Standards
& Testing
Agency

Key stage 2 science sampling 2014: methodological note and outcomes

Contents

1 Introduction

This document provides information on the outcomes of the key stage 2 (KS2) science sampling assessment in 2014. This was the first year of a new sampling methodology, the details of which are explained in this document. The change in methodology means that comparisons cannot be made with previous years. However, once the next KS2 science sampling assessment is completed in summer 2016, a new time series will be created.

2 Executive Summary

In June 2014, the first live administration of the new format biennial KS2 science sampling assessment took place. The new format follows a matrix sampling design similar to other large scale international sampling assessments such as the Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS) and the National Assessment of Educational Progress (NAEP) in the USA. These types of large scale sampling assessments seek to obtain valid and reliable measures of the achievement of the national population by administering assessments to a sample of pupils.

Since the objective is not to measure the achievement of individual pupils, a large pool of questions can be used, with different groups of pupils taking different combinations of these questions. This is known as matrix sampling, and has the key advantage of allowing test developers to cover a far greater proportion of the programme of study than could normally be covered in a single test instrument. This maximises the validity of the outcomes of the assessment. This approach for KS2 science sampling was recommended by Lord Bew's review[1] of KS2 2 testing, assessment and accountability. The review recognised that the interim arrangements put in place for 2010 to 2012 following the abolition of whole cohort testing in science at KS2 did not take advantage of the potential increase in validity which could be gleaned from a sampling assessment.

Whilst the new approach to science sampling can be considered a more valid measure of science attainment across the curriculum, it represents a large scale change in the design of the assessment, meaning that direct comparisons cannot be made with performance in previous years. The new design involved selecting a sample of approximately 9,500 pupils across 1,900 schools, with pupils taking different combinations of test booklets.

In the previous model used in 2010 to 2012, a sample of 750 schools was selected and all eligible pupils in those schools took the same test. Also in the previous model, schools were notified in February and the test took place in May alongside the other national curriculum tests at the end of KS2; in 2011 and 2012 they received their pupils' results.

In 2014, schools were notified in April and the assessment took place in June; results were not provided to schools. It is likely that the differences in design for 2014 have resulted in some changes in school behaviour and pupil motivation, which are likely to have had an effect on the reported outcomes.

The second administration will take place in 2016. This will follow the same design as the 2014 administration but will assess attainment against the new national curriculum. In

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/176180/Review-KS2-Testing_final-report.pdf

line with other KS2 assessments, reporting arrangements will change with the abolition of the previous national curriculum levels and the introduction of new scale scores and the setting of a new expected standard of attainment. Given the overlap in questions between the 2014 and 2016 assessments, we will be able to report the 2014 and 2016 results on the same basis. Hence, the new measure will be constructed as part of the analysis of 2016 sampling results and will form the basis of the future time series for KS2 science. As an interim solution for 2014, the following results are reported:

- An estimate of the overall performance of the national population in terms of a scale score based on the 2012 raw score scale (0 to 80), plus the percentage of pupils achieving each level.
- Overall performance by gender, English as an additional language (EAL) and eligibility for free school meals (FSM).
- Performance on the 4 attainment targets of the pre-2014 national curriculum (Sc1, Sc2, Sc3 and Sc4) plus the cognitive complexity strands of *knowledge and comprehension*, *application and analysis* and *synthesis and evaluation* from the cognitive domain published in the 2016 test framework (applied retrospectively to the 2012 and 2014 assessments). Sub-score performance is reported by gender.
- Item level information is provided for a selection of released questions in the separate question commentary.

As a result of the methodology change, the data from the 2014 assessment is not comparable historically and is only an interim solution prior to 2016.

Due to the large number of changes between 2012 and 2014, users should be extremely cautious when comparing results as there are a wide range of potential reasons for any observed difference.

In addition, it is difficult to draw implications from KS2 science outcomes in relation to future pupil performance. The science curriculum at KS2 is designed to be broad and balanced. However, evidence shows that when looking at future performance in key stage 4 (KS4) science (for those that go on to do it) performance in KS2 mathematics is a better predictor of performance (see Appendix 5 for further details).

Finally, although TIMSS items were included in the sample, it will not be possible to draw any conclusions on how these results relate to TIMSS performance until we have more data from the 2016 sample and the 2015 TIMSS results become available at the end of 2016.

