Standards & Testing Agency

Key stage 2 science sampling 2016

Methodology note and outcomes

July 2017

Contents

1 Introduction

This document provides information on the outcomes of the key stage 2 (KS2) science sampling assessment in 2016. It also links the 2016 outcomes to those of the 2014 science sampling assessment. It necessarily contains technical information on the matrix sampling method and analysis in addition to providing information on the outcomes of the analysis.

1.1 Background

The purpose of the KS2 science sampling assessment is to monitor national performance in science. It is not possible or appropriate to provide information on individual or school performance. The biennial KS2 science sampling assessment first took place in 2014. The new format follows a matrix sampling approach similar to large-scale international assessments such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). Matrix sampling assessments seek to obtain valid and reliable measures of the achievement of the national population by administering assessments to just a sample of pupils. Since measurement of individual pupils is not the aim of this approach, a large pool of questions can be used, with different groups of pupils taking different combinations of questions. This has the advantage of allowing test developers to cover a far greater proportion of the programme of study than would normally be covered in a single test instrument, while minimising the assessment burden on individual pupils.

The 2016 science sampling assessment at KS2 was the first taken by pupils who had studied the revised national curriculum, introduced in the 2014 to 2015 academic year. The curriculum covers the 4 years of KS2 and is designed to combine the best elements of the world's most successful school systems and have a greater focus on scientific knowledge. Pupils who took part in the science sampling assessment in 2016 only experienced 2 years of the revised curriculum.

After the administration of the KS2 science sampling assessment, as with the other KS2 tests in mathematics, English reading and English grammar, punctuation and spelling, a new expected standard equivalent to the scaled score of 100 was set by teacher panels. The range of scaled scores available for the science sampling assessment was 70 to 120, and is intended to stay the same in future years.

The sample design involved selecting a sample of approximately 9,500 pupils across 1,900 schools, with pupils taking different combinations of test booklets. Schools were selected such that each eligible pupil in the population had an equal chance of being selected. The selection was stratified by the proportion of pupils eligible for free school meals (FSM), school type and region to ensure representativeness of the population. In 2014 and 2016, schools were notified of their required participation in April and the assessment took place in June. Results were not provided to schools.

As there was a large overlap of questions between the 2014 and 2016 administrations, outcomes can be reported on the same scale. However, any differences in performance between 2014 and 2016 must be considered in the context of the changing primary curriculum. Pupils taking part in the 2014 matrix sample were not exposed to the newly-implemented national curriculum. Pupils taking part in the 2016 matrix sample had only accessed 2 years of the curriculum. Caution should be exercised in interpreting the findings.

2 Executive summary

In June 2014, the first live administration of the new-format biennial KS2 science sampling assessment took place. The second administration took place in June 2016 following the same design as the 2014 administration but assessing attainment against the revised national curriculum.

The assessment design 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. Lord Bew's review[1] of KS2 testing, assessment and accountability recommended this approach for KS2 science sampling. 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 matrix sampling approach.

Whilst the 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 design involved selecting a sample of approximately 9,500 pupils across 1,900 schools, with pupils taking different combinations of test booklets.

In line with other KS2 assessments, reporting arrangements changed with the abolition of the previous national curriculum levels and the introduction of new scaled scores and the setting of a new expected standard of attainment. The analysis herein includes items and pupils from both 2014 and 2016. There was a large overlap of items to allow the estimation of the 2014 cohort on the 2016 scale.

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

The 2014 and 2016 results are detailed below:

- An estimate of the overall performance of the national population in terms of a scale score based on the 2016 scaled score range (70-120), plus the percentage of pupils achieving the expected standard.
- Overall performance by gender, English as an additional language (EAL) and eligibility for free school meals (FSM).
- Performance on the 4 content sub-strands of the national curriculum (biology, chemistry, physics and working scientifically). Sub-score performance is reported by gender.

23% of pupils are estimated to have achieved the expected standard in 2016 and 28% of pupils were estimated to have achieved the expected standard in 2014.

Figure 1: Percentage of pupils achieving the expected standard in science sampling at age 11

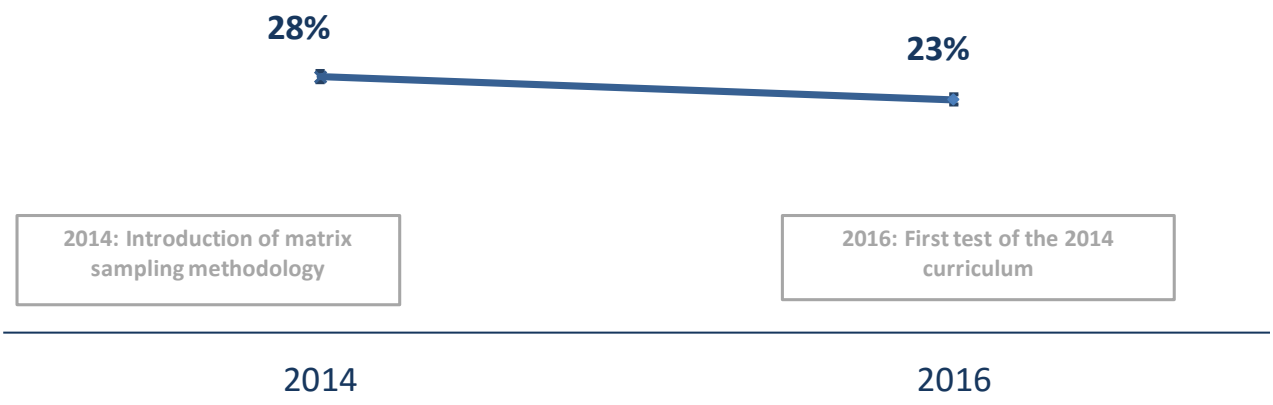


Figure 1: Percentage of pupils achieving the expected standard in science sampling at age 11

3. Design

This section details the matrix design and sample selection of the KS2 science sampling assessment.

