Estimating sub-domain scores within national assessments. A case study using the Key Stage 2 National Curriculum mathematics tests in England.

John Jerrim[1] (UCL Social Research Institute)

Natasha Plaister (FFT Education Datalab)

Dave Thomson (FFT Education Datalab)

[1] Social Research Institute, University College London, 20 Bedford Way London, WC1H 0AL. E-mail: j.jerrim@ucl.ac.uk (John Jerrim). +44 7590 761 755

Acknowledgements: We would like to thank our project advisory board for their helpful comments on the project. The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds and undertakes rigorous research, encourages innovation and supports the use of sound evidence to inform social and economic policy, and improve people's lives. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. This project has been funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily the Foundation. Find out more at: nuffieldfoundation.org.

Bluesky: @nuffieldfoundation.org

LinkedIn: Nuffield Foundation

This work was undertaken in the Office for National Statistics (ONS) Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners.

We are grateful to the Department for Education for data access.

Executive summary

Background

National assessments of mathematics, such as England's Key Stage 2 National Curriculum Tests (SATs), serve a wide range of purposes: measuring pupils' mastery of the curriculum, providing evidence for school accountability and tracking national standards over time. These assessments are influential, shaping the decisions of policymakers, schools and teachers alike. They also cost tens of millions of pounds annually and involve several hours of test time. Given their importance, the way results are reported has a profound impact on how the data are used.

Beyond overall performance, schools often seek more granular insights into their strengths and weaknesses. One common approach is Question Level Analysis (QLA), where teachers review how pupils answered specific items. While attractive in principle, QLA is vulnerable to measurement error, meaning that apparent patterns may simply reflect random variation.

Sub-domain scores, which summarise performance across different areas of the curriculum (e.g. algebra, geometry), represent an alternative. Schools can access such sub-domain scores from the Key Stage 2 tests via the Department for Education's Analyse School Performance tool. Yet the information provided is simply raw scores - percentage of correct responses to questions within each primary National Curriculum area – without any articulation of their reliability or uncertainty. There is consequently a risk that this information is misinterpreted by schools.

This project has explored a more psychometrically principled approach to producing Key Stage 2 sub-domain scores (for eight separate areas of the National Curriculum). In doing so, it has sought to investigate whether it is possible to produce Key Stage 2 mathematics sub-domain scores that are genuinely useful to teachers and schools.

Data

The analysis is based on Key Stage 2 mathematics SATs, taken annually by the majority of 10- and 11-year-olds in state schools in England. These assessments consist of three papers – one arithmetic and two reasoning papers – covering a total of around 110 marks and 80–90 questions. Each test maps onto eight National Curriculum domains: Algebra,

Calculations, Fractions, Geometry, Measurement, Number, Ratio, and Statistics, though the number of marks allocated to each domain varies significantly from year to year. The project has had access to the question-level marks of pupils, facilitating our construction of sub-domain scores and their associated psychometric properties. The analysis is based on a sample of 500 schools (totalling 76,033 pupils between 2018 and 2023). Results were also replicated very consistently across a larger sample (1,500 schools).

Methodology

The project first conducts a basic analysis of the Key Stage 2 item-level data broadly based on classical test theory (CTT). This includes producing summative scores and investigating their internal consistency (Cronbach's alpha). We also use factor analysis to explore the number of latent constructs that are measured within the Key Stage 2 mathematics test data, and the extent to which questions load onto the primary National Curriculum areas that they are supposed to capture.

Multi-dimensional Item Response Theory (MIRT) with latent regression is then used to produce Key Stage 2 sub-domain scores for each curriculum domain. This methodology improves on reporting raw sub-domain scores by accounting for measurement error in the data (e.g. due to the limited number of questions measuring each mathematics sub-domain) and producing indicators of measurement uncertainty. It is similar to the approach used in many large-scale international assessments, such as PISA, TIMSS and the National Reference Test.

