words/sentences. For example, some appeared to wait until all children had finished writing before reading the next word/sentence. In one school the children thought there was plenty of time to write down spellings in the AR version, but just enough time in the CD version. In another school one child preferred the CD version because he found the gaps to be longer between spellings.

Children can also take visual cues from an administrator, so they know when they are about to speak and move on to the next spelling. This is not possible with a CD. Some children thought the sentences on the CD started quite abruptly and they were not always ready. These differences will inevitably have had some effect on performance.

The clarity of the voice is also clearly a source of difference between the two modes of administration. For the CD version the quality of the amplifier and speakers (and placement of speakers) can have an effect on the clarity of the voice and the volume. Some children reported that the words on the CD version were not always clear, and sounded different in the sentence to when read on their own. However clarity can also be an issue for administrators. An administrator’s accent may have presented an issue in some cases, although the same could be said for the CD version. Also of concern is the potential for administrators to unintentionally provide clues to the spelling because of the way they read the word.

Errors are potential issues for both formats. For CDs there is the possibility of the CD skipping or breaking, or the CD system in the school not working. For an administrator human errors such as missing a spelling or mispronouncing a word are potential errors. In the trials researchers witnessed an administrator missing a spelling, a CD skipping and a CD system breaking.

Passage vs. single sentence

Two different types of spelling tests were included in the trial. One version featured a passage with words missing, while the other was made up of individual sentences, each with a word missing. Some children found the passage easier to follow. However, other children said that it was easy to lose your place in the passage, particularly if you missed a spelling. There were also instances where a sentence contained more than one spelling, which some children reported as being a source of difficulty. Some children also thought that separate sentences gave the impression that each spelling was a ‘different question’, so if you could not spell one word you could ‘start again’ on the next question.

Context

The children who were interviewed said that the context of the passage for spelling tests affected their level of engagement. Where they were interested in the subject they were more engaged, and found the spellings easier to put in context.

Handwriting

This section highlights potential issues with the different modes of administration and format, but in most cases there is no suggestion that these issues are sources of CIV. Any decisions regarding construct relevance would require consideration of whether the issues discussed below compromise the intended focus of the task.

Copying vs. dictation

Most children said they were used to handwriting tests where they were required to copy a passage, although some children wanted clarification on whether they should copy the title or not. There were no other reported difficulties with this format. Most children had finished the copied handwriting test within ten minutes and thought there was enough time to complete the test.

Children raised a number of issues in the dictated handwriting test. They told interviewers that they found this part of the test difficult, and were not sure it was really a handwriting test as they also had to remember what had been read, spell it correctly and punctuate it correctly. Most children said they concentrated on getting the information down and spelling the words correctly rather than their handwriting. Children said they could have done better in their handwriting if they had not had to focus on other demands such as spelling. One child said he thought about his handwriting to begin with but this became more difficult as the test went on and he was struggling to keep up. Two teachers thought the test gave children too much to think about and did not account for different abilities in terms of processing speed and writing speed.

Mode of administration

Some children thought that the instructions on the CD version were not always clear, particularly the instructions for when they should start writing. Children thought a short introduction stating when they should start writing would have been useful. The children also thought that the instructions should explain that sentences will be read slowly, with pauses so that children could write down the dictated text.

Children reported that the pauses between segments of text on the CD were too long, but the speed the sentences were read was excessive. Children who were interviewed said they struggled to remember everything that was read in the sentence. As the passage progressed some children struggled to keep up. Some children also thought the speed the text was read was inconsistent.

Finally, one child said he struggled to hear some of the words on the CD.

Although the researchers observed slight differences in the speed of reading between administrators there were no comments from children relating to the administrators' reading of the text.

In the administrator read version, one administrator thought the guidance could be clearer. One administrator said it could be unclear whether an administrator should read the punctuation.

Summary

This validity study was undertaken to examine the extent to which the questions, tests and administration methods in the trial test may be subject to CIV. In the short answer questions, only a small number of questions with potential CIV were uncovered in this study, suggesting that the vast majority of the 368 questions in the trial were effective in assessing the intended construct with little or no 'noise'.

Construct irrelevant easiness issues included examples within questions providing unintended clues to the correct answer, and information within questions providing clues to answer other questions in the same test that covered the same or related content. Construct irrelevant difficulty issues included the use of difficult non-technical vocabulary, the use of challenging contexts for sentences/passages, and lack of clarity of question requirements.

There were no sources of CIV in the content of the spelling and handwriting components, although there were a number of potential construct-related differences that were related to the mode of administration.

Key findings from this study were used to inform question selection for the 2013 live test, and will be used in future to inform question writers.

Annex 5: Modifiers of difficulty

This annex documents qualitative and quantitative analyses that have been undertaken to identify the main factors that moderate the difficulty of test questions in the short answer questions of the June 2012 technical pre-test. The annex includes some discussion on each of the factors, provides validation evidence and considers implications of the work.

There are three main purposes for this work:

1. Understanding the factors that moderate the difficulty of test questions provides important validity evidence. In particular, it may help to expose construct irrelevance issues. This is because, once the factors that appear to affect question difficulty have been identified, the construct relevance of those factors can be adequately considered.
2. To provide question writers and test developers with useful information regarding the construct relevant (and irrelevant) means by which question difficulty can be intentionally manipulated.
3. To aid question writers in making reasonable predictions about question difficulty when writing test questions in the future.

Methodology

Three sources were used in the process of identifying the factors appearing to affect the difficulty of questions. These are listed below.

Children's responses to technical pre-test questions

Responses to each of the short answer questions in the levels 3-5 technical pre-test were analysed, including at least 50 responses to each of the questions in levels 3-5 version 1. While analysis of children's responses can be a valuable tool in understanding the sources (both invalid and valid) of difficulty of test questions, the method is of somewhat limited value in selected-response questions. This is because children's responses to these questions tend to be limited to ticks, lines or other non-textual indications, which usually tell us little about the way in which they interpret test questions. In some of the less structured question types more responses were analysed, as more could be gauged about children's thinking in these questions.

Interviews with children

Approximately 40 children were interviewed immediately after taking the tests. A number of themes, relating to both difficulty and construct relevance, emerged from the analysis of these interviews. These are discussed in more detail in Annex 4. Evidence from these interviews was also used in this annex as a means of identifying factors affecting question difficulty.

Item Validation Trial (IVT) Report

The agency responsible for the administration of the Item Validation Trial (which took place in January 2012) produced a detailed report[19] on the performance of questions in the trial. While sample sizes were small, the authors were able to make use of qualitative evidence from interviews (with children) and questionnaires (completed by teachers and administrators).

Outcomes

Factors affecting Question difficulty

An analysis of evidence from the sources above yielded 18 factors that were found to affect the difficulty of test questions, as listed in Figure 23 below. The purpose was to find all factors affecting the difficulty of the test questions. At this stage, the construct relevance of the factors was not considered. Any implications of this annex for construct relevance are discussed in the Implications section. A more detailed discussion of construct irrelevance in the tests can be found in Annex 4.

	Factor	**Description**
1	Knowledge of technical language	The requirement to know the meaning of technical terminology (e.g. word classes).
2	Provision of examples	The provision of an example within a question can change the nature of the question. Question difficulty is often reduced, usually through a reduction in the knowledge of technical language required.
3	Question interaction effects	Evidence from interviews with children suggests that they were sometimes able to answer because the meaning of relevant technical terminology could be gauged from another question in the test.
4	Chance level	The probability of answering selected response questions in technical pre-test correctly through random guessing. As well as multiple choice and matching question types, this factor applies to questions where children are required to indicate (e.g. through circling) one or more words within a sentence or passage.
5	Number of features is specified	Stating the number of features to be indicated/selected (e.g. ‘Circle the two nouns in the sentence’) results in a reduced difficulty compared with questions that do not specify the number of features (e.g. ‘Circle all the nouns….’).

[19] This report cannot be released at this time as it contains confidential information about questions that will be used in future tests.