3 Design

A sample of approximately 9,500 pupils was selected from 1,900 schools, with each pupil taking one of 30 test combinations.

3.1 Assessment matrix

A number of items (test questions) comprising 330 marks were selected to cover the assessable areas of the programme of study. These items were split into 15 booklets of 22 marks each, with 5 booklets comprising questions in the context of each of the 3 core areas of biology, chemistry and physics. As part of the design, each pupil took a combination of 3 booklets (1 biology, 1 chemistry and 1 physics). The 15 booklets were organised into 15 combinations so that every booklet appeared in each of the 3 positions (first, second and third) and each combination included a booklet from each of the 3 core areas.

Additionally, 5 booklets of items from TIMSS were incorporated into the matrix design, as part of a research project, to study the link between performance on TIMSS and KS2 test materials. This created an additional 15 combinations, where each TIMSS booklet appeared in 3 combinations, once in each of the 3 positions. The KS2 test booklets each appeared in 2 of the additional 15 combinations. The complete matrix design is given in Appendix 1.

3.2 Sample selection

A sample of approximately 9,500 pupils was selected from 1,900 schools to take part in the 2014 live science sampling exercise. The selection of schools was stratified by:

- school type, split into: community schools, voluntary aided and voluntary controlled schools, foundation schools, academies and free schools, and special schools
- region, split into: London, South East, South West, North East, North West, Yorkshire and the Humber, East of England, East Midlands, West Midlands
- proportion of pupils eligible for free school meals (FSM), split into quintiles.

First, 1900 schools were selected with probability proportional to size so that each pupil in the population had the same chance of being selected. Within each of the selected schools, 5 pupils were randomly selected to take part. Some schools had fewer than 5 pupils eligible for selection and, in these schools, all pupils were selected. This gave rise to a selected sample of 9,482 pupils. 56 pupils were removed from the sample due to moving schools in the months before the tests took place, reducing the final achieved sample to 9426 pupils. Sample representation tables are given in Appendix 2 at school and pupil level.

4 Methodology

The analysis methodology was designed to replicate processes used to analyse data from international sampling assessments such as TIMSS, PIRLS and PISA. These studies also use matrix designs, where pupils sit different combinations of test blocks to allow a greater coverage of the whole curriculum or content domain than can be achieved within a single test. Analysis of these types of assessments involves a 3 stage process as described below.

4.1 Stage 1: Item Response Theory analysis

Firstly, we use a statistical model to determine the relative difficulty of the items across all of the booklets. The items were calibrated using an Item Response Theory (IRT) model which allows direct comparison between items which did not appear in the same booklets. The two-parameter Generalised Partial Credit (GPC) model was used; this was run using flexMIRT software. In order to report the 2014 outcomes in relation to those from 2012, 5 sets of data were incorporated into the IRT analysis: the 2012 and 2014 live datasets, plus data from 3 pretests which took place in 2009, 2012 and 2013. In total, data from over 40,000 pupils was included in the analysis. The 5 sets of data were matched together to run as a concurrent calibration (as opposed to analysing each dataset separately and then scaling the parameters, known as separate calibration).

IRT analysis relies on a number of assumptions about the data used in the analysis:

- all individual items fit the particular IRT model being used (in this case, the GPCM)
- local independence – scores on individual items are independent of each other once ability is taken into account
- unidimensionality – the items measure a single construct
- items used as 'anchors' to provide a link between different test administrations are sufficiently stable.

Each of these assumptions was tested empirically to ensure the validity of the analysis methodology. The results were presented at the Evidence Review meeting in January 2016 for members of the panel to validate the analysis prior to publication of results. The panel is chaired by the Deputy Director for Assessment Policy and Development and includes assessment and psychometric experts from the Standards and Testing Agency (STA) and the chair of the STA's Technical Advisory Group. This meeting is observed by representatives of various teacher associations and unions.

4.2 Stage 2: Latent regression model

As not all pupils attempted all items in all booklets, the next stage is to estimate their performance based on the items they attempted. In this process, we generate five plausible values for each pupil based on the model. Once all items had been calibrated

onto the same IRT scale, the DESI[2] software produced by the Educational Testing Service (ETS) was used to estimate the latent regression model and generate the 5 plausible values for each pupil. The live 2014 dataset was used, with variables to represent pupils' gender, EAL and FSM statuses included along with item scores. The purpose of including pupil characteristics in the latent regression model is to ensure that resulting sub-group comparisons based on those pupil characteristics are free from bias.

