Standards & Testing Agency

Key stage 2 science sampling 2018

Methodology note and outcomes

July 2019

Contents

Table of figures

Summary

This publication provides information on the methodology and outcomes of the key stage 2 (KS2) science sampling assessment in 2018. This publication:

- links the 2018 outcomes to the 2016 science sampling assessment outcomes
- contains technical information on the matrix sampling method and analysis
- provides information on the outcomes of the analysis

Expiry or review date

This publication will be reviewed before July 2021.

Who is this publication for?

This publication is for:

- measurement and assessment experts
- school leaders
- school staff
- governing bodies

It applies to all maintained schools, academies and free schools following the national curriculum in science.

Main points

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 approach is 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). These types of large-scale sampling assessments seek to obtain valid and reliable measures of the achievement of the national cohort by administering assessments to a sample of pupils.

The main areas of focus for the 2016 and 2018 science sampling assessments outcomes are detailed in below:

- an estimate of the overall performance of the national cohort in terms of a scaled score based on the 2016 scaled score range (70-120)
- the estimated percentage of pupils achieving the expected standard
- overall performance by gender

- overall performance by pupils with English as an additional language (EAL)
- overall performance by pupils eligibile for free school meals (FSM)
- performance on the 4 content sub-strands of the national curriculum (biology, chemistry, physics and working scientifically)
- performance on the 4 content sub-strands of the national curriculum (biology, chemistry, physics and working scientifically) by gender

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. This followed the same design as the 2014 administration but it assessed attainment against the revised national curriculum. This publication focuses on the third administration, which took place in June 2018.

In line with other KS2 assessments, reporting arrangements changed with:

- the removal of the previous national curriculum levels
- the introduction of new scaled scores
- the setting of a new expected standard of attainment in 2016

This analysis includes performance data from items and pupils in the 2016 and 2018 administrations. There was a large overlap of items to allow the outcomes estimation of the 2016 and the 2018 cohort on the 2016 scale.

The proportion of pupils estimated to be performing at the expected standard decreased slightly from 22.3% in 2016 to 21.2% in 2018. However, it should also be noted that the proportion of pupils not sitting the test also rose from 10% in 2016 to 14% in 2018. Since these pupils are counted as not performing at the expected standard, this will have affected the outcomes.

In both 2016 and 2018, pupils who were eligible for FSM performed significantly worse than their peers, as did pupils with EAL. There were no significant differences between boys and girls in terms of overall performance. Performance on biology and chemistry was very similar across the years. Pupils' performance dropped slightly in physics and slightly increased in working scientifically.

Design

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

Assessment matrix

A large pool of questions is used, with different groups of pupils taking different combinations of questions. This allows test developers to cover a far greater proportion of the programme of study than would normally be covered in a single test instrument. This is known as matrix sampling. This maximises the validity of the outcomes of the assessment, while minimising the burden on individual pupils.

Lord Bew's review[1] of KS2 testing, assessment and accountability recommended this approach for KS2 science sampling. The review recognised that the interim sampling arrangements put in place for 2010 to 2012 did not take advantage of the potential increase in validity that could be gleaned from a matrix sampling approach.

A number of questions comprising 330 marks were selected to cover the assessable areas of the programme of study. These questions were split into 15 booklets of 22 marks each, with 5 booklets covering questions in 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 (Appendix 1) 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 2016 and 2018 administrations, to link performance across those administrations.

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

Sample selection

The sample selection process was the same in 2016 and 2018. 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
- special schools

The selection of schools was also stratified by region, split into:

- London
- South East
- South West
- North East
- North West
- Yorkshire and the Humber
- East of England
- East Midlands
- West Midlands

The selection of schools was also stratified by the proportion of pupils eligible for FSM, split into quintiles[2].

1,900 schools were initially selected with probability of inclusion in the sample being proportional to school size so that each pupil in the population had the same chance of being selected. All schools with pupils eligible to take national curriculum tests are included in the sample, even if they have taken part in previous years. The 2018 sample included 294 schools who were also selected for the 2016 science sample test. This made up 15.5% of the total schools. This is a similar number to the overlap between 2014 and 2016, where 276 schools were selected in both. There were 53 schools in the 2018 sample who were also selected in 2014 and 2016.