	Factor	Description
6	Established misconceptions	Questions assessing speech misconceptions that are highly established such as ‘John and me’ vs. ‘John and I’.
7	Sentence length	The length of target sentences/passages within a question. In particular, passages (i.e. where sentences are combined) require children to assimilate more information.
8	Sentence complexity	The complexity of target sentences/passages.
9	Sentence vocabulary	The difficulty of vocabulary used within target sentences/ passages.
10	Sentence abstractness	The level of abstractness of ideas contained within target sentences.
11	Response type familiarity	The degree of familiarity children have with the response type. Most children will be familiar with many of the response types, including multiple choice and matching and circling, from their experience of other National Curriculum assessments and classroom assessments.
12	Explicitness of question requirements	The degree of explicitness of question requirements, particularly in relation to response types that are less familiar to children.
13	Own words required	The extent to which children are required to generate their own language in providing a response to a question.
14	Best answer vs. only one correct answer	Questions where there is more than one possible ‘correct’ answer, but only one ‘best’ (and creditworthy) answer, may be more difficult for children than questions where more than one answer could be considered correct.
15	More than one partially correct combination of answers	Particularly in cloze-type passage-based questions, there may be some words which can be correctly placed in more than one space, but only one correct combination of answers. This increases difficulty as children have to assimilate the whole passage in order to answer correctly, thereby increasing reading demand.
16	Number of repetitions required for credit	More repetitions mean a higher chance of an error even when children possess the required subject knowledge.
17	Complexity of instructions	The length and complexity of the question instructions.
18	Knowledge of punctuation rules	Knowledge of rules relating to the punctuating of sentences.

Figure 23: Factors affecting the difficulty of test questions in the trial (shown in no particular order)

A closer inspection of the 18 factors listed above showed that they could be categorised under three higher-level factors:

1. Knowledge of technical language and punctuation rules;
2. Response strategy; and
3. Sentence complexity.

These are listed in Figure 24 below, alongside the relevant factor number from Figure 23. Each of the three higher level factors is considered in more detail below.

Factor	Relevant factors from Figure 23
Knowledge of technical language and punctuation rules	1, 2, 3, 6, 16, 18
Response strategy	4, 5, 11, 12, 13, 14, 15, 17
Sentence complexity	7, 8, 9, 10

Figure 24: Higher level factors found to affect question difficulty

Knowledge of technical language and punctuation rules

In the definition used in this report, subject knowledge is split into two sub-components:

Knowledge of technical language (KTL): This refers to the extent to which knowledge of the meaning of technical terms (e.g. word classes) is required to answer grammar and punctuation questions. Target words in vocabulary questions are also considered here to be technical terms.

Knowledge of relevant punctuation rules (KPR): This refers to the extent to which knowledge of punctuation rules is required to answer punctuation questions.

The report on the pre-trial recognised the importance of the first of these factors, when it stated that ‘some pupils had difficulty identifying the grammatical categories for given words, clauses, phrases or sentences. It is likely that this difficulty has its origins in unfamiliarity with the terminology used and also in lack of grammatical knowledge (e.g. not knowing the definition of a technical term such as *contraction*)’.

The grammar questions in the test can be broadly divided into those that assess children’s ability to identify grammatical elements, and those that assess use of grammar. A high proportion of the latter require no KTL, while by definition all questions assessing the former do require KTL.

It was clear from interviewing children at both levels 3-5 and level 6, as well as from the qualitative analysis of children's responses, that the amount of KTL required to answer questions was a key modifier of difficulty. During an interview, one child stated that the questions tended to be 'either very easy or very difficult', and it became clear when prompted that questions were found to be difficult primarily when children lacked the KTL required to answer.

The difference between performance on questions requiring KTL and those not requiring KTL, when focusing on the same word class, was often marked. Despite being focused on the same content, the difference in performance between these two types suggests that the skills/knowledge being assessed by the questions are substantially different. Grammar questions requiring no KTL tended to be easy, and could be answered purely by considering what 'sounds' right. These questions appear to be assessing competence in speaking fluency, a less demanding skill than competence in written language, and one that the vast majority of children should be competent in by the end of Key Stage 2. A rare exception to this rule occurs in a small number of grammar questions that assess misconceptions that are widespread, even among adults in spoken English.

Questions assessing recall of a particular piece of content would usually be expected to be easier than questions assessing application of that content. Analysis of the question level data for the trial showed that this tended not to be the case with the grammar questions; questions assessing recall (in the form of KTL) tended to have lower facilities than those assessing use of grammar.

Additionally, in some cases, the KTL required in questions was either reduced or removed through the provision of an example.

Response strategy

The response strategy demand appears to be affected by three variables:

- complexity of response requirements;
- familiarity of response requirements; and
- explicitness of response requirements.

These are discussed in turn below.

Complexity of response requirements

While none of the question types require children to write more than a sentence, there is still a range of response strategy demands within the short answer questions in the tests. In broad terms, this ranges from selected response questions (at the lower end of the scale) to open response questions where children are required to generate their own language to answer the question. However even within selected response questions, demand can be manipulated in a number of ways, including:

- Increasing the number of options (distractors), so that the chance of answering correctly by guessing is reduced.
- Using questions where more than one option could justifiably be considered to be 'correct', but where there is a single 'best' (and creditworthy) answer, as these may be more difficult for children than those where there is no such consideration to be made.
- Not stating the number of words to be indicated, e.g. 'Circle the two nouns in this sentence' as opposed to 'Circle all the nouns in this sentence'.

Particularly in cloze-type passage-based questions, there may be some words which can be put in more than one space, but only one correct combination of answers. This increases difficulty as children have to assimilate the whole passage in order to answer correctly.

Familiarity of response requirements

Question types that are familiar to children (for example because they are employed in existing Key Stage 2 assessments) tend to be less demanding to children than those that are less familiar.

Explicitness of response requirements

Particularly in the case of less familiar question types, question difficulty is also affected by the extent to which requirements of the question are obvious to children. As always, there is a compromise to be made between minimising the length of a question (in terms of number of words) and being explicit about question requirements.

Sentence Complexity

This refers to the difficulty of the target phrases or sentences.

Sentence difficulty can be manipulated in a number of ways, such as increasing the:

- length of sentences;
- number of related sentences (i.e. those that constitute a 'passage' rather than discrete sentences);
- abstractness of the ideas within phrases/ sentences; and
- difficulty of vocabulary within sentences.

These are discussed in more detail below:

Vocabulary demands

The use of non-technical language that is unfamiliar to children may increase question difficulty in questions where the relevant target sentence/passage needs to be understood.

Sentence complexity

Children will inevitably find it more difficult to manipulate or identify features in more complex sentences.

Sentence length

Longer sentences will also be more difficult for children to interpret. In particular, the use of passages rather than discrete sentences will increase difficulty due to added demands on working memory of assimilating more information.

Sentence abstractness

The abstractness of ideas contained within sentences/passages also appears to affect question difficulty. Familiar contexts will be easier for children to access.

Question difficulty scale

A question difficulty rating scale was developed, using a three-point scale for each of the three factors. The definition of the three points for each of the factors is shown in the table below.

Factor	**1** **(Low)**	**2**	**3** **(High)**
Knowledge of technical language and punctuation rules	There is no requirement to know the meaning of technical language such as word classes or identify or use types of punctuation.		Children are required to identify less common linguistic features such as prepositions and different phrase types. Also includes use of more advanced punctuation (e.g. colons, semi-colons), and questions that assess very common speech misconceptions.
Response strategy	Questions are selected response, and use a question type likely to be familiar to children (such as multiple choice or matching).		Questions are constructed response, requiring children to generate their own language. Question type may be less familiar, with instructions either complex or not entirely explicit.
Sentence complexity	Question contains either no target sentences, or very simple target sentences with familiar contexts, simple language, and short sentences.		Sentences/ phrases are complex: they may contain abstract ideas, form a passage that needs to be assimilated by children (imposing additional load on working memory), or use difficult vocabulary.

Figure 25: Description of the question difficulty scale used to rate test questions

Validation of Question difficulty scale

In order to validate the rating scale, the 45 questions from version 1 of the levels 3-5 test were rated using the scale shown on the previous page. The relationship between the question facility values (based on a sample of approximately 500 children) and the ratings was analysed.

Figure 26 below shows the correlations between the ratings given for each of the three factors and question facility. In each case, there is a strong negative relationship evident, indicating that as the factors of question difficulty increase the question facility decreases. Technical Knowledge showed the highest correlation (-0.66), with correlations of -0.40 and -0.32 for response strategy and sentence complexity respectively.

	KTL and KPR	Response strategy	Sentence difficulty
Correlation (with question facility)	-0.66**	-0.40**	-0.32*

** significant at 0.005 level   * significant at 0.05 level.

Figure 26: Correlation between facility and the three higher-level factors affecting difficulty

A regression analysis was also undertaken. The adjusted R Square for the regression is 0.61, meaning that the 3-factor model explains about 61 per cent of the variability in the difficulty of questions.

The correlations, along with the R square value, suggest that the three factors are able to explain a high proportion of the difficulty of the questions. The three point scale employed is quite simplistic, and it is quite likely that the predictive power of the three factors would be even higher if the scale was extended to four or five points.

Summary

There are two key implications that arise from this work. These are discussed in turn below.