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. Nine of these booklets were the same in 2014 and 2016 administrations and were used to link performance across those administrations.

Additionally, 5 booklets of items from TIMSS were incorporated into the matrix design in 2014 and 2016, 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 provided in Appendix 1.

3.2 Sample selection

The sample selection process was the same in 2014 and 2016. A sample of approximately 9,500 pupils was selected from 1,900 schools to take part in the 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

1,900 schools were initially 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. In 2016, this gave rise to a selected sample of 9,480 pupils. 71 pupils were removed from the sample due to moving schools in the months before the tests took place, reducing the final achieved sample to 9,409 pupils. In 2014, the selected sample included 9,482 pupils. Of these, 56 pupils were removed from the sample due to moving schools prior to

the administration, reducing the final achieved sample to 9,426 pupils. Sample representation tables are provided 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 curriculum or content domain than can be achieved within a single test. Analysis of these types of assessments involves a 3-stage process, detailed below.

4.1 Stage 1: Item Response Theory (IRT) analysis

A statistical model was used to determine the relative difficulty of the items across all of the booklets. The items were calibrated in flexMIRT software using the Graded Response Model (GRM). Data from both 2014 and 2016 were incorporated into a single analysis (concurrent calibration). The common items, from 9 booklets, which were the same across the 2 years, served as a link so that all items could be directly compared, despite not appearing in the same booklets or even within the same year. Note that, although the same TIMSS items were administered in the 2014 and 2016, these were not used to link performance between years. The reason for this is that it would be inappropriate for a change in performance on TIMSS items to be reflected as a change in performance on KS2 science, given that the TIMSS content domain is slightly different.

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 GRM)
- local independence – scores on individual items are independent of each other once ability is taken into account
- the construct being measured is unidimensional – 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.

4.2 Stage 2: Latent regression model

As each pupil took a subset of the overall pool of items, the next stage of analysis involves estimating their performance based on the items they were given. The same datasets were used as for the first stage, with variables included to represent pupils’ gender, EAL and FSM status. These variables are included in the model to ensure that resulting sub-group comparisons based on those pupil characteristics are free from bias.

Once the latent regression model was run, plausible values were generated. These are random draws from the probability distribution (referred to as the posterior distribution) of a pupil's ability. They are used as a way of reflecting the measurement error inherent in the process. For this analysis, 5 plausible values were generated for each pupil. The latent regression and generation of plausible values were run in flexMIRT.

4.3 Stage 3: Outcomes analysis

Following the previous stages, all pupils (who took the test) from both 2014 and 2016 had 5 plausible values. These plausible values are generated on a relatively arbitrary but standard IRT ability scale, which centres around 0 and ranges from around -3 to 3. In order to translate these plausible values into meaningful outcomes the expected standard for KS2 science needed to be applied to them.

A standard setting exercise was conducted in September 2016 using the Bookmark approach, the same used for the other 2016 national curriculum tests. This process involves convening teacher panels to recommend where the standard should be set, using an ordered item booklet based on live administration data. The outcome of the process is to agree a raw score on the ordered item booklet that represents the threshold for the standard. Usually the ordered item booklet would be based on the single test instrument but, since the KS2 science design does not fully allow for this, the ordered item booklet was a 'virtual' test, based on a selection of items from the 2016 item pool totalling 100 marks. The ordered item booklet was made up of 1 actual booklet combination of biology, chemistry and physics with additional questions to improve the curriculum representation in the booklet.

The outcome of the standard setting approach was that a raw score of 62 marks on the ordered item booklet would represent the threshold for the expected standard. Using the IRT parameters derived from the analysis described in sections 4.1 and 4.2, it was possible to estimate the ability parameter of a pupil who would have an expected score of 62 marks on the selection of items comprising the ordered item booklet. This ability value was then used to represent the expected standard. In line with the whole cohort national curriculum tests, a score scale was required such that 100 would represent the expected standard. The range of scaled scores available was 70 to 120.

For each pupil, each of their 5 plausible values was converted to a 'plausible scaled score' and then to a 'plausible outcome' (such as has met the expected standard or has not). All statistics (for example the percentage at the expected standard, average scaled score) calculated for reporting the outcomes were calculated on each set of plausible values and then averaged. Measurement error was calculated by taking the variance of the statistic across the 5 plausible values.

In addition to measurement error, sampling error was estimated in order to account for the fact that only a sample of pupils took the assessment. This 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 the sampling error.