Several analyses are conducted to investigate the reliability and utility of school-level sub-domain scores. This includes investigating their robustness to different model specifications, their stability across academic years and the strength of the correlations across the different National Curriculum areas. The strength of the correlations across the different domains is then compared to the analogous results from TIMSS. Subgroup analyses are then conducted where we explore demographic gaps in attainment across domains, comparing these against TIMSS (where possible) as an external benchmark.

Results

First, sub-domain raw scores are problematic. The Key Stage 2 mathematics test appears to essentially be unidimensional, indicating that any attempt to measure eight separate

factors may be challenging. Test questions do not seem to clearly cluster together within their proposed National Curriculum areas. Moreover, while reliability estimates were high for some domains with many items (e.g. Calculations, Fractions), they were much lower for domains with fewer questions (e.g. Algebra, Geometry, Statistics). This means that raw sub-domain scores are not suitable for school-level reporting.

Second, reliable Key Stage 2 mathematics sub-domain scores can be generated using MIRT latent regression that can be used for school-level reporting, particularly when pooling data over multiple years. However, the school-level correlation in these sub-domains is very strong (Pearson correlation often close to 0.99) – to such an extent that they provide little unique information over and above one another (or indeed overall Key Stage 2 mathematics scores). In other words, it is not possible to distinguish schools that have a comparative strength in one area (e.g. Statistics) and a comparative weakness in another (e.g. Geometry). School-level correlations of a similar magnitude across sub-domains are observed for the TIMSS assessment.

Third, there is moderate stability of sub-domain school-level scores across academic years. The level of stability is similar to overall mathematics scores, with inter-year correlations around 0.65–0.70, indicating that schools' relative performance remains moderately consistent from year to year. Pooling data across years would lead to greater stability in primary school performance measures (both overall and sub-domain scores), while also ensuring broad content coverage within each sub-domain.

Finally, differences across demographic groups for each sub-domain were of similar size and direction to those based on overall Key Stage 2 mathematics scores. There are two possible interpretations of this result. Either demographic groups genuinely perform very similarly across different aspects of mathematics, or the current design of the Key Stage 2 test is not sufficiently refined to detect such subtle differences across domains.

<u>Recommendations</u>

1. Discontinue the provision of Key Stage 2 sub-domain raw scores.

The Department for Education should stop providing schools with sub-domain raw scores within its Analyse School Performance tool. These figures are presented without any accompanying indication of uncertainty, making them prone to misinterpretation. School leaders and staff are not experts in the statistical nuances of such data, and the provision of raw, unqualified results carries a serious risk of schools reaching erroneous and potentially harmful conclusions.

2. Prioritise fewer, higher-quality indicators.

The Department for Education should more broadly review the type and quantity of information that is fed back to schools. In this context, less is often more. Making many pieces of information available – such as sub-domain scores – risks diverting schools' attention and resources. Schools would be better served by the provision of a smaller number of carefully selected, robust pieces of information that they can act upon with confidence. The sub-domain scores provide a clear example where the availability of more data is counterproductive, creating noise rather than clarity.

3. Reform sub-domain reporting where demand exists.

If, due to user demand, the Department for Education chooses to continue reporting sub-domain scores within the Analyse School Performance tool, these must be constructed using a more robust methodology. The current practice of reporting raw scores without measures of uncertainty is inadequate and misleading. The methodology set out in this report offers a more principled approach. Explicitly accounting for measurement error enables schools to understand their comparative position across curriculum areas in a way that is both reliable and informative. It would thereby reduce the risk of schools making misinformed decisions.

4. Redesign Key Stage 2 tests if they are to provide diagnostic information.

If the Department for Education wishes Key Stage 2 assessments to serve a diagnostic purpose, then the tests themselves need to be fundamentally redesigned. In their current form, the assessments are not fit to inform schools of their (and their pupils') relative strengths and weaknesses. Attempting to repurpose them as diagnostic tools without structural change is unwise and risks generating misleading conclusions about pupil

attainment in specific curriculum areas. A redesigned assessment would require a stronger balance of questions across domains, as well as items that more effectively discriminate between pupils' skills across different areas of the National Curriculum.