Although the full set of instruments administered in 2014 included 5 booklets of TIMSS material, this was for research purposes only. It is not appropriate for overall national reporting of performance to be based on performance on TIMSS items as this could distort the outcomes, given that TIMSS is based on a different curriculum. Hence, only parameters for the 2014 live material were provided to DESI and the DESI data file only contained data on the 2014 live items. Note that this means that, for the pupils who took a TIMSS booklet, the regression model and plausible values were based on 2 test booklets rather than 3.

In order to create sets of plausible values for each sub-score, the DESI process was re-run for each of the 7 sub-sets of items contributing to each sub-score using the original calibrated item parameters. Again, 5 plausible values were generated for each sub-score.

4.3 Stage 3: Outcomes analysis

The plausible values were generated on the ability scale from the IRT model. To present results in relation to performance in 2012, these plausible ability values were transformed into scores on a 0-80 scale corresponding to raw scores on the 2012 test using IRT true-score equating: essentially this means using the IRT model to predict the distribution of scores the 2014 cohort would have achieved on the 2012 test. Since these scaled scores were equivalent to the raw scores on the 2012 tests, the 2012 level thresholds could simply be applied to those scaled scores to generate an estimated distribution of levels for the 2014 sample.

All statistics (e.g. percentages at each level, average scores) calculated for reporting the outcomes were calculated on each set of plausible values and then averaged. The standard error for each statistic, and therefore the confidence interval, was calculated to take into account both sampling error and measurement error. Measurement error was calculated by taking the variance of the statistic across the 5 plausible values (plausible scaled scores or plausible levels). Sampling error was calculated using bootstrapping: 500 re-samples were taken from the original sample, with replacement (to achieve 500 samples of the same size as the original sample), and the statistics of interest were calculated based on each re-sample. The variance of each statistic across the bootstrap samples provides an indication of sampling error. The estimates of sampling variance

[2] Direct Estimation Software Interactive 4.0.0 (2013) Educational Testing Service, New Jersey, USA. Used with kind permission from ETS.

and measurement variance were combined together to produce an overall estimate of the variance using the following formula (Foy et al, 2008):

$$Var(\hat{T}) = \bar{U} + (1 + M^{-1})\, B_M$$

Where:

- $\hat{T}$is the estimate of the statistic of interest (eg the mean scaled score)
- $\bar{U}$ is the average sampling variance across the 5 plausible values (those derived from bootstrapping)
- *M* is the number of plausible values (5)
- B_M is the variance of the estimate of T across the plausible values (the measurement error).

The overall standard error, the square root of Var(T^), was then used to generate confidence intervals to be reported around the statistics.

4.4 Item level analysis

A question commentary will be published shortly containing a selection of released questions and their mark schemes with some commentary on how pupils tended to respond. The report will show the estimated percentage of pupils in the national cohort who got the question part (item) correct (or scored 1/2/3 marks) or who gave each type of response. Since different groups of pupils took different combinations of the test booklets, this could not be calculated as a straightforward percentage of the pupils who were given that item. Instead, the overall sample was split into 20 quantiles (‘ventiles’) based on ability. For each item, the percentage of pupils achieving each mark point was calculated within the 20 quantiles. An average was then taken across the 20 quantiles; since there would not be an equal number of pupils in each quantile for a particular item, this is effectively a weighted average which gives an estimate of how the whole sample would be expected to perform on that item.

5 Outcomes for 2014

Attainment in the 2014 science sampling exercise is summarised in table 1 for all pupils and split by sub-groups. 63% of pupils are estimated to have achieved level 4 and above and 11% to have achieved level 5. Note that results are referred to as estimates. This is because in the matrix sample design each pupil was given a subset of questions. It is not appropriate to assign levels to individual pupils and aggregate them to calculate a standard percentage. Instead, statistical modelling is used to estimate the performance of the sample as a whole, as described in section 4.