The estimates of sampling variance and measurement variance were combined together to produce an overall estimate of the variance using the following formula[2]:

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

Where:

- $\hat{T}$ is the estimate of the statistic of interest (for example 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 $Var(\hat{T})$, was then used to generate confidence intervals to be reported around the statistics.

4.4 Sub-strand analysis

Sub-scores were estimated for curriculum content areas, biology, chemistry, physics and working scientifically. A bifactor model in flexMIRT was used in this analysis. This is a type of multidimensional IRT model in which there is one general factor, on which all items load (overall KS2 science performance), and several specific factors, on which subsets of the items can load (for example the content areas). Items can load on, at most, 2 factors (for example the general factor and one sub-factor).

The model was run in flexMIRT in much the same way as the main analysis described in sections 4.1 and 4.2, with the addition of being set up as a bifactor model. This resulted in each pupil being assigned 5 plausible values for the general factor and each of the sub-factors. The plausible values for the general factor in the bifactor analysis were ignored, as the plausible values from the main analysis were used to determine the overall measure of science performance. As with the main analysis, a bootstrapping procedure was run to estimate the sample variance and the measurement variance was derived from the variance of the plausible values. These were then combined together to form confidence intervals for the average sub-strands. The sub-strands are reported on

[2] Foy, P., Galia, J. and Li, I. (2008). 'Scaling the Data from the TIMSS 2007 Mathematics and Science Assessments', TIMSS 2007 Technical Report, 11, 225-279 [online]. Available: https://timssandpirls.bc.edu/TIMSS2007/PDF/T07_TR_Chapter11.pdf [26 April, 2017].

the IRT ability scale, which centres around zero. Values that are below zero indicate lower attainment in the sub-strands, values that are over zero indicate higher attainment in the sub-strands. As there is no expected standard at sub-strand level it is not possible to create a score scale for the sub-strands that works in the same way as that for the overall score.

5 Outcomes for 2014 and 2016

Attainment in the 2014 and 2016 science sampling exercise is summarised in table 1 for all pupils and split by sub-groups. Overall attainment in the 2016 science sampling exercise is estimated to be lower than in the previous sampling exercise in 2014. The percentage of pupils estimated to have achieved the expected standard in 2016 is down by 5 percentage points from 2014. 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 outcomes to individual pupils and aggregate them to calculate a standard percentage. Instead, statistical modelling is used to estimate the performance of the population as a whole, as described in section 4.

	Estimated percentage achieving expected standard in 2014	95% confidence interval in 2014	Estimated percentage achieving expected standard in 2016	95% confidence interval in 2016
All Pupils	28.06	26.82 – 29.31	22.77	21.61 – 23.92
Boys	27.86	26.03 – 29.68	22.73	21.02 – 24.45
Girls	28.28	26.26 – 30.31	22.80	20.80 – 24.80
FSM	13.14	10.75 – 15.53	9.00	6.80 – 11.19
Non-FSM	31.35	29.93 – 32.76	25.28	23.99 – 26.58
EAL	23.17	20.25 – 26.09	17.41	14.64 – 20.17
Non-EAL	29.06	27.63 – 30.49	23.77	22.49 – 25.04

Table 1: Estimated percentage of pupils achieving the expected standard based on KS2 science sampling assessments in 2014 and 2016

The scaled score distribution for 2014 and 2016 is shown in figure 2. The scaled score range is 70 to 120. The shapes of the distributions are very similar but it also shows greater numbers of pupils achieving the high scaled scores in the 2014 sample and greater numbers of pupils achieving the lowest scaled scores in the 2016 sample.