5. Base Key Stage 2 accountability on multi-year averages.

Beyond the issue of sub-domain reporting, our analysis also highlights the wider problem of volatility in Key Stage 2 results at the school level. Year-on-year variation is substantial, driven largely by the small size of primary school cohorts (e.g. on average, there were 42 Year 6 pupils per school in 2024). Therefore, reliance on single-cohort results is inappropriate for accountability purposes. The Department should reform the accountability framework so that school-level performance is assessed using multi-year averages. This would provide a more stable and accurate measure of school performance and reduce distortions arising from school cohort sizes. For further discussion of this issue, see Menzies and Jerrim (2021).

1. Introduction

Many jurisdictions conduct national or state assessments measuring pupils' mathematics knowledge, understanding and skills during primary school (OECD, 2023). These examinations are developed with different purposes in mind. While some seek to summarise what individual pupils have learned of the set curriculum (e.g. Key Stage 2 and GCSE mathematics tests in England), others are primarily designed to draw inferences at the group (e.g. state, country, school) level (e.g. the National Reference Test in England; PISA, TIMSS and PIRLS internationally). Many also feed into systems of school accountability, where school-level results are published and may help inform parental school choice (Allen & Burgess, 2011). It is also, of course, possible for assessments to be used for multiple purposes at once (Newton, 2007).

Some of these assessments do not, however, come cheap – both in terms of financial commitment and test time. A prime example comes from England – the empirical setting of this paper – where Key Stage 2 National Curriculum Tests (SATs) are conducted at the end of the primary school course (when pupils are age 10/11) each year. These cover English and mathematics, run over four consecutive school days and total 110 minutes of test time dedicated to mathematics. The most recent data available suggests running these tests costs tens of millions of pounds each year (Ward, 2017). In return, the data the SATs generate are used for several purposes. As well as attempting to measure pupils' mathematics ability, they also have a prominent role in both primary and secondary school accountability, measuring national trends in primary pupils' mathematics skills and are widely used as a key outcome measure in educational research.

Given the high-stakes nature of the Key Stage 2 SATs, it is little wonder that schools closely watch the results. For most schools, the focus is overall results (e.g. school average mean scores and the percentage achieving the expected standard) as these form the headline measures of school accountability. Yet item (question) level data is also provided back to schools, leading to interest in QLA. This is where schools examine response patterns to individual questions to try to better understand their relative strengths and weaknesses. Indeed, commercial tools are now available to assist schools with such endeavours (e.g. https://www.primarytools.co.uk/ks2-qla/; https://daisi.education/qla-results/). A recent Teacher Tapp poll in England indicated that

half of teachers in England have had to manually enter data over the last academic year to conduct a QLA (Teacher Tapp, 2024). This may, in theory, lead to teachers and schools changing the amount of time they spend on different topics, altering aspects of their curriculum or altering their pedagogical approach to certain topics.

Of course, the challenges with such QLA are well-known within the educational assessment literature. This includes the substantial measurement error that is inherent in individual item responses (Allen, 2021). Such limitations are, however, likely to remain underappreciated or poorly understood amongst practitioners within schools (and, indeed, more broadly outside of the psychometric community). Consequently, the QLAs undertaken by schools could lead to mistaken inferences being drawn.

Ideally, then, what may be needed is a halfway house – school-level estimates that provide richer insights than overall test scores but provide a more robust basis for making inferences than performance on individual test questions. This is the role of "sub-domain" scores – measures that capture pupils' performance on certain aspects of a test (e.g. a specific area of the National Curriculum). Users of tests (e.g. teachers and schools) often want such information to inform key aspects of their teaching and curriculum development. Yet – just like QLA – there is a risk of such scores doing more harm than good. If they are not produced with sufficient reliability – and any uncertainty in their measurement clearly articulated – then it could lead users to draw erroneous inferences (e.g. believing that their pupils are performing poorly in one particular area when this is not the case). This could lead users to invest their time and effort into the wrong areas, including those that do not require any change. Therefore, for the reporting of sub-scores to be justifiable, they must meet at least three important conditions: (1) they are measured with sufficient reliability; (2) they add additional information over and above the overall test score; and (3) they include an indication of measurement uncertainty. A principled and well-tested approach to creating such sub-domain scores is thus needed.