Construct relevance of the three factors

Although construct relevance is considered in more detail in Annex 4, it is important here to consider the construct relevance of the three factors. In the case of KTL and KPR, provided that the knowledge being assessed is included within the content domain, the construct relevance of the factor cannot be questioned. The construct relevance of the other two factors, which are not directly related to the test content, are worthy of further consideration.

Response strategy

While there are advantages to including a variety of response types within an assessment, a fundamental question to consider is the extent to which children's performance for some question types is affected by difficulties in understanding the task requirements rather than because they lack the subject knowledge to answer. While many of the question types (e.g. multiple choice, matching, circling) will likely be familiar to children, there are others where they may be less certain about what they are expected to do. Only two questions in version 1 of the levels 3-5 test was given a rating of three for this factor, suggesting that only a small number of question types may carry this concern.

Uncovering precisely which question types may have had such effects is difficult in most cases because the responses required by children are often non-textual (for example ticks, circles), and such responses often tell us little about the way in which questions have been interpreted. Interviews with children did provide evidence that certain question types resulted in task requirements being difficult to interpret. These are discussed in more detail in Annex 4. It may be of value, in the near future, to conduct some further small-scale trialling of any question types where there is uncertainty regarding children's understanding of task requirements. Including such questions in sample materials for schools may be effective in ensuring children have an understanding of requirements in those questions.

Sentence difficulty

Unlike with response strategy, it is easier to identify questions where this factor was likely to have negatively impacted on children's performance. The question is whether this increased demand is construct relevant. In other words, is it valid to manipulate question difficulty by increasing sentence difficulty? Fundamentally, this is a decision that must be made by subject experts.

It may be worth providing specific guidance to question writers regarding what level of sentence difficulty is appropriate. This guidance should also be used by question writers when developing future questions in the subject. The guidance could cover:

- difficulty of vocabulary in target sentences;
- appropriate and inappropriate contexts for use in target sentences (and the relevant issue of concrete vs. abstract contexts);
- passage/sentence lengths; and
- use of complex sentences in questions not specifically assessing complex sentences.

The effect of practice on test performance

The factor that was found to correlate best with (and explain the variation in) question facility is KTL/KPR. The body of knowledge within the subject is not vast, and it is anticipated that once children are specifically taught this content, performance on questions assessing KTL in particular will increase, perhaps markedly. Because questions assessing *use* of grammar tended to be easy (for reasons explained above), there is a strong possibility that children's performance will increase markedly in the tests (even relative to the 'standard' improvements in test performance seen after the introduction of a new assessment) in the next couple of years, with most questions displaying very high facilities. If the assessment becomes largely a competency based test, with one threshold which the majority of children are expected to achieve, then this is not a concern. On the other hand, if the purpose of the assessment is to discriminate adequately between a broad range of ability (such as the case in the current levels 3-5 tests in mathematics and English reading), then consideration may need to be given for how this can be achieved in future test cycles.

Annex 6: Identified issues for children with special educational needs

This annex examines potential accessibility issues of questions, tests and administration methods for children with special educational needs (SEN). The annex includes a brief overview of literature, evidence from a small trial of questions with children with SEN and interviews with experts in the fields of visual impairment, hearing impairment and dyslexia.

Methodology

For the purpose of this research the special educational needs identified and investigated were visual impairment, hearing impairment and dyslexia. Three approaches were used to identify potential accessibility issues for children with these special educational needs. These approaches are listed below.

Literature review

A high level literature review was conducted to identify research on technical English accessibility, skills and testing for children with special educational needs including visual impairment, hearing impairment and dyslexia.

Interviews with experts

Experts in visual impairment, hearing impairment and dyslexia were interviewed by researchers from STA to identify potential accessibility issues for children with SEN. Four interviews were conducted with six experts, two for each identified educational need. Interview questions were focused on identifying general issues with the concept of testing grammar, punctuation and spelling with children with different educational needs. However, interviewees were given access to test materials in the interview, allowing them to identify specific sections and questions which may cause issues. Interviewers made detailed notes of the interviews.

Trialling of test materials

Three researchers from STA visited three schools to trial test questions. The schools selected were all secondary schools, and included a specialist hearing impaired school, a specialist school for children with visual impairments and a private school with a specialist dyslexia resource. Two researchers visited each school and three to five children participated in the trial at each school.

Booklets from the technical pre-test were trialled with children with hearing impairments and children with dyslexia. A selection of questions from the trial was converted into braille and modified large print to form a trial test for blind and partially sighted children.

Due to time constraints the children did not complete a full test, however, a sample of each section of the test was trialled at every school.

The researchers made notes of any questions which were raised by children during the test and recorded administration issues. The researchers also interviewed children following each section of the test and made detailed notes of the interview. The children's test booklets were available to refer to during these interviews. Interview questions focused on individual questions in the test and aimed to highlight any questions that were easier or more difficult than intended, particularly due to accessibility issues. In some schools researchers also interviewed teachers to collect their thoughts on the tests. Detailed notes were taken by the researcher.

Outcomes

This section is divided into three parts, each part focusing on one of the identified special educational needs:

Visual impairment

As part of the research a short literature review was conducted to ascertain whether there are any specific accessibility issues relating to the testing of grammar, punctuation and spelling with visually impaired children. Researchers from STA also met with two experts in the field of testing for visually impaired children from the Royal National Institute for Blind People (RNIB). They provided general feedback on the concept of testing grammar, punctuation and spelling with blind and partially sighted children as well as feedback on different sections of the test. It is worth highlighting that much of the research is from small scale studies, and findings are often based on trials with a small number of children. Due to the relatively low numbers of children who are blind or partially sighted it is unsurprising that sample sizes are often low, but it does mean the findings should be used with caution.

A review of literature found that some studies had been conducted into the early development of language amongst blind and visually impaired children (Andersen, Dunlea and Kekelis, 1993[20]; Dunlea, 1989[21]; Petrez-Pereira, 1994[22], 2001[23]). The research suggests that blind children develop speech formulaically, by learning chunks of language and imitating them through verbal routine. They are less likely to use an analytical approach to learning words, where a child will identify a word and begin using it with other words they have previously analysed. However, the research suggests that

[20] Andersen, E. S.; Dunlea, A. & Kekelis, L. S (1993). The impact of input: language acquisition in the visually impaired. First Language, 13, 23-49.

[21] Dunlea, A. (1989). Vision and the emergence of meaning: blind and sighted children's early language. Cambridge University Press, Cambridge, UK.

[22] Perez-Pereira, M. (1994). Imitations, repetitions, routines, and the child's analysis of language: Insights from the blind. Journal of child language, 21, 317-337.

[23] Perez- Pereira, M. (2001). First grammar in blind, visually impaired and sighted bilingual children: do they follow different routes? Research on child language acquisition, 1196-1206.

over time (and as early as the age of 3) blind children's use of imitative speech and verbal routines will have decreased. It is therefore unlikely that any differences in the approach to early language acquisition will be a factor by the time blind children are at Key Stage 2.

Whilst the research suggests visual impairment is unlikely to have a long term effect on speech development, there are still barriers to a child who has a visual impairment's development of language, vocabulary and reading skills. Experts from the RNIB suggested that blind and partially sighted children often develop language and reading skills slower than sighted children due to a lack of incidental learning. Sighted children will learn informally by seeing and reading words in everyday life, whereas blind and partially sighted children, particularly braille users, may not have as many opportunities to access written words in everyday life. This reduced level of incidental learning may have an effect on blind and partially sighted children's vocabulary and grasp of technical English.

Little research was found which looked at the grammar, punctuation or spelling skills of children with a visual impairment. One piece of research from Holland did look at the spelling skills of blind and partially sighted children when compared to their sighted peers (van Bon, Adriaansen et al. 2000[24]). The research found that whilst spelling performance was lower for students with a visual impairment, the difference between groups was relatively small. The difference narrowed with age, approaching zero towards the end of elementary school. This suggests that visual impairment only interferes with the acquisition of orthographic knowledge for a limited time.

Much research has been conducted on blind and partially sighted children's reading speed and comprehension skills. The research finds that children who are visually impaired, particularly those who read braille, read at a much slower rate than sighted children (Douglas et al, 2002[25]). Whilst sighted children's reading speed and reading level develops quickly, a child who is blind will often lag behind in terms of reading speed and reading level. However, research into visually impaired children's level of reading comprehension has had mixed results; some research suggests reading comprehension (like reading speed) is lagging (Douglas et al, 2002[26]), whereas other research suggests that whilst processing speed is slower, the level of comprehension is comparable with sighted children of the same age (Mohammed and Omar, 2011[27]).