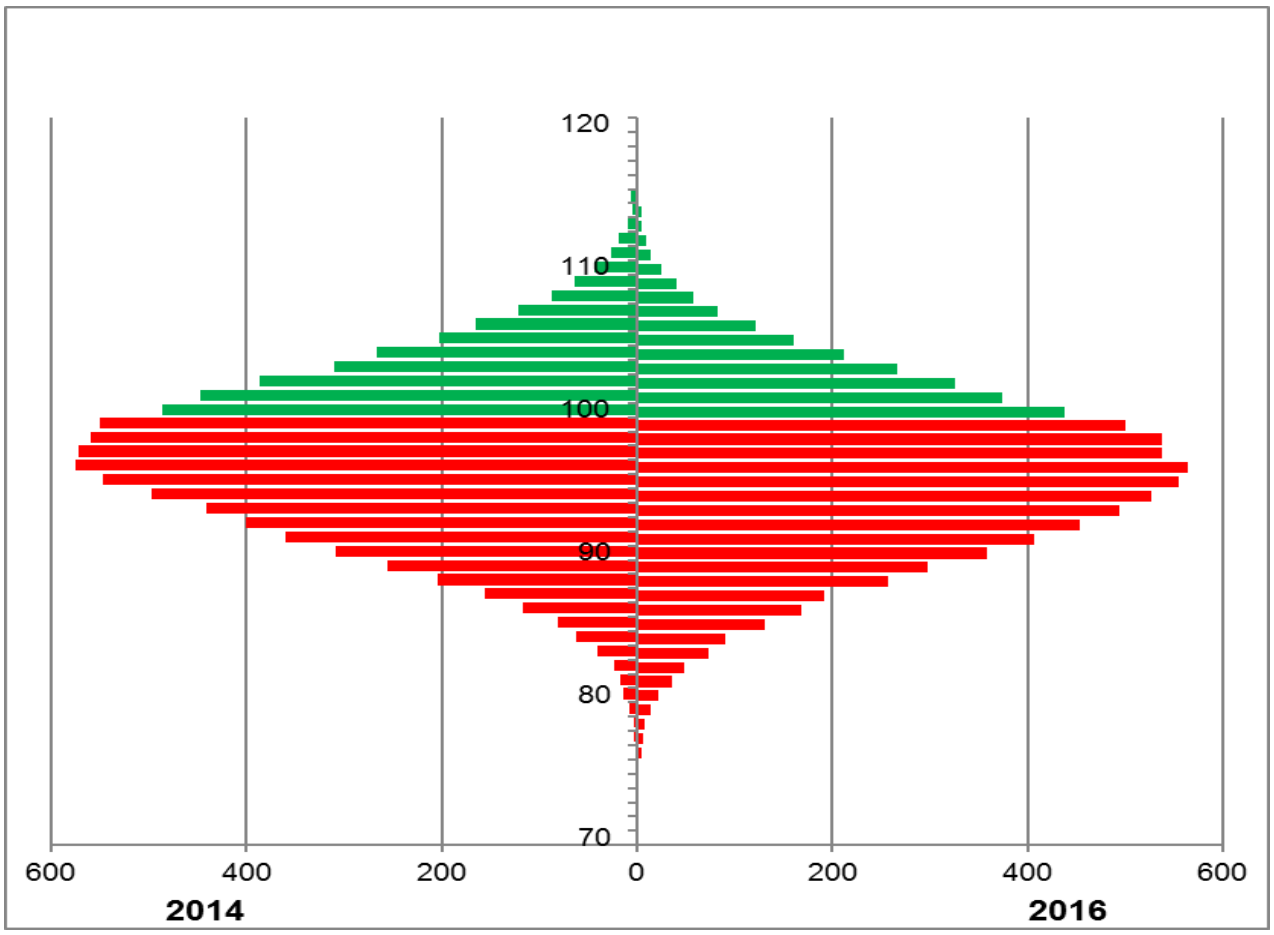
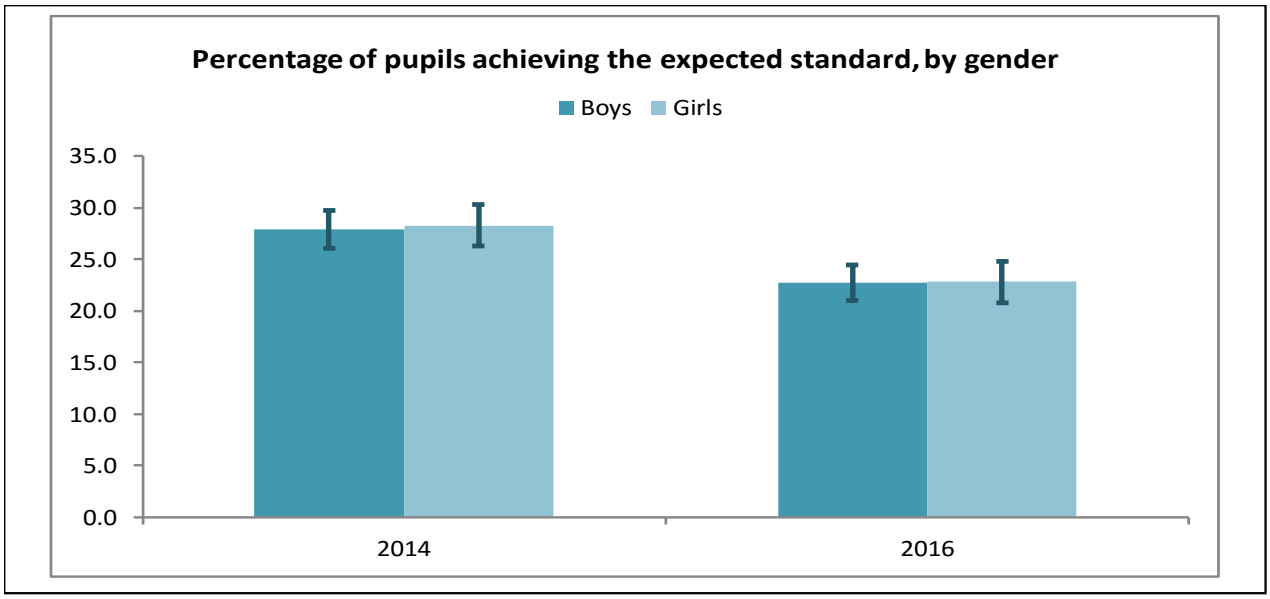


Figure 2: Distribution of scaled scores

5.2 Gender

As in previous years, performance between girls and boys was very similar. There was no significant difference in the percentages of girls and boys achieving the expected standard in either 2014 or 2016. In 2014, just over 27% of boys were estimated to have reached the expected standard, compared to just over 28% of girls. In 2016, just over 22% of boys and girls were estimated to have reached the expected standard.