Table 1: Estimated percentage of pupils achieving level 4 and above and level 5 based on KS2 science sampling tests in 2014

	Estimated percentage achieving level 4 or above	Level 4 or above 95% confidence interval	Estimated percentage achieving level 5	Level 5 95% confidence interval
All pupils	63.5	±1.3	10.7	±1.1
Boys	62.4	±1.7	10.8	±1.9
Girls	64.6	±1.9	10.5	±2.2
FSM	43.2	±3.3	3.5	±3.1
Non-FSM	67.9	±1.5	12.2	±1.6
EAL	54.3	±3.7	8.0	±3.5
Non-EAL	65.2	±1.5	11.2	±1.6

Performance between boys and girls was very similar, with no significant difference in the percentage of boys and girls achieving level 4 and above and level 5. Pupils eligible for FSM performed significantly lower than other pupils, with fewer than 4% achieving level 5. Pupils with EAL also performed significantly lower than other pupils.

Performance was broken down into sub-scores representing each of the 4 attainment targets and these results are shown in table 2. Note that attainment target scores cannot be compared *across* the attainment targets. This is because it's impossible to separate how well pupils performed on a particular content area from how difficult that content area is. Only scores *within* a given attainment target can be meaningfully compared.

Given that the new scaled scores scale cannot be developed until after the 2016 sampling test of the new national curriculum has been administered, we have scaled the outcomes from 2014 to report against the 0-80 raw score scale that was reported in 2012. The results for 2014 show that boys performed significantly better than girls on Sc4

(physical processes). Girls performed slightly better than boys on the other strands but none of these differences were statistically significant.

Table 2: Average scores achieving on the KS2 science sampling tests by attainment target

	Sc1: Scientific enquiry (Max 31 marks)	Sc2: Life processes and living things (Max 16 marks)	Sc3: Materials and their properties (Max 17 marks)	Sc4: Physical processes (Max 16 marks)	Overall test score (Max 80 marks)
All pupils	18.7 (±0.2)	9.9 (±0.1)	9.3 (±0.1)	8.7 (±0.1)	46.6 (±0.4)
Boys	18.6 (±0.2)	9.8 (±0.1)	9.3 (±0.2)	8.9 (±0.1)	41.1 (±0.1)
Girls	18.9 (±0.2)	10 (±0.1)	9.3 (±0.2)	8.4 (±0.2)	42.5 (±0.1)

Performance was also broken down into sub-scores representing each of the 3 levels of the cognitive complexity rating strand of the cognitive domain. More information can be found in the KS2 science sampling test framework at www.gov.uk/government/publications/key-stage-2-science-sampling-test-framework.

Note that the cognitive domain strands were introduced for the 2016 test framework but have been applied retrospectively to the 2014 assessmment. Again, scores are presented on the 2012 raw score scale. The results for 2014 show that boys performed slightly better on the items assessing knowledge and comprehension and girls performed slightly better on application and analysis and synthesis and evaluation but none of these differences were statistically significant.

Table 3: Average scores achieving on the KS2 science sampling tests by cognitive complexity rating