Experts at the RNIB thought that there can be an increase in memory load for children with a visual impairment when reading and this could contribute to a reduced level of

[24] van Bon, W., Adriaansen, L., Gompel, M., & Kouwenberg, I. (2000). The reading and spelling performances of visually impaired Dutch elementary school children. Visual Impairment Research, 2, 17-31.

[25] Douglas, G., Grimley, M., Hill, E., Long, R., and Tobin, M. (2002). The use of the NARA for assessing the reading ability of children with low vision. British Journal of Visual Impairment, 20 (2), 68-75.

[26] Ibid.

[27] Mohammed, Z., and Omar, R. (2011). Comparison of reading performance between visually impaired and normally sighted students in Malaysia. British Journal of Visual Impairment, 29 (3), 196-207.

comprehension. When considering memory load it is useful to consider the three load types of cognitive load theory:

- intrinsic load – mental load requisite for completing a task;
- germane load – cognitive demands that are not necessary for gaining essential knowledge but enhance learning; and
- extraneous load – the manner in which information is presented to learners. This load can be attributed to the design of the instructional materials.

Because there is a limit to the cognitive capacity, using it to process the extraneous load reduces the capacity available to process the intrinsic load (required to complete a task) (Kettler, Elliott and Beddow, 2009[28]). This can be a particular problem for children with a visual impairment, particularly those with very low vision and those who are blind. Even a simple task, such as answering a multiple choice question, will have an increased extraneous load because a blind child cannot use visual cues. Experts at RNIB thought that tasks such as those requiring the identification of the correct grammar and punctuation from a choice of sentences will require a much higher level of memory load for blind children than sighted children. Through question modifications, such as the reduction of the number of options in a multiple choice question, STA attempt to reduce the level of extraneous load for blind and partially sighted children.

Not all blind or partially sighted children at Key Stage 2 will have had the same level of sight since birth. If a child's sight has deteriorated over time their access to text may have changed. Tobin (1994[29]) explains that students who began reading print, but converted to braille as their sight deteriorated, may be progressing satisfactorily but their reading speed and accuracy may be lagging behind their comprehension. If a child's reading speed and accuracy are below average it is possible their grasp of grammar, punctuation and spelling may also be lagging.

A person learning braille will usually start with grade 1, uncontracted braille, where each written letter is replaced with a braille letter. As a person's braille skills progress they will begin to use grade 2 'contracted' braille, where groups of letters are replaced with braille symbols. Contracted braille is also quicker to read and write than uncontracted braille. For children who are experienced braillists reverting to uncontracted braille can be difficult, particularly if the writing demands are high. However, if the child were to use braille with contractions their spelling skills would not be tested as effectively. This could cause potential issues if writing skills and spelling skills are assessed in the same task.

[28] Kettler, R. J., Elliott, S. N., Beddow, P. A. (2009). Modifying achievement test items: a theory-guided and data-based approach for better measurement of what students with disabilities know. Peabody Journal of Education: Issues of Leadership, Policy, and Organizations, 84 (4), 529-551.

[29] Tobin, M.J (1994). Assessing visually handicapped people: An introduction to test procedures. David Fulton Publishers.

Research visit

As part of the research two researchers visited a specialist secondary school for visually impaired children in Birmingham to trial test questions with blind and partially sighted children. The researchers made two visits; on the first visit the questions were trialled with three partially sighted children; on the second visit the questions were trialled with two blind children. Test questions were modified by the RNIB into modified large print for the partially sighted children and braille for the blind children. A total of 19 levels 3-5 short answer questions were trialled. These were selected to cover a range of question types. The children were also asked to undertake a spelling test, consisting of 10 spellings, and half a handwriting exercise. The children were given 40 minutes to complete the short answer section and as long as they needed for the spelling and handwriting sections. Children provided feedback after each section of the test.

Short answer questions

Overall the modified large print users and braille users liked the test. The modified large print users said that their preferred question layouts were questions where they had to circle the answer, tick a box, or join boxes. They liked the fact that there was not much writing in the test and the questions were generally easy to understand. There were, however, a few issues that the children pointed out.

There were a number of examples where the layout or format of a question was not always user-friendly. Two children taking the modified large print test thought that one of the questions which required a connective to be circled could have had more space, as they were worried that what they drew may not be accurately placed. One way of solving this could be to have larger spaces between words on questions which require the marking up of sentences. However, it still needs to read naturally and one question which increased spacing between letters and words was found to be difficult to separate words and decode when trialled.

One child found that he was not sure whether one question used the letter ‘I’ or the number ‘1’ and had to seek clarification from a teacher. He suggested that it may be easier to distinguish the letter as an ‘I’ if it had serifs (i.e. ‘I’). It is worth noting that the number 1 does have a serif at the top in the font used in the text.

Both braille and modified large print users found they missed, or misunderstood parts of some questions, and in some cases this was due to format or layout. There were a number of examples where braille users had not seen parts of the question until after they had begun answering. In one example, both braillists thought they had to write an explanation because they did not realise it was a multiple choice question. One child realised in time, the other only realised after he had written an answer. The children said it would have been useful if the question had highlighted that there were multiple choice options, for example have a line that said ‘choose from A, B, C or D’. The suggestion from the children seems an appropriate modification to highlight the fact it is a multiple choice question.

The use of examples also confused the braille users for two questions. In a modified version of a connect-the-box question, one braille user was confused by the question layout, and did not like having a list followed by another list. The children said the example was confusing; after the two lists it said ‘one has been done for you’ but the ‘1C’ did not make sense to the children, who turned over the page looking for the example. One child left the question out, the other child eventually realised what he had to do, but wrote the words next to each other rather than the number and letter.

Children taking the modified large print test said that a question was difficult because the commas in the sentence were a bit too small, and they had to concentrate really hard to see the difference between the sentences. This was not a problem for the braille users though, who said the question made them think but they did not have any problems holding the information in their memory.

Experts from the RNIB also raised some issues with the print size and font, particularly where the print size a partially sighted child normally uses is not available in the test. Whilst the child may be able to read in a smaller print size than they are used to, the punctuation may be less easy to differentiate. By nature punctuation is often small anyway and if the text is smaller than the child is used to punctuation may become too small for the child to identify and use.

As discussed in the literature review, experts at the RNIB explained that some question formats are not as instant a format in braille as they are in print. Connecting boxes and tables are examples of formats which can be more difficult for braille users as there is more memory involved when attempting the questions. There may be a similar impact for children using modified large print, depending on the nature of their sight problems and size of text required.

Spelling

A shorter version of the spelling tests from the trial booklets was used with the blind and partially sighted children. Due to time constraints the children were asked to complete 10 spellings rather than 20. The children completed the single sentence version of the spelling test, but feedback was also sort on the alternative passage version. A researcher from STA read out the spellings.

Braille users completed the spelling test using grade 1, uncontracted braille. Using grade 1 braille for spelling tests is a standard approach in schools and children are used to using it in this context. The children answered onto a blank piece of paper, so did not have the sentence to read. It was noticeable that the braillers which the children used were quite noisy. This makes it difficult to know when to start and stop reading the sentence so that it was not competing with the noise of the brailler. For the second read through the researcher waited until the children had stopped using the brailler, but it was noticeable that the children had finished writing and were just waiting for the next sentence. The children did not see the noise of the brailler as a problem and the teacher said that other children in the class get used to the noise and it does not distract them.

Modified large print users answered on an enlarged version of the answer sheet. Whilst extra time was required to allow the children to locate the next spelling none of the children needed any help to find the correct question and answer space. The children found the numbers useful for locating where they were, and one child said locating where a new sentence starts would have been difficult without the numbers. The same child explained that normally they just wrote their spellings on a blank piece of paper, but she preferred this approach and she thought that the sentences helped when she did not hear the word clearly.

The children were told about the alternative ‘passage’ spelling test, but the children thought this sounded harder as it would be difficult to keep their place. This opinion was echoed by the experts at the RNIB who thought that the single sentence spellings would be easier to locate. They also thought that the numbers could be used as markers and there was less chance of children getting lost.

Experts from the RNIB also thought that an administrator read spelling test was preferable as a CD test has a set length of time for the pause between readings. Because blind and partially sighted children are often slower at writing the gap may not be long enough. An administrator can adjust the gap to ensure everyone has finished writing, which is not possible with a pre-recorded reading.

The RNIB representatives also explained that there is the potential for misspelling when using a brailler because it is possible to hit the wrong key by mistake. If this happens it is very difficult to decipher what the child was trying to type, as it can come out completely different if one keystroke is incorrect.