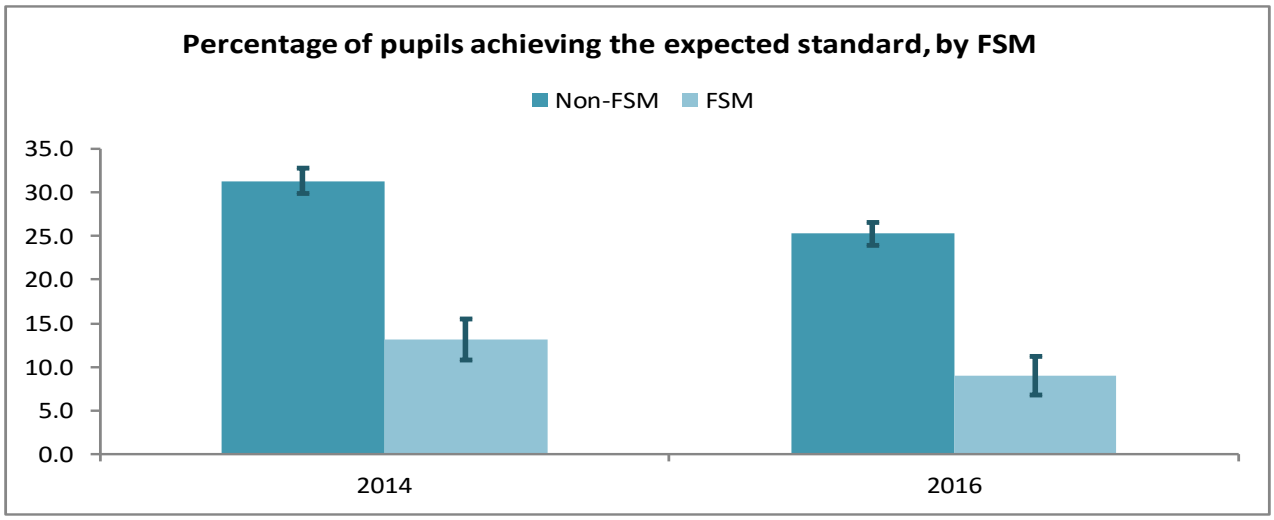


Note: Bars around data points indicate approximate 95% confidence intervals.

Figure 3: Attainment in science by gender

5.2 Free school meals (FSM)

The performance of pupils eligible for FSM was significantly lower than other pupils. In 2016, 9% of FSM eligible pupils were estimated to have reached the expected standard compared to just over 25% of non-FSM pupils. In 2014, just over 31% of non-FSM pupils were estimated to have reached the expected standard compared to just over 13% of FSM pupils.

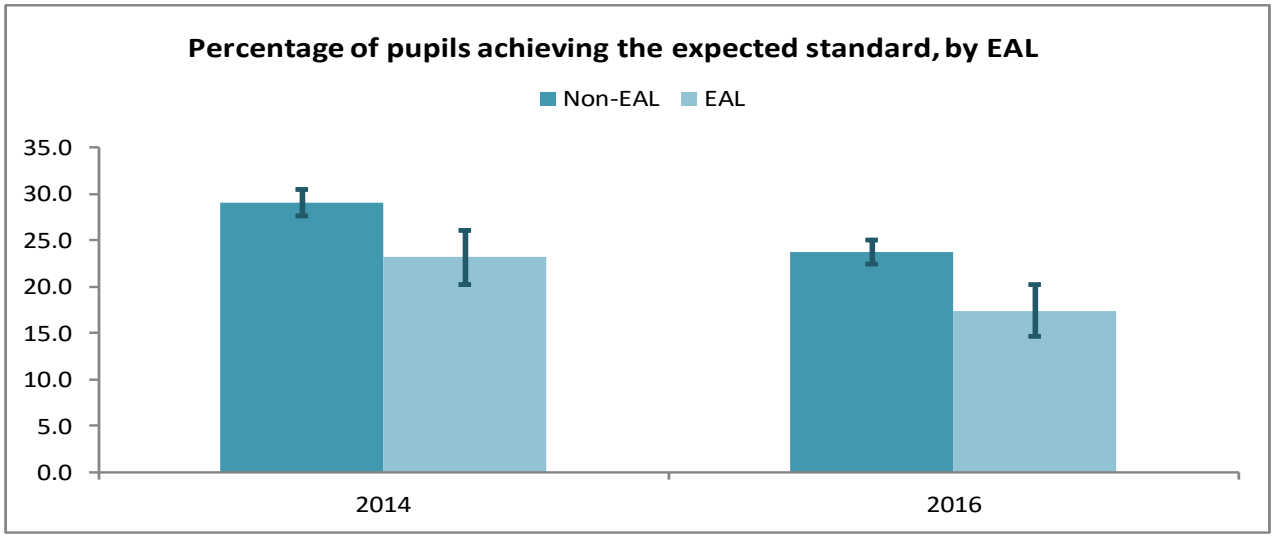


Note: Bars around data points indicate approximate 95% confidence intervals.

Figure 4: Attainment in science by FSM eligibility

5.3 English as an additional language (EAL)

The performance of pupils with EAL was significantly lower than other pupils. In 2016, 17% of pupils with EAL compared to 23% of non-EAL pupils were estimated to have reached the expected standard. In 2014, 23% of pupils with EAL were estimated to have reached the expected standard, compared to 29% of non-EAL pupils.