Within each of the selected schools, 5 pupils were randomly selected to take part. Some schools had fewer than 5 pupils eligible for selection. In these schools, all pupils were selected. In 2018, the selected sample included 9,481 pupils. 51 pupils were removed

[2] The stratifier was split into fifths.

from the sample due to moving schools in the months before the tests took place, reducing the final achieved sample to 9,430.

In 2016, the selected sample included 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 both years, some of those pupils were:

- absent
- working below the level of the test
- at the level of the test but unable to access it

This meant that these pupils did not actually take the test. This left a total of 8,428 test takers in 2016 and 8,139 test takers in 2018.

The proportion of pupils not sitting the test has been increasing over time. Under the interim sampling arrangements (2010 to 2012), pupils not sitting the test accounted for less than 4% of the population. This rose to over 10% when matrix sampling was introduced in 2014 and maintained similar levels in 2016. This proportion has risen again in 2018 to almost 14%. 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.

Sample representation tables are provided in Appendix 2 at school level and Appendix 3 at pupil level.

Methodology

The methodology for analysing matrix sampling assessments involves a 3-stage process, detailed below.

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 item response theory model (GRM). Data from both 2016 and 2018 were incorporated into a single analysis (concurrent calibration). The common items from 9 booklets were the same across the 2 years. These acted as a link so that all items could be directly compared.

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, science attainment, 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.

Stage 2: latent regression model

As each pupil took a subset of the overall pool of items, the next stage of analysis involved estimating pupils’ 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. This was 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 to reflect the measurement error, which is inherent in the process. For this analysis, 10 plausible values were generated for each pupil. The latent regression and generation of plausible values were run in flexMIRT.

Stage 3: outcomes analysis

The 10 plausible values for each pupil were generated on an IRT ability scale, which centred 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[3].

The outcome of the standard-setting approach was that a raw score of 62 marks on the 100 mark ordered item booklet would represent the threshold for the expected standard. Using the IRT parameters derived from the analysis described above, it was possible to estimate the ability parameter of a pupil with 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, with 100 representing the expected standard. The range of scaled scores available was 70 to 120.

For each pupil, each of their 10 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 met the expected standard'. All statistics for reporting the outcomes were calculated on each set of plausible values and then averaged. These statistics included the percentage at the expected standard and average scaled score. Measurement error was calculated by taking the variance of the statistic across the 10 plausible values.

[3] Further detail on the KS2 standard setting process can be found at this link: https://www.gov.uk/guidance/key-stage-2-tests-standard-setting

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:

1. 600 re-samples were taken from the original sample, with replacement (to achieve 600 samples of the same size as the original sample).
2. 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.
3. The estimates of sampling variance and measurement variance were combined together to produce an overall estimate of the variance using the following formula[4]:

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

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 10 plausible values (those derived from bootstrapping)*
M *is the number of plausible values (10)*
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.

Sub-strand analysis

Sub-strand scores on the IRT scale were estimated for the curriculum content areas:

- biology
- chemistry
- physics
- working scientifically

A bifactor model in flexMIRT was used in this analysis. This is a type of multidimensional IRT model. There is one general factor, overall KS2 science performance, on which all items load. There are several specific factors, for example, the content areas, on which

[4] 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].

subsets of the items can load. Items can load on 2 factors at most, the general factor and one specific factor.

The model was run in flexMIRT in much the same way as the main analysis described above, with the addition of being set up as a bifactor model. This resulted in each pupil being assigned 10 plausible values for the general factor and 10 plausible values for 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. 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-strand scores.

The sub-strand scores are reported on 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.

Outcomes for 2016 and 2018

Attainment in the 2016 and 2018 science sampling exercise is summarised in table 1 for all pupils and split by sub-groups. Overall attainment in the 2018 science sampling exercise is estimated to be slightly lower than in the previous sampling exercise in 2016, but this difference is not statistically significant.

Within each year, the percentage of pupils achieving the expected standard was not significantly different by gender. It was, however, significantly different by FSM and EAL. These pupils were significantly out-performed by their peers. There is no statistical difference between the outcomes across years. For example, the difference in performance for pupils with FSM is not statistically significant between 2016 and 2018.

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.