Handwriting

The dictated handwriting task was delivered in line with the administrators’ instructions for both the modified large print users and the braille users. However, only the first three sentences of the test were read, and the researcher reading the script allowed children to ask for sentences to be re-read where required (each sentence was read one extra time).

Initially the researchers considered not conducting the handwriting test with braille users but the RNIB suggested it should be conducted as braille is their writing medium. Whilst cursive writing and neatness are harder to assess, a different mark scheme could ensure the test is useful for rating a child’s writing progress.

The two braille users were asked to complete the handwriting portion of the test using grade 1 ‘uncontracted’ braille. They found this extremely difficult as they were both experienced grade 2 (contracted) braille users and both slipped into contracted braille in parts of the test because it was more natural. They found it more difficult to keep up with their writing when using uncontracted braille and it was clear they struggled to store the sentence in their memory long enough to write it out.

Each child asked for at least one sentence to be re-read. One teacher suggested that being asked to complete the task in grade 1 braille placed extra pressure on the child through an increased cognitive load. However, if the child were to use braille with contractions their spelling skills would not be tested as effectively. This could cause potential issues in some sections of the test such as handwriting, writing and spelling, as children will need to be told where and when to use uncontracted braille. One suggestion from a teacher was to allow children to complete the task in their preferred contracted braille, but ask them to spell some key words in uncontracted braille after the dictation to test their spelling.

The modified large print users' writing speeds were also quite slow; the child with the quickest writing speed was the child with the lowest level of sight loss, the slowest was the child with the highest level of sight loss. The children had some accuracy mistakes, and had missed out various punctuation marks. The children struggled to hold the whole sentence in their head whilst writing, and therefore asked for sentences to be re-read. It is therefore likely that the children's accuracy may have been improved because of the extra read through for each sentence.

Other comments

Blind and partially sighted children are not a homogenous group; they will have a variety of visual impairments and children's preferred way of accessing text can vary considerably. Blind and visually impaired children are therefore likely to access text in a variety of ways; some will use the original text with their own access technology or with a reader/scribe, some will use varying levels of enlarged print or modified large print and some will use braille.

STA currently provide tests in large print, modified large print and braille, but there are likely to be some children who still struggle to access the tests in these formats. Some schools will request early opening of tests so they can edit them into an appropriate format for their students. Experts at the RNIB suggested that papers could be sent out to schools electronically on request. This would help schools who need to enlarge the paper for visually impaired children. By providing teachers with an electronic copy of the test they can edit the tests more easily and quickly (as they will not need to scan the paper to create an electronic copy). If the test is provided in an appropriate format it would also allow staff at schools to edit the test so it is in the appropriate format for the individual child.

Experts at the RNIB thought that when reporting marks, it would be very useful to be able to isolate the different sections of the test, allowing teachers to see how children scored on different tasks and how their scores on these made up their overall score. This is particularly important to teachers when asked to factor the results of the test into an assessment of a child's overall writing level, as a visual impairment may have a negative impact on some aspects of their test performance.

Hearing impairment

As part of the research, a literature review was conducted to identify specific accessibility issues relating to the testing of grammar, punctuation and spelling with children with a hearing impairment. While there is a wealth of research that looks at relevant cognitive deficits for deaf children or children with profound hearing difficulties, there is considerably less research focusing on children with mild to severe hearing impairments (Moeller et al. 2007[30]). In addition, while research tends to show a close relationship between the extent of any cognitive deficit and degree of hearing impairment, findings from research specifically focused on children with mild or moderate hearing impairments tend to be somewhat equivocal, with some studies showing no deficit. Finally, it is worth pointing out that many studies are based on very limited sample sizes and/or use non age-matched participants. While this is not entirely surprising given the relative lack of children with a hearing impairment in mainstream schools, it does mean that, in tandem with the lack of research on children with less severe hearing impairments, any findings should be treated with some caution. Most of the relevant research relates to more general linguistic skills and capacities, such as reading comprehension and working memory. These are discussed in turn below.

Vocabulary

Most studies suggest that children with mild to severe hearing loss perform less well on standardised vocabulary assessments than children with no hearing impairment (NH). Some research suggests that even very mild hearing loss will delay vocabulary development (e.g. Davis et al. 1986[31] and Wake et al. 2004[32]), but this view is by no means unanimous, and other studies conclude that many children with mild to moderate hearing loss perform comparably to age-matched peers with NH (Gilbertson & Kamhi, 1995[33]).

Coppens et al. (2011[34]) examined the vocabulary knowledge of children aged eight to 11, and in particular the relative reading vocabulary disadvantage of children with a hearing impairment. The performance of 394 children with NH and 106 children with a hearing

[30] Moeller, M., Tomblin, B., Yoshinaga-Itano, McDonald Connor, C., Jerger, S. (2007). Current state of knowledge: language and literacy of children with hearing impairment. Ear Hear, 28 (6), 740-53.

[31] Davis, J., Elfenbein, J., Schum, R., Bentler, R. (1986). Effects of mild and moderate hearing impairments on language, educational, and psychological behaviour of children. Journal of Speech and Hearing Disorders, 51, 53–62.

[32] Wake M, Hughes E, Poulakis Z, Collins C and Rickards W., (2004). Outcomes of children with mild-profound congenital hearing impairment at 7-8 years: a population study. Ear and Hearing 25, 1-8.

[33] Gilbertson M, Kamhi A (1995). Novel word learning in children with hearing impairment. Journal of Speech and Hearing Research. 38, 630–642.

[34] Coppens, K., Tellings, A., Verhoeven, L., & Schreuder, R. (2011). Depth of reading vocabulary in hearing and hearing-impaired children. Reading and Writing, 24 (4), 463-477.

impairment was examined on two vocabulary assessment tasks. The results showed that most NH children reached the expected norm, whereas most children with a hearing impairment did not. In addition, results showed that children with a hearing-impairment not only knew fewer words, but that they also knew them less well.

According to American Speech-Language-Hearing Association (ASHA[35]), vocabulary develops more slowly in children who have hearing loss. They argue that children with hearing loss learn concrete words like ‘cat’ more easily than abstract words like ‘before’ or ‘jealous’, and have difficulty understanding words with multiple meanings. They also state that the gap between the vocabulary of children with normal hearing and those with hearing loss widens with age. According to many researchers, a key cause of any vocabulary deficit lies in the fact that hearing impaired children encounter fewer words than NH children because they are relatively deprived of linguistic input, at least with respect to spoken language.

Reading comprehension

Reading comprehension is ‘the active process of constructing meaning from text; it involves accessing previous knowledge, understanding vocabulary and concepts, making inferences and linking key ideas’ (Vaughn & Linan-Thompson, 2004[36]). Clearly, these skills are central to the success of children taking any written assessment.

Luckner and Handley (2008[37]), after an analysis of the literature, suggests that deaf or hard of hearing children tend to demonstrate one or several of the following behaviours:

- effortful word recognition;
- limited vocabulary;
- a lack of understanding of figurative language;
- weak topic knowledge;
- a slow reading rate;
- inadequate understanding of syntax;
- limited knowledge of different genres;
- a lack of awareness of text organisation;
- a limited repertoire of comprehension strategies;
- failure to monitor comprehension;
- lack of motivation; and
- avoidance of reading as much as possible.

The literature suggests that in general children with a hearing impairment show lower levels of reading comprehension than their hearing peers. For example, Wauters et al.

[35] Effects of Hearing Loss on Development. American Speech-Language-Hearing Association (ASHA). Retrieved November 25 2012, accessed at: http://www.asha.org/public/hearing/disorders/effects.htm

[36] Vaughn, S., & Linan-Thompson, S. (2004). Research-based methods of reading instruction, grades K–3. Alexandria, VA: Association for Supervision and Curriculum Development.

[37] Luckner, J., and Handley, C. (2008). A summary of the reading comprehension research undertaken with students who are deaf or hard of hearing. American Annals of the Deaf, 153 (1).

(2006[38]) showed that on average hearing impaired participants between seven and 20 years old performed at the reading comprehension level of seven-year-old hearing participants. They also found that only 4 per cent of the students with hearing impairments in their study were reading at an age-appropriate level. Reading difficulties for children with hearing impairments tend to be linked to a number of factors, including the vocabulary deficit described in the previous section.

ASHA[39] state that children with hearing loss:

- have poorer comprehension than children with normal hearing; and
- often have difficulty understanding complex sentences, such as those with relative clauses (e.g. 'the teacher whom I have for math was sick today') or passive voice (e.g. 'the ball was thrown by Mary').