Note: Bars around data points indicate approximate 95% confidence intervals.

Figure 5: Attainment in science by EAL

6 Performance of sub-strands

The model used to compute performance on the content sub-strands produces a scale centred around zero. The values in the tables below are averages on the scale for each of the content domains for all pupils and broken down by gender. Values below zero indicate lower attainment in the strands, values over zero indicate higher attainment in the strands.

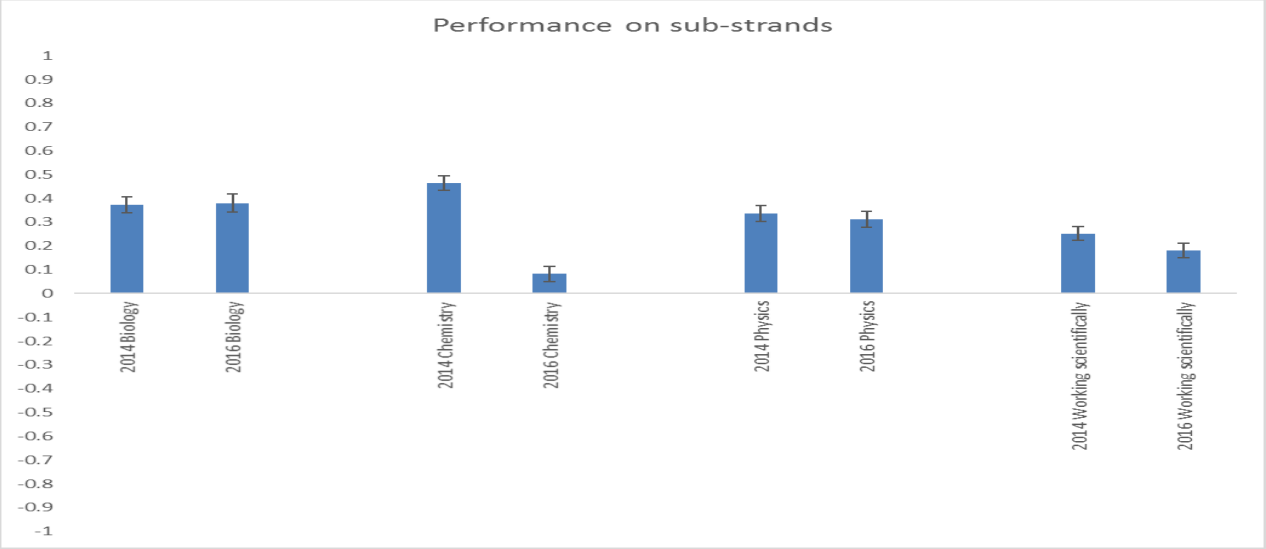
Performance on biology and physics is relatively similar from 2014 to 2016. There is a slight fall in working scientifically and there is a marked drop in chemistry. This indicates that pupils' estimated ability in chemistry has fallen between 2014 and 2016. This is illustrated in table 2 and figure 6.

While the new curriculum provides guidance on when each curriculum strand should be taught, the old curriculum did not. From 2014 onwards, chemistry content is being taught in years 3, 4 and 5, with no year 6 content. Because of this, pupils sitting the 2016 science sampling assessment may have not have received chemistry teaching in year 6, nor might they have experienced the new chemistry curriculum in full.

	Biology	Chemistry	Physics	Working scientifically
2016	0.38 (0.34, 0.42)	0.08 (0.05, 0.12)	0.31 (0.28, 0.35)	0.18 (0.15, 0.21)
No. marks assessing strand	97	81	81	119
2014	0.37 (0.34, 0.41)	0.46 (0.43, 0.49)	0.34 (0.30, 0.37)	0.25 (0.22, 0.28)
No. marks assessing strand	82	81	76	136

Note: ranges given in brackets indicate approximate 95% confidence intervals.

Table 2: Performance on sub-strands



Note: Bars around data points indicate approximate 95% confidence intervals.

Figure 6: Overall performance by sub-strand

6.1 Gender by sub-strand

While boys consistently outperform girls across the content sub-strands in both 2014 and 2016, girls outperform boys in working scientifically. As each test booklet contains a significant proportion of working scientifically marks, it is important to note that pupils see a significantly larger number of marks attributed to the working scientifically strand relative to the others. This means overall performance is similar, as seen in table 2 and figure 6.

There has been a decrease in performance on the working scientifically items for boys in 2016, while girls' attainment is similar between 2014 and 2016. The drop in chemistry is visible in both boys' and girls' performance. While attainment in biology remains similar, in physics boys perform better in 2016 while girls perform worse. See table 3 and figure 7.

	Biology	Chemistry	Physics	Working scientifically
2016 Boys	0.47 (0.42, 0.51)	0.21 (0.17, 0.25)	0.58 (0.53, 0.63)	0.11 (0.06, 0.16)
2014 Boys	0.46 (0.41, 0.51)	0.56 (0.52, 0.61)	0.49 (0.45, 0.54)	0.23 (0.18, 0.27)
2016 Girls	0.29 (0.24, 0.34)	-0.05 (-0.09, 0.00)	0.04 (0.00, 0.09)	0.26 (0.21, 0.30)
2014 Girls	0.28 (0.24, 0.33)	0.36 (0.32, 0.41)	0.17 (0.12, 0.23)	0.28 (0.24, 0.32)

Note: ranges given in brackets indicate approximate 95% confidence intervals.