Characteristic	Estimated percentage achieving expected standard in 2016	95% confidence interval in 2016	Estimated percentage achieving expected standard in 2018	95% confidence interval in 2018
All Pupils	22.3[5]	21.1 – 23.5	21.2	20.0 – 22.4
Boys	22.3	20.5 – 24.1	21.1	19.4 – 22.8
Girls	22.3	20.6 – 24.0	21.3	19.7 – 23.0
FSM	9.0	6.9 – 11.0	9.4	7.3 – 11.5
Non-FSM	24.8	23.4 – 26.1	23.3	22.0 – 24.7
EAL	17.2	14.7 – 19.7	16.5	13.9 – 19.1
Non-EAL	23.3	21.9 – 24.6	22.4	21.0 – 23.7

Table 1: Estimated percentage of pupils achieving the expected standard based on KS2 science sampling assessments in 2016 and 2018

[5] The proportion of pupils estimated to have reached the expected standard in the 2016 analysis was reported at the time as 22.8, which rounded to 23%. That analysis included pupils who took the sample test in 2014, under the previous curriculum. The 2018 analysis included data from pupils who took the assessment in 2016 and 2018, excluding pupils who took the assessment in 2014. This resulted in a small change in the estimate of the 2016 percentage achieving the expected standard, 22.3, which rounds down to 22%.

The scaled score distribution for 2016 and 2018 is shown in figure 2. The scaled score range is 70 to 120. The shapes of the distributions are very similar.