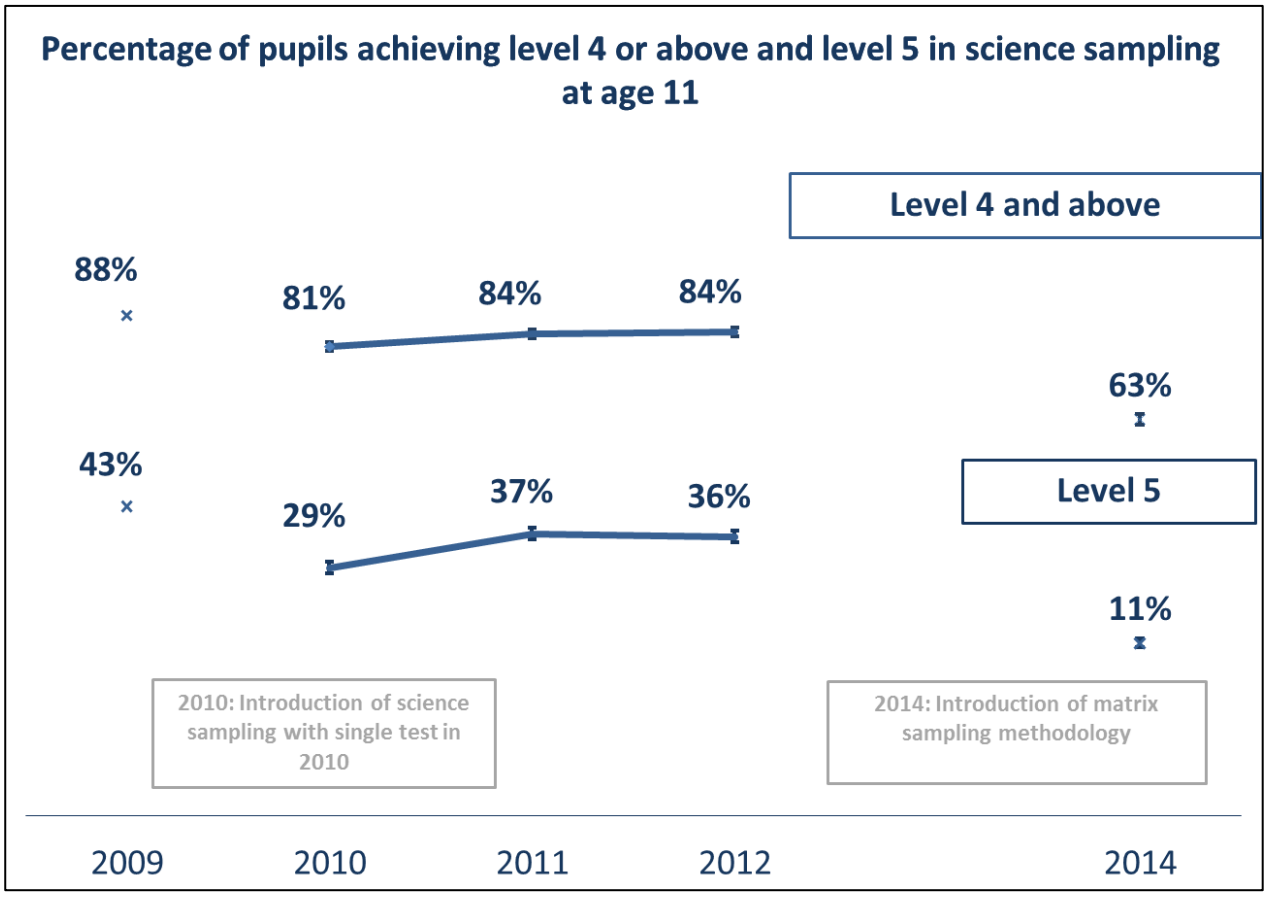
	Knowledge and comprehension (Max 24 marks)	Application and analysis (Max 44 marks)	Synthesis and evaluation (Max 12 marks)	Overall test score (Max 80 marks)
All pupils	14.8 (±0.1)	24.7 (±0.2)	7.2 (±0.1)	46.6 (±0.4)
Boys	14.9 (±0.2)	24.6 (±0.3)	7.1 (±0.1)	41.1 (±0.1)
Girls	14.7 (±0.2)	24.8 (±0.3)	7.2 (±0.1)	42.5 (±0.1)

6 Historical performance

The considerable change in the way KS2 science achievement is now being measured means that valid comparisons cannot be made with performance in previous years. Once the 2014 results are published on a consistent basis with 2016 methods, 2014 will be considered a new baseline for comparison, although the actual measure for comparison will not be set until 2016 given the additional changes made for assessing the new national curriculum (i.e. the removal of levels and the introduction of scaled scores).

Figure 1 below shows the time series.