Table 3: Performance on sub-strands by gender

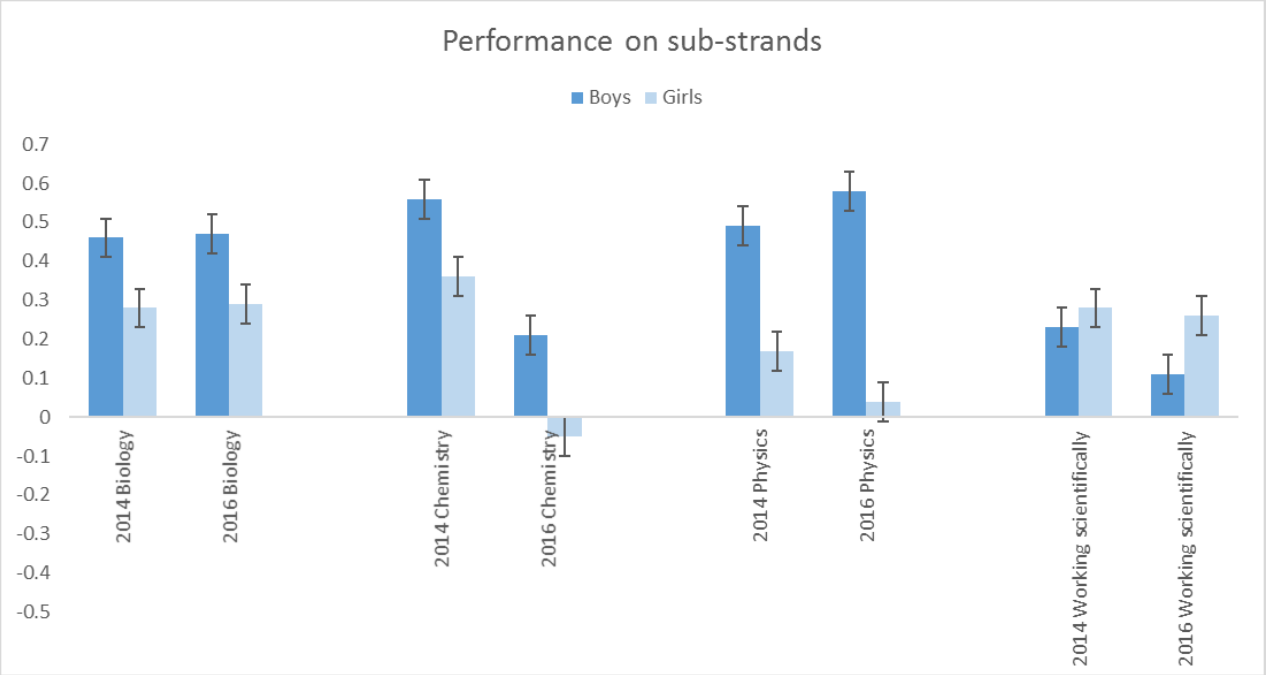


Figure 7: Performance on sub-strands by gender

7 Discussion

The purpose of the KS2 science sampling assessment is to monitor national performance in KS2 science. The context of a changing curriculum must be considered when analysing the outcomes of the matrix sampling exercise. Pupils in 2014 had no exposure to the revised curriculum, however, because of the large overlap in items presented between 2014 and 2016, it is possible to put the 2014 cohort on the same scale as the 2016 cohort.

The 2016 cohort and their teachers would have faced the challenge of implementing the revised science curriculum in year 6 alongside other curricular changes. Therefore, the differences between the 2014 and 2016 outcomes need to be considered in light of this, and any comparisons made with caution.

The administration model of the matrix sample meant that schools did not know until April that they were required to take part in the KS2 science sampling assessment. Schools did not know until 1 week before the assessment which 5 pupils had been randomly selected to participate. Also, no pupil or school level results are available and as a result, there was little opportunity or incentive for schools to undertake preparation activities.

Some evidence for the potential impact of the change to matrix sampling had on school behaviour and pupil motivation comes from the rise in the proportion of pupils not sitting the test. This was stable prior to 2012. 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, increased between 2012 and 2014. In 2012, this was just under 4% of the population. In 2014, it was just over 10%. It remains just over 10% in 2016.

8 Quality assurance and future reporting arrangements

A series of papers to agree details of the matrix design, sample selection, analysis procedures and reporting were presented to STA's technical sub-programme board meetings 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 STA's technical advisory group in February 2014 and again in February 2017, prior to analysis. All analysis was quality checked by a second psychometrician.

The science sampling assessment takes place biennially, with the next administration in June 2018. It is envisaged that reporting the 2018 science sampling outcomes will take place in the summer of 2019 via a further methodological statement set out in a similar structure to this document.

Appendix 1: Test booklet combinations

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.