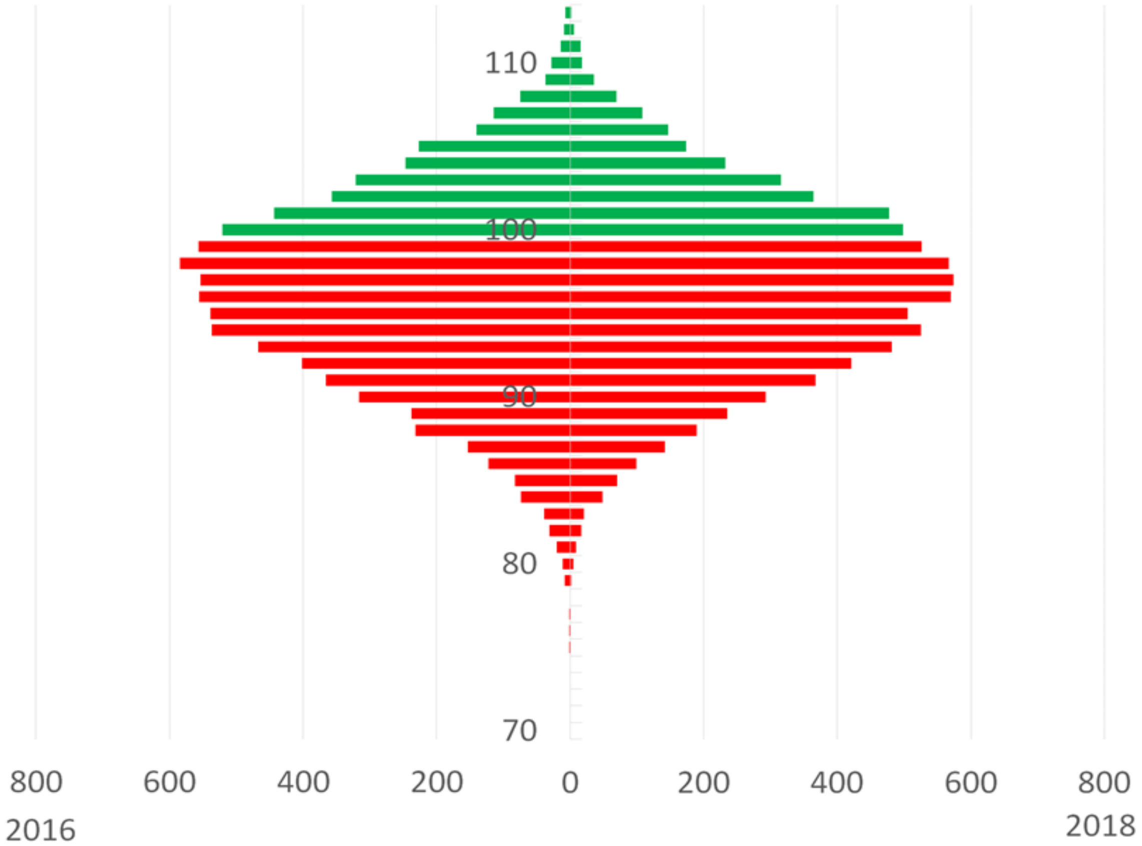
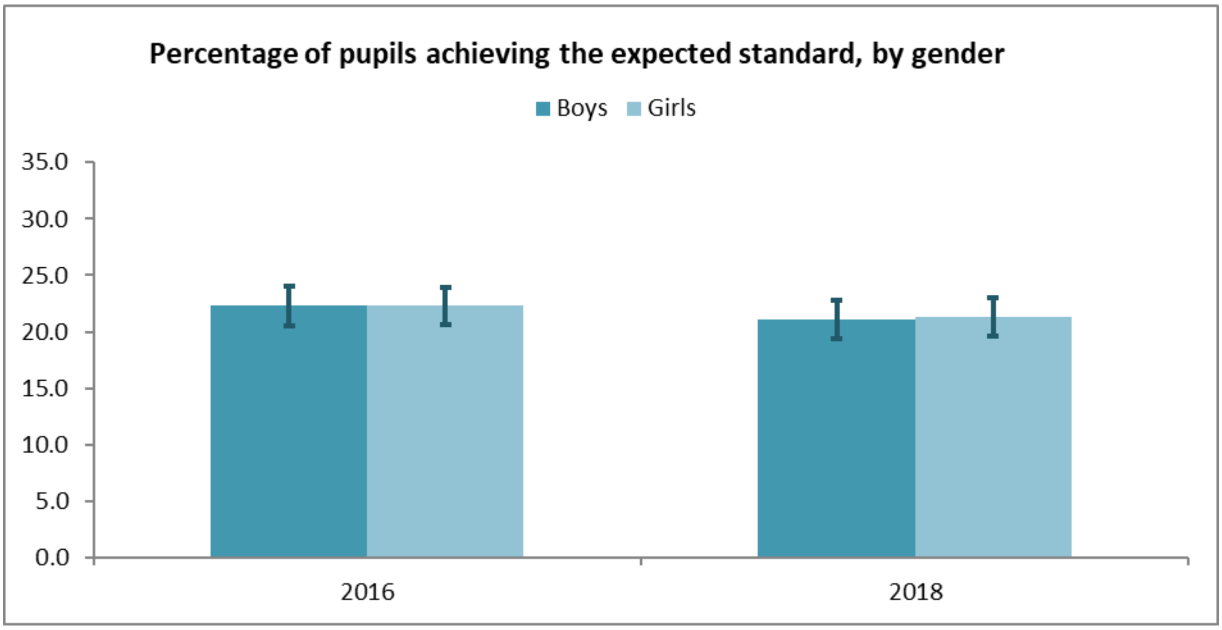


Figure 2: Distribution of scaled scores

Gender

As in previous years, performance of girls and boys was very similar. There was no significant difference in the percentages of girls and boys achieving the expected standard in either 2016 or 2018. In 2016, just over 22% of boys and girls were estimated to have reached the expected standard. In 2018, just over 21% of boys and girls were estimated to have reached the expected standard.