Working memory

Working memory is considered to impinge on a range of linguistic skills, including vocabulary acquisition, sentence comprehension and reading. Marschark et al. (2011[40]) argue that 'studies of memory consistently indicate that deaf or hearing impaired individuals have shorter memory spans than hearing age-mates. That is, when questions have to be remembered in a particular order, hearing impaired children and adults will remember fewer of them than hearing age mates'. In addition, Alamargot et al. (2007[41]) found that deaf students had lower phonological (memorising letter series) and executive (written production span) capacities than hearing pupils.

Grammar and punctuation

There is evidence from several studies that children with mild to severe hearing loss experience delays in morphological development, including the acquisition of graphical morphemes (markers such as the past participle 'ed' used in the past tense, the present participle 'ing' used in the present progressive, or third person singular 's'). This, argues ASHA[42], leads to misunderstandings and misuse of verb tense, pluralisation, non-agreement of subject and verb, and possessives.

In a study by McGuckian and Henry (2007[43]) a group with hearing impairments produced possessive -s and plural -s significantly less frequently than the controls but produced progressive -ing, articles and irregular past tense significantly more frequently than the

[38] Wauters, L., van Bon, W., Tellings, A., van Leeuwe, J. (2006). In search of factors in deaf and hearing children's reading comprehension. American Annals of the Deaf, 151(3), 371-380.
[39] "Effects of hearing loss on development". American Speech-Language-Hearing Association (ASHA). Retrieved November 25 2012, from: http://www.asha.org/public/hearing/disorders/effects.htm
[40] Marschark, M., Spencer, P., Adams, J., Sapere, P (2011) Teaching to the strengths and needs of deaf and hard-of-hearing children. European Journal of Special Needs Education , 26 (1), 17–23
[41] Alamargot, D., Lambert, E., Thebault, C, Dansac, C. (2007). Text composition by deaf and hearing middle-school students: The role of working memory. Reading and Writing, 20, 333–360.
[42] "Effects of hearing loss on development". American Speech-Language-Hearing Association (ASHA). Retrieved November 25 2012, from: http://www.asha.org/public/hearing/disorders/effects.htm
[43] McGukian, M., & Henry, A. (2007). The grammatical morpheme deficit in moderate hearing impairment. International Journal of Language & Communication Disorders, 42 (S1), 17–36.

controls. One possible explanation they raised is that of ‘perceptual saliency’: the more perceptually salient a graphical morpheme is, the more often it will be perceived and the easier it will be acquired.

Writing

Antia et al. (2005[44]), in an analysis of the literature, argue that:

- because of difficulty accessing and learning English syntactical and morphological structures, deaf or hard of hearing pupils make numerous errors at the sentence level;
- because deaf or hard of hearing pupils have difficulty with reading, their exposure to models of good writing may be limited; and that
- because teachers of deaf or hard of hearing children often emphasise an approach to writing that focuses on producing basic sentences, their writing may be uninteresting, uninformative and not coherent.

While stressing that the vast majority of research focuses on children who are deaf or have profound hearing loss, they add that the research suggests that:

- While the grammatical complexity of deaf and hard of hearing students’ writing may increase over time, current research shows that they may experience difficulty with grammatical constructions throughout their school years.
- Deaf and hard of hearing students may exhibit difficulty with cohesion of ideas in writing, with pupils able to communicate main ideas but without additional elaboration or detail.

Yoshinaga-Itano and Downey (1996[45]) examined the written language of 461 pupils who were deaf or hard of hearing and 94 hearing students aged between seven and 18. They reported increased delays in written language with increased degree of hearing loss. They also reported students with mild and moderate hearing losses were delayed in written language compared to hearing peers up to age 13, but showed performance similar to hearing peers by high school. On the other hand, students with moderate or severe hearing loss showed delays compared to hearing peers at all ages, with the delay growing progressively greater as hearing loss increased.

ASHA[46] state that children with hearing loss:

- produce shorter and simpler sentences than children with no hearing loss; and
- often have difficulty writing complex sentences, such as those with relative clauses (e.g. ‘the teacher whom I have for math was sick today’) or passive voice (e.g. ‘the ball was thrown by Mary’).

[44] Antia, S., Reed, S., & Kreimeyer, K. (2005). Written language of deaf and hard-of-hearing students in public schools. Journal of Deaf Studies and Deaf Education, 10 (3), 244-255.

[45] Yoshinaga-Itano, C., & Downey, D. M. (1996). The psychoeducational characteristics of school-aged students in Colorado with educationally significant hearing losses. The Volta Review, 98, 65–96.

[46] “Effects of hearing loss on development”. American Speech-Language-Hearing Association (ASHA). Retrieved November 25 2012, from: http://www.asha.org/public/hearing/disorders/effects.htm

Small scale trial and expert interviews

As part of this research, a small scale trial was carried out with four Year 7 children from a mainstream secondary school with a specialist hearing impairment unit. The trial lasted an hour, and during this time, the children attempted a range of questions from the short answer questions section of one of the technical pre-test booklets, as well as briefly attempting versions of the handwriting and spelling assessments. STA researchers observed the children during this time, and led guided discussions to obtain their feedback between each section of the test.

In addition, researchers from STA conducted one extended interview with the head of the sensory education service of a Greater London Borough. (This interviewee is referred to as the primary interviewee, or PI, in the remainder of this section). The interview lasted for two and a half hours, including approximately forty five minutes for the interviewee to familiarise herself with samples of the range of test materials used in the technical pre-test at levels 3-5 and level 6. In addition, a shorter (forty five minute interview) was conducted with the two specialists who ran the unit for children with a hearing impairment used in the small scale trial.

The expert interviewee supported evidence from the literature review by stating that children with hearing impairments were likely to have much more limited vocabularies, and made the case that any non-technical language should be simplified as far as possible. While this may be considered standard practice in written assessment in general (non-technical language that impedes performance would normally be considered a source of CIV), the threshold beyond which she felt that language may impede performance for hearing impaired children appeared to be lower than would normally be the case in Key Stage 2 assessments.

In relation to the technical pre-test short answer questions, it is clear that any vocabulary deficit for children with a hearing impairment would affect performance in questions specifically assessing vocabulary, as well as any questions using unfamiliar language that assess other elements of the programme of study. When the expert looked through the short answer questions in one of the tests, it was clear that she felt language could be simplified in both the question stems and target sentences. For example, she recommended using the word ‘sentences’ instead of ‘passage’.

The PI was also concerned that difficulties children with a hearing impairment have with comprehension and vocabulary would mean that interpreting the requirements of the various question types used in the short answer questions would be very challenging for children.

All three interviewees raised a number of concerns regarding children’s ability to understand question requirements, reflecting evidence from the literature regarding the reading comprehension abilities of children with hearing impairments. The PI felt that in the short answer questions, task requirements were often likely to be too complex for many hearing impaired children to follow, even when they possessed the required subject knowledge. Of particular concern were more innovative question types and questions containing relatively large amounts of text describing what children were required to do. It is important to state that the issues raised were not with the question types themselves, but rather in comprehending written instructions within questions.

These concerns were also evident in the small scale trial. At least two of the children tended to avoid reading any of the question stems, attempting instead to infer the task requirements from the layout of the question. For many question types, including multiple choice and matching questions, this was a largely effective strategy.

All three expert interviewees were also concerned that the quantity and complexity of instructions that were provided in the test booklets (particularly the ‘General instructions’ on page 3) were inappropriate, with the PI describing them as ‘very intimidating’. All were concerned that children would fail to assimilate most of the information, and the PI suggested the instructions should be removed entirely. If the instructions were deemed mandatory, she showed a strong preference for diagrammatic exemplification rather than textual description of question types.

Although it is not possible to separate any potential vocabulary acquisition or reading comprehension deficit with a working memory deficit, there were a number of questions where two children in the small scale trial appeared to find the amount of information that needed to be assimilated excessive.

The PI suggested that because those with hearing impairments would be likely to have difficulties comprehending question requirements:

- it would be more important for questions to be ordered by question type rather than by anticipated increasing difficulty; and
- it would be more effective to use bold text for key instructions (e.g. command words) in questions, rather than for word types and other technical terms.

The PI supported evidence from the literature review suggesting that children with mild to severe hearing loss experience delays in morphological development. For example, she pointed out that children with a hearing impairment tend to struggle to distinguish between tenses due to the subtlety in differences between words such as ‘she’s’ and ‘she’ll’.

All three interviews suggested it was highly unlikely that any children with moderate or severe hearing impairments would take the level 6 test, which contains a writing task. However, the PI suggested that the sorts of writing deficits discussed in the literature review would also affect performance of children with a hearing impairment in the open response questions within the short answer questions section.