Figure 1: Historical achievement in KS2 science



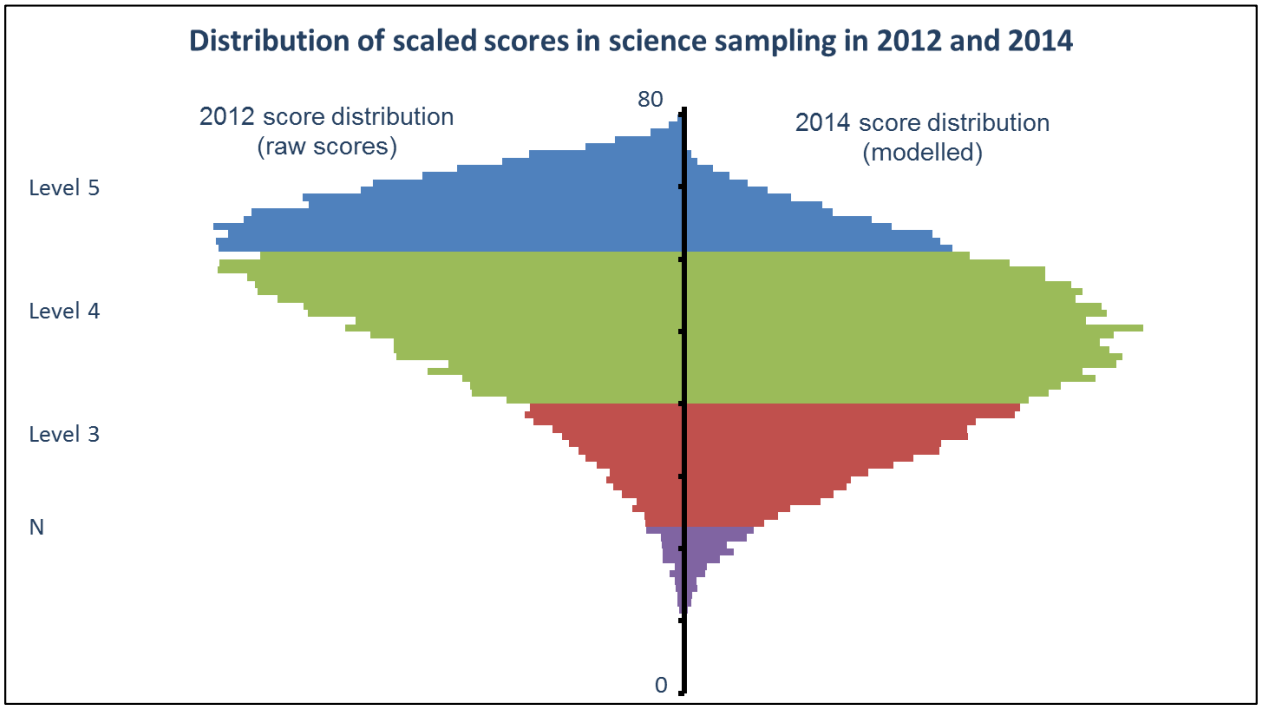
There are a number of ways in which the change in design of the science sampling assessment may account for the drop in performance. In the previous model, schools were told in February that they would be part of the sample, the whole class was involved and the teacher received the results. This allowed schools to undertake preparation activities with the class before the test if they wanted to, which is likely to have improved performance. The test also took place during KS2 test week so pupils were 'test ready'. In the new model, schools are told in April that they will be in the sample but they are not told which 5 pupils will take part in the sample until 5 working days before the test and no pupil level results are provided to the school. This gives less opportunity and less incentive to prepare the whole class when only 5 pupils will be selected. The test also

takes place in June at a different time from the other KS2 tests. These changes are likely to have had some impact on school behaviour and pupil motivation.

Some evidence for the potential impact this change has had on school behaviour and pupil motivation comes from the rise in the proportion of pupils not sitting the test. The proportion of pupils not sitting the test due to being absent, working below the level of the test or working at the level of the test but unable to access it has increased between 2012 and 2014. In 2012 this was just under 4% of the population. In 2014 it was just over 10%. Since these pupils are considered part of the sample and are included in the denominator when percentages are calculated, this would automatically have the effect of reducing the reported performance. Of the pupils who actually sat the test, 71% were estimated to have achieved level 4 or above and 12% to have achieved level 5. This compares with 87% and 37% respectively in 2012.

The distribution of scores for those who sat the test is shown in figure 2. The 2012 score distribution was skewed, with the largest proportions of pupils around the level 5 threshold. The 2014 scores, modelled to be equivalent to the raw scores on the 2012 test for the purpose of comparison, follow something much closer to a normal distribution, with the majority of pupils falling around the middle of the level 4 band. The reasons for this change cannot be attributed due to the large number of possible reasons.

Figure 2: Distribution of scaled scores: 2012 and 2014



7 Quality assurance and sign-off

A series of papers to agree details of the matrix design, sample selection, analysis procedures and reporting were presented to the STA Technical Sub-programme Board meeting between December 2012 and March 2013.

The complex nature of these types of matrix sampling assessments means that traditional methods of analysis, setting of level thresholds and reporting are no longer appropriate, and techniques new to STA needed to be employed, as has been presented in this paper. The analysis methodology was reviewed by the Technical Advisory Group in February 2014. All analysis was quality checked by a second psychometrician. Details of the analysis, including the results of the IRT assumption checking, were presented at the Evidence Review meeting held in January 2016 and chaired by the Deputy Director for Assessment Policy and Development. Participants at the meeting included the Chair of the Technical Advisory Group. The meeting agreed the analysis was technically sound and signed the outcomes off for publication.