Currently, schools in England can access their Key Stage 2 sub-domain scores from the Department for Education's Analyse School Performance tool. However, these are provided in the form of "raw scores" (percentage correct) compared to national averages, with no measure of statistical uncertainty. An example of how this information is

presented to schools can be found in Table 1 below. As noted in the paragraph above, presenting these data to schools in this way carries significant risks in terms of overinterpreting apparent differences in performance across different parts of the mathematics curriculum.

Table 1. Example of the sub-domain information reported back to a school in the 2024/25 academic year (fictional data)

	Marks available	Correct response %	National %	Difference
Calculations	42	56	73	-17
Algebra	4	53	58	-5
Fractions, decimals and percentages	25	64	63	1
Geometry - position and direction	3	63	79	-16
Geometry - properties of shapes	4	53	56	-3
Measurement	9	68	63	5
Number and place value	11	69	83	-14
Ratio and proportion	7	65	52	13
Statistics	5	60	71	-11
Total	110	67	68	-1

Notes: Fictional data.

The central aim of this paper is to set out and test a more robust approach to creating Key Stage 2 mathematics sub-domain scores. Specifically, we investigate the psychometric properties – and potential insights to be gained – of producing separate Key Stage 2 SATs mathematics scores across the eight National Curriculum areas - Algebra, Calculations, Fractions, Geometry, Measurement, Number, Ratio, Statistics. While we recognise that the reporting of these scores will not be appropriate at the individual level (due to the large measurement error in individual scores when sub-tests are short), there is the potential that they may provide a reasonable basis for making group-level comparisons. We investigate this possibility at both the national level (e.g. the average difference between boys and girls in the eight National Curriculum areas of mathematics) and the local level (e.g. differences in performance on the eight National Curriculum areas across schools).

Specifically, we address the following research questions:

- Research question 1. Is it possible to use the SATs results to create scores for each area of the national mathematics curriculum that provide a robust basis for making comparisons across schools?
- Research question 2. How big are the gender, Free School Meal (FSM) eligibility, English as an Additional Language (EAL) and special educational needs gaps across the different areas of the national mathematics curriculum?

2. Data. The Key Stage 2 mathematics test.

The data we use are drawn from England's National Curriculum Tests (SATs), which pupils sit in May of Year 6 (age 10/11) at the end of primary school, focusing on the mathematics assessment. They are taken by most state school pupils in England, although they are not mandatory in independent (private) schools. They are designed to perform two key functions. First, they provide information on individual pupils' achievement in mathematics, as specified in England's National Curriculum. Second, the results play a key role in school accountability; "league tables" of schools are produced using the results and are considered by OFSTED (the school inspectorate) in their inspection of schools.

The mathematics test is divided into three separate papers, one focused on mental arithmetic and two on mathematics reasoning. In most years, each of these papers includes around 25 to 35 questions with around 35 to 40 marks available in each. Each paper has a time limit (30 minutes for the mental arithmetic paper and 40 minutes for each reasoning paper). Thus, across the three papers, there are a total of around 80 to 90 questions and 110 marks, with a total test time of 110 minutes.

The test covers eight National Curriculum areas of mathematics:

- Algebra.
- Calculations.
- Fractions.
- Geometry.
- Measurement.
- Number.

- Ratio.
- Statistics.

As illustrated by Table 2, these are not allocated equal test time. In particular, the test is weighted towards Calculations and Fractions. The different National Curriculum areas also vary in the position that questions appear on the test papers and their difficulty. For instance, Number questions tend to be easier and appear towards the start of papers, while Ratio questions tend to be harder and thus are towards the end (as the papers are generally designed to be ramped in difficulty). Most questions are worth a single mark, although there are a small number of partial credit items (between 2017 and 2023, 76% of items were worth one mark, 23% two marks and 1% three marks). All tests are externally marked, with the marking reliability high (Ofqual, 2024). Copies of the tests are available from the Department for Education website (e.g. https://www.gov.uk/government/publications/key-stage-2-tests-2024-mathematics-test-materials), with further technical information on the questions published by the Standards and Testing Agency (STA). This includes a mapping between each test item and the eight National Curriculum areas. See STA (2025) for further information.