All three interviewees supported the main finding from the literature review that primary level children with a hearing impairment would demonstrate poorer spelling than same-age non-hearing impaired children. In addition, they felt that children with a hearing impairment would show a preference for the individual sentence rather than passage version of the spelling test. The PI felt that when children were faced with challenging-to-spell words, their confidence would be affected more strongly in the passage than in the individual sentence version. This was because in the latter the fact that the each question was discrete would give the feeling of 'being able to start again'. In addition, she also felt that the sentences should be as short as possible (while still providing sufficient context to aid children's understanding of the target word). She also suggested an alternative model where rather than use a whole sentence for each target word, children are simply given the target word plus one or two synonyms. A small number of questions in one administrator-read sentence and one administrator-read passage spelling test were trialled. The children showed a preference for the individual sentence version, for the reasons predicted by the experts.

The experts all showed a strong preference for the copying rather than dictation handwriting task. The PI made the point that any working memory limitations would mean that even the task of copying text may be slower for many children with a hearing impairment compared with NH children, however concern regarding the accessibility of the dictation task was far greater. This was due to the multiple cognitive and literacy-related challenges involved in the task (for example vocabulary knowledge, working memory demands and spelling), many of which are potential sources of performance deficit for children with a hearing impairment. This concern was exacerbated by the fact that the task is administered under strict time conditions.

Dyslexia

Literature and expert review

As part of the research, a literature review was conducted. This looked at the difficulties associated with dyslexia that might have an effect on performance in different content areas, different question formats and different types of administration. Researchers from STA also met with two academic experts in the field of dyslexia, one from Oxford University and one from the Dyslexia Trust. They provided insight into the development of grammar, punctuation and spelling skills in children with dyslexia, as well as feedback on different sections of the test and individual test questions. The findings of these two research strands are summarised below.

Although dyslexia is primarily associated with difficulties in reading and spelling, a range of "downstream effects" are observed in dyslexics' written and spoken language (Shaywitz et al. 2008[47]). Research has found that children with dyslexia are less skilled at producing complex sentences than age-matched peers (Puranik et al., 2007[48]; Du Pre et

[47] Shaywitz, S., Morris, R. & Shaywitz, B. (2008). The education of dyslexic children from childhood to young adulthood, Annual Review of Psychology, 59(1), 451-475.

[48] Puranik, C., Lombardino, L. & Altmann, J. (2007) Writing through retellings: an exploratory study of language-impaired and dyslexic populations, Reading and Writing Interdisciplinary Journal 20(3), 251-272.

al. 2008[49]) and that dyslexic children's written work may have omitted words, sentences that barely make sense and incorrect uses of tenses and prepositions (Poustie et al. 1998[50]). Other research has found that dyslexic writers frequently fail to use enough connective words and that people with dyslexia tend to write sentences with words in the wrong order (Hornsby 1997[51]).

Whilst the experts tended to agree with much of the research findings in the literature there were some findings which the experts challenged or disagreed with. For example, some research found that children with a reading disability (such as dyslexia) showed a significant lag in the development of grammatical sensitivity (Siegel, & Ryan, 1988[52]) but this was challenged by the interviewees. Similarly, some sources conclude that writing in an impersonal style can be difficult for dyslexics (Du Pre et al. 2008[53]) although the interviewees disputed this, and thought it posed no greater problems for those with dyslexia.

One area where there is a lack of current research evidence is on dyslexia and punctuation. It is widely acknowledged in the literature that dyslexic writers' use of capital letters and punctuation marks is poor (Reid, 2007[54]), but few empirical or specific studies are available. Following consultation with academic experts, some further undergraduate-level research may become available in this area within the next two years.

In terms of question types, the academic experts raised a concern about 'matching' tasks. Their primary issue was that children with dyslexia find the visual scanning, working memory and grapho-motor demands of this question type problematic. Furthermore, children with dyslexia find it difficult to follow the lines that they have already drawn in order to check their answer or eliminate parts of the question already used. An additional concern was raised about the use of the instruction word 'match', which implies (particularly to weaker readers, who may be more literal in their interpretation) that the two parts to be 'matched' are exactly alike. The over-arching concern about the question type was reflected in subsequent teacher interviews and the trial with children.

There is also some literature which suggests that children with dyslexia find multiple-choice questions difficult, due to associated working memory problems and the increased likelihood of reading inaccuracies when shifting visual focus up and down as well as left

[49] Du Pre, L. Gilroy, D. & Miles, T. (2008) Dyslexia at college, Routledge.
[50] Poustie, J. et al (1998) Solutions for specific learning difficulties, 3rd ed., Great Britain: A Next Generation Publication.
[51] Hornsby, B. (1997) Overcoming dyslexia, Vermillion/Prospect House.
[52] Siegel, L. & Ryan, E. (1988) Development of grammatical-sensitivity, phonological, and short-term memory skills in normally achieving and learning disabled children, Developmental Psychology 24(1), accessed at http://psycnet.apa.org/journals/dev/24/1/28/
[53] Du Pre, L. Gilroy, D. & Miles, T. (2008) Dyslexia at college, Routledge.
[54] Reid, G.(2007) Dyslexia (The Sage Handbook of Dyslexia), SAGE.

to right (Rack, 2008[55]). This is a matter of controversy within the dyslexia community, and was not supported by the experts consulted.

Of the research that has been conducted on performance across different question formats no significant differences were found in dyslexics' performance, although this was a small sample of GCSE science candidates (Crisp et al. 2011[56]). Overall:

- Bullet points - while both dyslexic and non-dyslexic students said they preferred longer texts to be broken up into bullet points, exam performance was inconclusive. One analysis showed that dyslexic students performed better without bullet points.
- Tick boxes - dyslexic students benefited from answering in a tick box format rather than having to write the answer, but so did the control non-dyslexic group, although to a lesser extent.

It is widely recognised that poor spelling is a characteristic of dyslexia (Puranik et al. 2007[57]; Connelly et al. 2011[58]; Berninger et al. 2008[59]; Sterling et al. 1997[60]; Coleman et al. 2009[61]). The specific spelling errors which are common in dyslexia include:

- phonological errors (e.g. f/ph);
- inconsistent use of letters with similar sounds (e.g. s/z);
- incorrect word endings, especially where a choice of 'y' or 'ie' is required; and
- double consonants incorrectly used or omitted (Reid, 2007[62]).

There is a lack of consensus in the literature as to whether dyslexic children have impaired handwriting (Berninger et al. 2008[63]) or not (Connelly et al. 2011[64]). This dichotomy was confirmed by the academic experts consulted. Having examined the literature, it appears that some children with dyslexia have poor handwriting, whilst others have neat handwriting but an excessively slow speed of production (Rose, 2009[65]; Reid,

[55] Dr John Rack, Dyslexia Action, and Sue Flohr, BDA, quoted in BBC news article 29 July 2008, accessed at http://news.bbc.co.uk/1/hi/magazine/7531132.stm

[56] Crisp, V., Johnson, M. & Novaković, N. (2011) The effects of features of examination questions on the performance of students with dyslexia, British Educational Research Journal.

[57] Puranik, C., Lombardino, L. & Altmann, J. (2007) Writing through retellings: an exploratory study of language-impaired and dyslexic populations, Reading and Writing Interdisciplinary Journal 20(3), 251-272.

[58] Connelly, V., Dockrell, J. & Barnett, A. (2011) Children challenged by writing due to language and motor difficulties, accessed at http://psych.brookes.ac.uk/ewsc/connelly2011.pdf

[59] Berninger, V., Nielson, K. & Abbott, R. (2008) Writing problems in developmental dyslexia: under-recognized and under-treated, Journal of School Psychology 46(1), accessed at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2344144/

[60] Sterling, C., Farmer, M., Riddick, B., Morgan, S. & Matthews, C. (1997) Adult dyslexic writing. Dyslexia 4, 1-15.

[61] Coleman, C., Gregg, N., McLain, L. & Bellair, L.W. (2009) A comparison of spelling performance across young adults with and without dyslexia. Assessment for effective intervention, 34, 94-105.

[62] Reid, G.(2007) Dyslexia (The Sage Handbook of Dyslexia), SAGE.

[63] Berninger, V., Nielson, K. & Abbott, R. (2008) Writing problems in developmental dyslexia: under-recognized and under-treated, Journal of School Psychology 46(1), accessed at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2344144/

[64] Connelly, V., Dockrell, J. & Barnett, A. (2011) Children challenged by writing due to language and motor difficulties, accessed at http://psych.brookes.ac.uk/ewsc/connelly2011.pdf

[65] Rose, J. (2009) Identifying and teaching children and young people with dyslexia: an independent report for the Secretary of State for Children, schools and families, HMSO.