Appendix 1: Test Booklet Combinations Used in 2014

Combination	1st booklet	2nd booklet	3rd booklet
1	ST001B	ST010C	ST011P
2	ST002B	ST008C	ST012P
3	ST003B	ST007C	ST013P
4	ST004B	ST006C	ST015P
5	ST005B	ST009C	ST014P
6	ST010C	ST013P	ST002B
7	ST008C	ST015P	ST003B
8	ST007C	ST014P	ST004B
9	ST006C	ST011P	ST005B
10	ST009C	ST012P	ST001B
11	ST011P	ST003B	ST009C
12	ST012P	ST004B	ST010C
13	ST013P	ST005B	ST008C
14	ST015P	ST001B	ST007C
15	ST014P	ST002B	ST006C
16	ST00T1	ST009C	ST013P
17	ST001B	ST00T1	ST014P
18	ST004B	ST008C	ST00T1
19	ST00T2	ST006C	ST012P
20	ST002B	ST00T2	ST015P
21	ST007C	ST011P	ST00T2
22	ST006C	ST013P	ST00T3
23	ST010C	ST00T3	ST005B
24	ST00T3	ST014P	ST003B
25	ST008C	ST00T4	ST001B
26	ST00T4	ST003B	ST010C
27	ST012P	ST005B	ST00T4
28	ST015P	ST00T5	ST009C
29	ST011P	ST004B	ST00T5
30	ST00T5	ST002B	ST007C

The 15 KS2 test booklets are denoted ST001 to ST015, with a B, C or P suffix to indicate the core content area assessed. The TIMSS booklets are denoted ST00T1 to ST00T5.

Appendix 2: Sample Representation Tables

Table A2:1 shows the representation of the sample in terms of the three school-level stratifiers of school type, region and FSM band. It confirms that the sample was representative of these school level characteristics.

Table A2.1: School level sample representation

		In sample frame		In sample	
		Frequency	%	Frequency	%
School Type	Community Schools	7267	46.9	891	46.9
	Voluntary aided and voluntary controlled schools	5257	33.9	642	33.8
	Foundation schools	581	3.7	72	3.8
	Academies and free schools	1732	11.2	212	11.2
	Special schools	671	4.3	83	4.4
Region	East Midlands	1493	9.6	182	9.6
	East of England	1720	11.1	211	11.1
	London	1705	11.0	208	10.9
	North East	792	5.1	97	5.1
	North West	2457	15.8	301	15.8
	South East	2219	14.3	272	14.3
	South West	1754	11.3	217	11.4
	West Midlands	1646	10.6	201	10.6
	Yorkshire and the Humber	1722	11.1	211	11.1
FSM band	Lowest	3112	20.1	382	20.1
	Second Lowest	3093	19.9	376	19.8
	Middle	3097	20.0	378	19.9
	Second Highest	3110	20.1	382	20.1
	Highest	3096	20.0	382	20.1
Total		15,508		1900	

The final achieved sample consisted of 9426 pupils. Some of those pupils were absent (code A), working below the level of the test (code B) or at the level of the test but unable to access it (code T) so did not actually take the test. They are still considered part of the sample for the purposes of reporting outcomes such as the proportion of pupils at a given level. The proportion of pupils in each of these categories was considerably greater than in 2012 and left a total of 8449 test takers.

Table A2.2: Pupil level sample representation

		In population (%)	In sample (%)	In test takers (%)
Gender	Female	48.8	48.5	49.3
	Male	51.2	51.5	50.7
FSM	No FSM provision	81.9	81.4	83.0
	FSM provision	17.1	17.6	16.2
EAL	English as first language	80.9	83.0	83.5
	English as additional language	18.2	16.0	15.7
	Total			

*Note that 1% of pupils in the sample were missing FSM and EAL information

Table A2:2 shows the representation of the sample at pupil level in terms of the 3 reporting characteristics of gender, EAL and FSM. For gender, the sample and the sub-sample who took the test were both representative of the overall population[3]. For FSM, the sample was representative but the proportion within the group of test takers was significantly lower than in the population. The proportion of pupils with EAL was significantly lower both in the sample and group of test takers than in the population. The sample was not designed to be stratified at pupil level, but this difference should be born in mind when interpreting results.