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

Appendix 2: Sample representation tables

Tables A2.1 and A2.2 shows the representation of the 2016 and 2014 samples, respectively, in terms of the 3 school-level stratifiers of school type, region and FSM band. It confirms that the samples were representative of these school level characteristics.

		Frequency in sample frame	% in sample frame	Frequency in sample	% in sample
School type	Community schools	6,657	42.6	810	42.6
	Voluntary aided and voluntary controlled schools	4,851	31.1	591	31.1
	Foundation schools	611	3.9	73	3.8
	Academies and free schools	2,812	18.0	342	18.0
	Special schools	692	4.4	84	4.4
Region	East Midlands	1,497	9.6	180	9.5
	East of England	1,763	11.3	216	11.4
	London	1,715	11.0	209	11.0
	North East	800	5.1	100	5.3
	North West	2,469	15.8	300	15.8
	South East	2,235	14.3	271	14.3
	South West	1,762	11.3	214	11.3
	West Midlands	1,656	10.6	200	10.5
	Yorkshire and the Humber	1,726	11.0	210	11.1
FSM band	Lowest	3,102	19.9	378	19.9
	Second lowest	3,134	20.1	379	19.9
	Middle	3,149	20.2	384	20.2
	Second highest	3,122	20.0	379	19.9
	Highest	3,116	19.9	380	20.0
Total		15,623		1,900	

Table A2.1: 2016 school-level sample representation

		Frequency in sample frame	% in sample frame	Frequency in sample	% in sample
School type	Community schools	7,267	46.9	891	46.9
	Voluntary aided and voluntary controlled schools	5,257	33.9	642	33.8
	Foundation schools	581	3.7	72	3.8
	Academies and free schools	1,732	11.2	212	11.2
	Special schools	671	4.3	83	4.4
Region	East Midlands	1,493	9.6	182	9.6
	East of England	1,720	11.1	211	11.1
	London	1,705	11.0	208	10.9
	North East	792	5.1	97	5.1
	North West	2,457	15.8	301	15.8
	South East	2,219	14.3	272	14.3
	South West	1,754	11.3	217	11.4
	West Midlands	1,646	10.6	201	10.6
	Yorkshire and the Humber	1,722	11.1	211	11.1
FSM band	Lowest	3,112	20.1	382	20.1
	Second lowest	3,093	19.9	376	19.8
	Middle	3,097	20.0	378	19.9
	Second highest	3,110	20.1	382	20.1
	Highest	3,096	20.0	382	20.1
Total		15,508		1,900	

Table A2.2: 2014 school-level sample representation

Tables A2.3 and A2.4 show the sample representation at pupil level. The column indicating the full sample contains all pupils in the sample. The achieved sample includes just those pupils who took the test or were designated as absent, working below the level of the test or at the level of the test but unable to access it. The test takers column includes just those pupils who were present and took all three booklets.

In 2016, the final achieved sample consisted of 9,409 pupils. The final achieved sample in 2014 consisted of 9,426 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, leaving a total of 8,428 test takers in 2016 and 8,449

test takers in 2014. Having previously accounted for less than 4% of the population, the proportion of pupils not sitting the test rose in 2014 to over 10%. The percentage of pupils who did not sit the test maintained similar levels between 2014 and 2016. 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.

		Freq in full sample	% in full sample	Freq in achieved sample	% in achieved sample	Freq in test takers	% in test takers
Gender	Female	4,600	48.5%	4,567	48.5%	4,178	49.6%
	Male	4,880	51.5%	4,842	51.5%	4,250	50.4%
FSM	No FSM provision	7,838	82.7%	7,790	82.8%	7,134	84.6%
	FSM provision	1,497	15.8%	1,476	15.7%	1,179	14.0%
	Missing	145	1.5%	143	1.5%	115	1.4%
EAL	English as first language	7,756	81.8%	7,702	81.9%	6,944	82.4%
	English as additional language	1,573	16.6%	1,558	16.6%	1,364	16.2%
	Missing	151	1.6%	149	1.6%	120	1.4%
Total		9,480	100.0%	9,409	100.0%	8,428	100.0 %

Table A2.3: 2016 pupil-level sample representation

		Freq in full sample	% in full sample	Freq in achieved sample	% in achieved sample	Freq of test takers	% of test takers
Gender	Female	4,586	48.4%	4,568	48.5%	4,167	49.3%
	Male	4,896	51.6%	4,858	51.5%	4,282	50.7%
FSM	No FSM provision	7,710	81.3%	7,676	81.4%	7,010	83.0%
	FSM provision	1,677	17.7%	1,659	17.6%	1,372	16.2%
	Missing	95	1.0%	91	1.0%	67	0.8%
EAL	English as first language	7,866	83.0%	7,828	83.0%	7,052	83.5%
	English as additional language	1,521	16.0%	1,507	16.0%	1,330	15.7%
	Missing	95	1.0%	91	1.0%	67	0.8%
Total		9,482	100.0%	9,426	100.0%	8,449	100.0%

Table A2.4: 2014 pupil-level sample representation

Standards & Testing Agency

About this publication:

enquiries www.education.gov.uk/contactus

download www.gov.uk/government/publications

Send any enquiries about this publication to science.KS2@education.gov.uk

Reference: ISBN: 978-1-78644-835-4 STA/17/7954/e