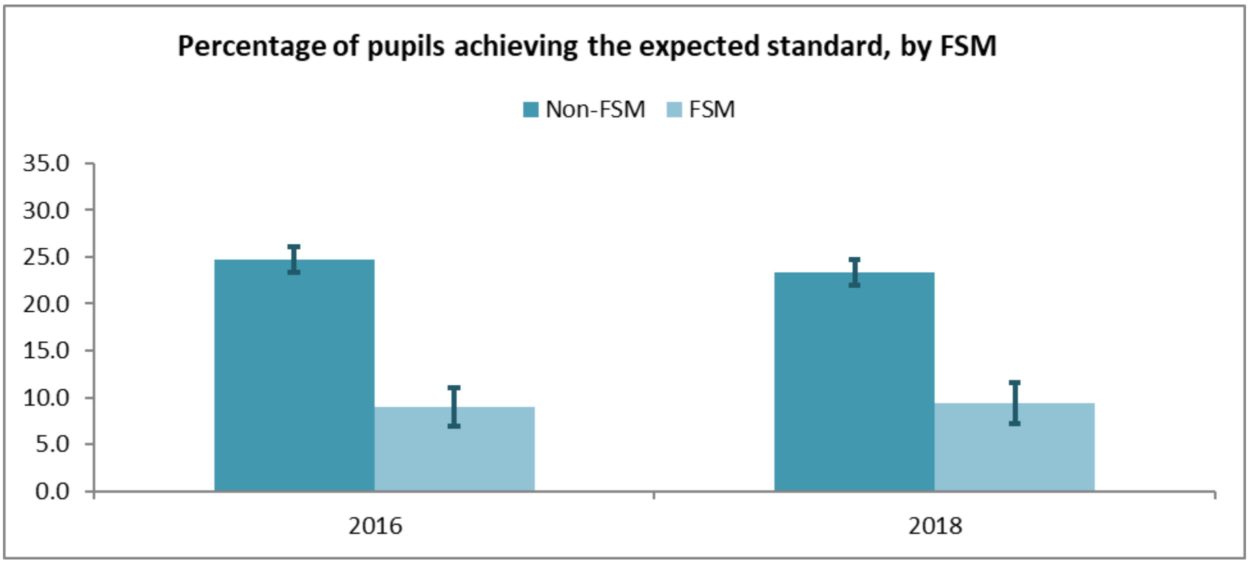


Note: bars around data points indicate approximate 95% confidence intervals.

Figure 3: Percentage of pupils achieving the expected standard, by gender

Free school meals

The performance of pupils eligible for FSM was significantly lower than other pupils. In 2018, just over 9% of FSM eligible pupils were estimated to have reached the expected standard. This is compared to just over 23% of non-FSM pupils. In 2016, just under 9% of FSM pupils were estimated to have reached the expected standard. This compared to nearly 25% of non-FSM pupils.

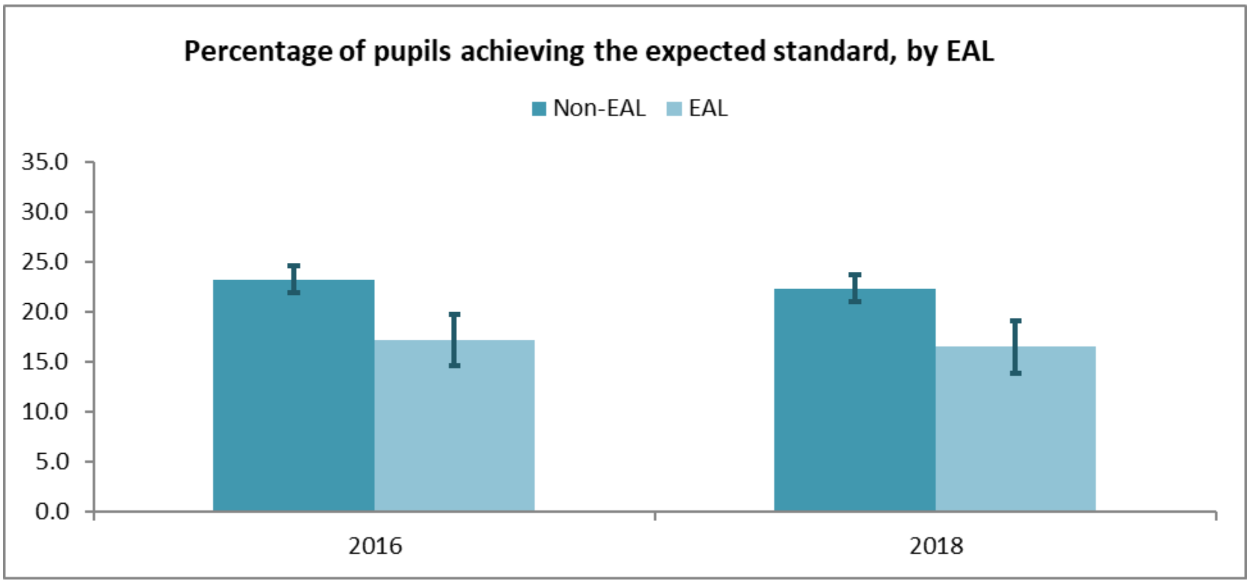


Note: bars around data points indicate approximate 95% confidence intervals.