[3] Here the 'overall population' is defined as the key stage 2 national curriculum pupil registration data used to determine the sample frame for selection.

Appendix 3: Historical Results Tables

Table A3.1: Estimated percentage of pupils achieving level 4 and above and level 5 based on KS2 science sampling tests over time

	Year	Estimated percentage achieving level 4 or above	Estimated percentage achieving level 5
All pupils	2009	88	43
	2010	81	28
	2011	84	36
	2012	84	36
	2014	63	11
Boys	2009	88	43
	2010	80	29
	2011	83	35
	2012	84	36
	2014	62	11
Girls	2009	89	43
	2010	81	28
	2011	85	38
	2012	85	35
	2014	65	11
FSM	2014	43	3
Non-FSM	2014	68	12
EAL	2014	54	8
Non-EAL	2014	65	11

Appendix 4: Sub-score Results 2012 to 2014

Table A4.1: Average scores achieving on the KS2 science sampling tests by attainment target: 2012 to 2014

	Year	Sc1: Scientific enquiry (Max 31 marks)	Sc2: Life processes and living things (Max 16 marks)	Sc3: Materials and their properties (Max 17 marks)	Sc4: Physical processes (Max 16 marks)	Overall test score (Max 80 marks)
All pupils	2012	21.8	11.3	11.1	10.4	54.6
	2014	18.7 (±0.2)	9.9 (±0.1)	9.3 (±0.1)	8.7 (±0.1)	46.6 (±0.4)
Boys	2014	18.6 (±0.2)	9.8 (±0.1)	9.3 (±0.2)	8.9 (±0.1)	41.1 (±0.1)
Girls	2014	18.9 (±0.2)	10 (±0.1)	9.3 (±0.2)	8.4 (±0.2)	42.5 (±0.1)

Table A4.2: Average scores achieving on the KS2 science sampling tests by cognitive complexity rating: 2012 to 2014

	Year	Knowledge and comprehension (Max 24 marks)	Application and analysis (Max 44 marks)	Synthesis and evaluation (Max 12 marks)	Overall test score (Max 80 marks)
All pupils	2012	17.2	29.1	8.3	54.6
	2014	14.8 (±0.1)	24.7 (±0.2)	7.2 (±0.1)	46.6 (±0.4)
Boys	2014	14.9 (±0.2)	24.6 (±0.3)	7.1 (±0.1)	41.1 (±0.1)
Girls	2014	14.7 (±0.2)	24.8 (±0.3)	7.2 (±0.1)	42.5 (±0.1)

Appendix 5: The relationship between attainment at KS2 and attainment at KS4

This appendix provides information on the relationship between KS2 results and KS4 EBacc science attainment. More specifically, it looks at which of the KS2 results is a better predictor of KS4 EBacc science.

Table A5.1 below sets out figures on the correlation between KS2 science, maths and English fine grades, and KS4 EBacc science from 2010/11 to 2014/15. The results show that KS2 maths is a slightly better predictor of KS4 EBacc science than KS2 science and this is consistent over time.

Table A5.1: Correlation between KS2 English, maths and science fine grade and EBacc Science points for EBacc entrants only, 2011 to 2015

KS2 year	2005/06	2006/07	2007/08	2008/09	2009/10
KS4 EBacc Science year	2010/11	2011/12	2012/13	2013/14	2014/15
KS2 Maths	0.62	0.61	0.60	0.60	0.60
KS2 Science	0.61	0.59	0.59	0.59	N/A*
KS2 English	0.56	0.56	0.53	0.54	0.56

*Note that the correlation between KS2 science fine grade and KS4 EBacc science score is not available for the 2014/15 KS4 cohort because KS2 science tests were discontinued after 2008/09

Standards & Testing Agency

2016 Key stage 2 science sampling 2014: methodological note and outcomes
Electronic version product code: STA/16/7528/e ISBN: 978-1-78644-118-8

Any enquiries regarding this publication should be sent to us at science.ks2@education.gsi.gov.uk.

Central newsdesk for media enquiries: 020 7783 8300

This document is also available for download at: https://www.gov.uk/government/organisations/standards-and-testing-agency.