2007[66]). Dyslexia is also associated with limitations of short-term memory (Siegel, & Ryan, 1988[67]). In a dictation based handwriting test the combination of slow handwriting and lower levels of short term memory could have a negative effect on performance in assessments.

Thirty-five to forty per cent of dyslexic children also suffer from Specific Language Impairments (SLI) (Rose, 2009[68]), which is linked to slow handwriting. Dockrell (2009[69]) suggests that this is probably due to their inability to coordinate and manage linguistic information. The research did not suggest how much slower a child with SLI at Key Stage 2 may write, although Graham et al. (1998[70]) found that at age 16, the handwriting speed of SLI sufferers was typical of children seven years younger. Although much of the literature surrounding SLI focuses on oral communication difficulties, the condition is also known to affect written language (Connelly et al. 2011[71]). Children with SLI are particularly prone to making verb errors, particularly in forming the regular past tense, third person singular verb forms and the use of auxiliary verbs.

Dysgraphia and graphomotor dyspraxia are distinct specific learning difficulties that impair handwriting, and may be present alongside dyslexia. Comorbidity rates are discussed in Pauc (2005[72]). Children with dysgraphia may find writing physically painful. Their work is characterised by absent joints, the collision of letters, insufficient word pacing and ambiguous letter forms. The difficulties these children have with the mechanics of writing mean that they may not be able to reproduce writing at speed or from memory (Poustie, 1998[73]).

The difficulties listed above caused the experts consulted to recommend that handwriting is not included in this assessment. They also endorsed the principle that wherever possible the questions in this test should be selected-response or require only minimal writing.

Research visit

A visit to a secondary school in Oxfordshire was undertaken by two STA researchers, in order to trial test questions with pupils with dyslexia.

[66] Reid, G.(2007) Dyslexia (The Sage Handbook of Dyslexia), SAGE.
[67] Siegel, L. & Ryan, E. (1988) Development of grammatical-sensitivity, phonological, and short-term memory skills in normally achieving and learning disabled children, Developmental Psychology 24(1), accessed at http://psycnet.apa.org/journals/dev/24/1/28/
[68] Rose, J. (2009) Identifying and teaching children and young people with dyslexia: an independent report for the Secretary of State for Children, schools and families, HMSO.
[69] Dockrell, J. E., Lindsay, G., & Connelly, V. (2009) The impact of specific language impairment on adolescents' written text. Exceptional Children, 75(4), 427-446.
[70] Graham, S., Berninger, V. W., Weintraub, N., & Schafer, W. (1998) Development of handwriting speed and legibility in grades 1–9. Journal of Educational Research, 92, 42–56.
[71] Connelly, V., Dockrell, J. & Barnett, A. (2011) Children challenged by writing due to language and motor difficulties, accessed at http://psych.brookes.ac.uk/ewsc/connelly2011.pdf
[72] Pauc R. (2005) Comorbidity of dyslexia, dyspraxia, attention deficit disorder (ADD), attention deficit hyperactive disorder (ADHD), obsessive compulsive disorder (OCD) and Tourette's syndrome in children: A prospective epidemiological study; Clinical Chiropractic (2005) 8, 189—198.
[73] Poustie, J. et al (1998) Solutions for Specific Learning Difficulties, 3rd ed., Great Britain: A Next Generation Publication.

At the start of the visit, two specialist dyslexia teachers were given the opportunity to review a test booklet and make some initial comments. They were asked to predict which questions might contain problematic curriculum content or question formats.

Three Year 7 pupils with dyslexia were then given 40 minutes to work through a levels 3-5 test booklet. They were not expected to finish the questions in this time. The children were also given two extracts from spelling tests, each consisting of five spellings due to time constraints: one was presented as a passage and one was presented as discrete sentences. Finally, the children attempted one sentence of a handwriting task, delivered by dictation.

STA researchers observed the children during the written tasks, and led guided discussions to obtain their feedback between each section of the test. After the trialling, STA researchers reviewed the children's responses with the two dyslexia teachers.

Short answer questions

Specific questions arose as pupils worked through the booklets. The pupils attempted an average of 30 questions each, during which time they reported a total of 11 issues in which they were unable to understand the question. Some broad themes can be identified by grouping the questions raised during testing together with the results of the discussion which followed.

Page layout and design appeared to impact on the pupil experience. Pupils reported that they would like one question per page, to help them identify where each new question started. Several instances of pupils erroneously attempting to use part of a question from the top of a page to answer an unrelated question further down the same page were observed during the post-test analysis of the pupils' responses by specialist teachers. It was felt that a larger gap between questions would be advantageous for dyslexic pupils. Many are weaker readers and are therefore less able to scan accurately when there is a large amount of text on a page.

Question design is a further area of concern, particularly where the layout was unfamiliar to children from classroom practice and other National Curriculum assessments.

In some questions, *poor reading skills* appeared to affect pupil performance. In interview, the teachers believed that more able readers would be advantaged in this test as they would be better at interpreting the instructions and the target vocabulary. This concern may be mitigated in part by the provision of a reader during the test, where this is permitted as an access arrangement for eligible pupils as defined by the Key Stage 2 *Assessment and reporting arrangements*. However, it is not possible to counteract the lesser exposure to literature, particularly that containing more complex vocabulary and grammatical features, which may be a result of some dyslexic children's difficulties with reading.

Examples of reading difficulties were also gathered during pupil observation.

- In one question, pupils had to locate a ‘comma’. Two pupils independently read this to researchers as ‘coma’, and were subsequently confused by the request.
- Another question was omitted by two pupils, who were apparently put off after being unable to read an unusual word in the target sentence. This was despite this content being unrelated to the grammatical operation that they were asked to perform. The third child was also unable to decode the word, but still attempted the question.
- In other questions, children ignored the question instructions so that, in one instance, instead of adding a suffix to a word, the pupil wrote a synonym.

As predicted by academic experts and teacher interviews, *matching* questions appeared problematic. One matching question was not reached by one pupil and omitted by one pupil. The remaining pupil attempted the question, albeit incorrectly and with a great deal of crossing-out.

It appeared that little use was made of *examples* within the test. These were intended to scaffold answers in some areas of content and support pupil interpretation of question response formats. During the pupil discussion, several comments were made that children ‘didn’t see’ the example given. It may be that, for dyslexic pupils, an example appears to add additional textual clutter on the page and is skimmed over; the relevance of given text is not drawn out.

During the pupil discussion, researchers drew the children’s attention to the example given for a particular question and read it aloud with them. When questioned about whether they now knew the answer to the question, all three children could supply the correct response, despite their inability to do so prior to the example having been read aloud.

Spelling

In the pre-interview, teachers anticipated that pupils would prefer the sentence version of the spelling test. Two of the three pupils interviewed reported a preference for the passage version, because it was “all about one topic” and “more like writing a story”. The third pupil preferred sentences because you could hear the same word three times. However, in addition to limitations of the very small sample size of this research, it should also be noted that the children attempted the spelling test versions at the end of the day and after a long test session. Levels of concentration and fatigue may therefore have been influencing factors.

Teachers expressed concern that the poor handwriting associated with dyslexia might further disadvantage pupils in this section of the test. Despite handwriting it not being the skill assessed, markers may be unable to read what has been written, particularly where time is limited.

Handwriting

Only a very short part of the dictated handwriting task was attempted during the research visit. All three pupils found it impossible to transcribe more than two consecutive words with any accuracy of recall (aside from issues of spelling or punctuation). During the following discussion, all three pupils felt they had ‘done their best’ handwriting during the task, but subsequent teacher reflections suggested that the children would have been capable of better performance in other circumstances.

Summary

The most problematic questions for children with SEN were those with unfamiliar language, complex or unclear instructions, a high word count and high working memory requirement. Questions with an unfamiliar layout and questions being too close together on the page were also challenging. However, the number of questions highlighted as being problematic was generally low. This was in part due to the work already done to make the questions clear, concise and with simple language.

Many of the issues raised are also contributors to CIV. The work done on reducing this and producing questions with simple language, clear instructions and in an accessible layout will help to make questions more accessible. More details on research looking at construct irrelevant variance in test questions can be found in Annex 4.

The researchers made it clear that just because questions were included in the report it did not mean they should not be considered for selection. The experts involved in test selection could use the findings from the research to inform their decisions on question selection. The findings from this research were therefore used during question selection and test construction for the 2013 English grammar, punctuation and spelling test. Based on issues found during the research some questions have been changed and were available for selection. Other questions were not considered for the 2013 test, and will be reviewed and potentially modified before they are included in future years.

Standards
& Testing
Agency

Any enquiries regarding this publication should be sent to us at assessments@education.gov.uk.

This document is also available from our website at: www.education.gov.uk.