Figure 4: Percentage of pupils achieving the expected standard, by FSM eligibility

English as an additional language

The performance of pupils with EAL was significantly lower than other pupils. In 2018, just under 17% of pupils with EAL compared to just over 22% of non-EAL pupils were estimated to have reached the expected standard. In 2016, just over 17% of pupils with EAL were estimated to have reached the expected standard, compared to just over 23% of non-EAL pupils.



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

Figure 5: Percentage of pupils achieving the expected standard, by EAL

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, chemistry and physics is relatively similar from 2016 to 2018. There is a small increase to pupils' estimated ability in working scientifically between 2016 and 2018. This is illustrated in table 2.

Subject	2016 sub score performance	2018 sub score performance	Marks assessing strand in 2016	Marks assessing strand in 2018
Biology	-0.01 (-0.05, 0.03)	-0.01 (-0.04, 0.03)	77	84
Chemistry	0.10 (0.07, 0.14)	0.14 (0.11, 0.18)	68	62
Physics	0.43 (0.39, 0.47)	0.33 (0.29, 0.36)	69	78
Working scientifically	-0.01 (-0.05, 0.02)	0.09 (0.06,0.12)	115	106

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

Table 2: Performance on sub-strands

Gender by sub-strand

While boys consistently outperform girls across the content sub-strands in both 2016 and 2018, girls outperform boys in working scientifically. 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.

In both years, boys significantly outperformed girls in chemistry and physics but girls outperformed boys in working scientifically. Boys also performed better than girls in biology. This difference was significant in 2016 but not in 2018. See table 3.

Year	Gender	Biology	Chemistry	Physics	Working scientifically
2016	Boys	0.06 (0.00, 0.11)	0.23 (0.17, 0.28)	0.71 (0.66, 0.76)	-0.09 (-0.14, -0.04)
2018	Boys	0.02 (-0.03, 0.06)	0.33 (0.27, 0.38)	0.69 (0.64, 0.73)	-0.01 (-0.05, 0.03)
2016	Girls	-0.08 (-0.13, -0.03)	-0.03 (-0.08, 0.02)	0.15 (0.10, 0.21)	0.07 (0.02, 0.11)
2018	Girls	-0.03 (-0.08, 0.02)	-0.04 (-0.09, 0.00)	-0.04 (-0.08, 0.01)	0.19 (0.14, 0.24)

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

Table 3: Performance on sub-strands by gender

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 purpose of these papers was to agree details of:

- the matrix design
- sample selection
- analysis procedures
- reporting

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. STA needed to use new techniques, as detailed 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 every two years. The next administration is planned to take place in June 2020. STA expect to report on the 2020 science sampling outcomes in the summer of 2021. This will take place using a methodology paper, set out in a similar structure to this document.

Appendix 1: Test booklet combinations

The 15 KS2 test booklets are denoted below, with a B, C or P suffix to indicate the core content area assessed.

Combination	1st booklet	2nd booklet	3rd booklet
1	ST016B	ST010C	ST020P
2	ST022B	ST024C	ST026P
3	ST003B	ST019C	ST021P
4	ST017B	ST018C	ST015P
5	ST023B	ST025C	ST027P
6	ST010C	ST021P	ST022B
7	ST024C	ST015P	ST003B
8	ST019C	ST027P	ST017B
9	ST018C	ST020P	ST023B
10	ST025C	ST026P	ST016B
11	ST020P	ST003B	ST025C
12	ST026P	ST017B	ST010C
13	ST021P	ST023B	ST024C
14	ST015P	ST016B	ST019C
15	ST027P	ST022B	ST018C

Appendix 2: School sample representation tables

Tables A2.1 and A2.2 show the representation of the 2016 and 2018 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.

Stratifier	Frequency in sample frame	% in sample frame	Frequency in sample	% in sample
School type—community schools	6,657	42.6	810	42.6
School type—voluntary aided and voluntary controlled schools	4,851	31.1	591	31.1
School type—foundation schools	611	3.9	73	3.8
School type—academies and free schools	2,812	18.0	342	18.0
School type—special schools	692	4.4	84	4.4
Region—East Midlands	1,497	9.6	180	9.5
Region—East of England	1,763	11.3	216	11.4
Region—London	1,715	11.0	209	11.0
Region—North East	800	5.1	100	5.3
Region—North West	2,469	15.8	300	15.8
Region—South East	2,235	14.3	271	14.3
Region—South West	1,762	11.3	214	11.3
Region—West Midlands	1,656	10.6	200	10.5
Region—Yorkshire and the Humber	1,726	11.0	210	11.1
FSM band—lowest	3,102	19.9	378	19.9
FSM band—second lowest	3,134	20.1	379	19.9
FSM band—middle	3,149	20.2	384	20.2

Stratifier	Frequency in sample frame	% in sample frame	Frequency in sample	% in sample
FSM band—second highest	3,122	20.0	379	19.9
FSM band—highest	3,116	19.9	380	20.0
Total	15,623	-	1,900	-

Table A2.1: 2016 school-level sample representation

Stratifier	Frequency in sample frame	% in sample frame	Frequency in sample	% in sample
School type—community schools	5,730	36.3	691	36.4
School type—voluntary aided and voluntary controlled schools	4,450	28.2	537	28.3
School type—foundation schools	530	3.4	63	3.3
School type—academies and free schools	4,339	27.5	523	27.5
School type—special schools	715	4.5	86	4.5
Region—East Midlands	1,529	9.7	184	9.7
Region—East of England	1,783	11.3	216	11.4
Region—London	1,745	11.1	211	11.1
Region—North East	811	5.1	98	5.2
Region—North West	2,469	15.7	298	15.7
Region—South East	2,269	14.4	274	14.4
Region—South West	1,767	11.2	211	11.1
Region—West Midlands	1,665	10.6	200	10.5
Region—Yorkshire and the Humber	1,726	10.9	208	10.9
FSM band—Lowest	3,187	20.2	385	20.3
FSM band—Second lowest	3,119	19.8	376	19.8
FSM band—Middle	3,142	19.9	377	19.8
FSM band—Second highest	3,230	20.5	389	20.5
FSM band—Highest	3,086	19.6	373	19.6
Total	15,764	-	1,900	-

Table A2.2: 2018 school-level sample representation

Appendix 3: Pupil sample representation tables

Tables A3.1 and A3.2 show the sample representation at pupil level. The columns indicating the full sample contain all pupils in the sample. The achieved sample columns include just those pupils who:

- took the test
- were designated as absent
- were working below the level of the test or at the level of the test but were unable to access it.

The test takers column includes just those pupils who were present and took all 3 booklets.

Stratifier	Frequency in full sample	% in full sample	Frequency in achieved sample	% in achieved sample	Frequency in test takers	% in test takers
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
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 FSM provision	145	1.5	143	1.5	115	1.4
Non-EAL	7,756	81.8	7,702	81.9	6,944	82.4
EAL	1,573	16.6	1,558	16.6	1,364	16.2
Missing EAL	151	1.6	149	1.6	120	1.4
Total	9,480	-	9,409	-	8,428	-

Table A3.1: 2016 pupil-level sample representation

Stratifier	Frequency in full sample	% in full sample	Frequency in achieved sample	% in achieved sample	Frequency of test takers	% of test takers
Female	4,581	48.3	4,565	48.4	4,063	49.9
Male	4,900	51.7	4,865	51.6	4,076	50.1
No FSM provision	7,998	84.4	7,969	84.5	7,057	86.7
FSM provision	1,360	14.3	1,339	14.2	997	12.2
Missing FSM provision	123	1.3	122	1.3	85	1.0
Non-EAL	7,683	81.0	7,643	81.0	6,606	81.2
EAL	1,671	17.6	1,661	17.6	1,444	17.7
Missing EAL	127	1.3	126	1.3	89	1.1
Total	9,481	-	9,430	-	8,139	-

Table A3.2: 2018 pupil-level sample representation

Standards & Testing Agency

About this publication:

enquiries www.education.gov.uk/contactus
download www.gov.uk/government/publications

Reference: ISBN: 978-1-78957-449-4 STA/19/8346/e