Table 2. The number of marks available in each National Curriculum area by year

	2017	2018	2019	2022	2023
Algebra	5	6	5	3	5
Calculations	37	38	38	42	37
Fractions	24	27	24	24	26
Geometry	8	7	10	9	8
Measurement	13	11	11	11	10
Number	11	10	9	9	10
Ratio	8	7	8	7	9
Statistics	4	4	5	5	5

Notes: Author's calculations based on information published by the STA (2025).

The findings we present are based on a random sample of 500 schools drawn from around 15,000 that administer the tests each year. This leads to a total pupil sample size of 76,033 pupils between 2018 and 2023. We have tested the robustness of our key findings using a larger sample of 1,500 schools, and very similar results emerge.

3. Initial investigations of the psychometric properties of the Key Stage 2 tests

Question positions, difficulty and items not reached

Table 3 begins by presenting information on the average position that questions appear within the Key Stage 2 test papers, according to their primary National Curriculum domain. Specifically, within each of the three test papers, the first question is assigned a rank position of 1, the second question a rank position of 2, and so forth (through to a rank position of n for the final item). These rank positions of items are then averaged across the three test papers. Note that – as highlighted within the Key Stage 2 technical appendix (Department for Education, 2025) – the test is designed so that easier questions tend to appear towards the start of the papers, with harder questions towards the end.

Questions covering the different domains are not evenly distributed across the papers. Those covering the Number domain appear early, with an average position of 6th place. In contrast, Ratio questions tend to be towards the end, with an average position of 24th. The average positions of Calculations, Statistics, Geometry, Measurement and Algebra questions tend to be similar – typically being between 13th and 18th position. Average positions of items across the domains appear to be reasonably stable across years, although with slightly more fluctuation in domains such as Statistics, Geometry and Algebra (as expected, given these domains are measured using fewer questions).

Table 3. The average position of items within test papers by National Curriculum area

	2017	2018	2019	2022	2023	Average item position
Number	10	8	4	5	5	6
Calculations	12	13	14	13	15	13
Statistics	5	13	17	16	18	14
Geometry	14	12	18	19	9	14
Measurement	16	17	16	14	16	16
Algebra	19	16	10	20	26	18
Fractions	21	20	20	19	20	20
Ratio	24	21	24	26	24	24

Notes: Figures refer to the average position of questions on a given topic within test papers. For instance, the typical Number question is around the 6th question the pupil faces on the test (i.e. they tend to be amongst the first questions answered). On the other hand, Ratio questions tend to come much later in the test, with the average question in this domain being the 24th question the pupil faces on a given paper.

Sub-domain scores based on classical test statistics

Next, we present information from the 2023 assessment on item-rest correlations and Cronbach's alpha. The former can be considered a measure of how well each question

captures the area of the National Curriculum it is designed to measure. The latter indicates the correlation between items within each National Curriculum area, which are as follows:

- Calculations = 0.92
- Fractions = 0.92
- Number = 0.78
- Ratio = 0.75
- Measurement = 0.74
- Algebra = 0.70
- Geometry = 0.61
- Statistics = 0.51

In the Calculations and Fractions domains – the two domains including the most test questions – the reliability of the measure (as captured by Cronbach's alpha) is high at 0.92. This provisionally indicates that – in these two domains – obtaining reliable sub-domain scores should be possible. In contrast, for most of the other areas, Cronbach's alpha is relatively low – again reflecting the smaller number of test questions and marks available. This suggests that – as expected – sub-domain scores using a CTT approach are likely to prove unreliable and will be particularly ill-suited to individual-level reporting.

Table 4 then presents correlations across the summative (raw) sub-domain scores. The association between pupils' scores in the Calculations and Fractions domains is high, standing at 0.90. However, for most of the other comparisons, the associations stand between around 0.6 and 0.75. For the Algebra domain, the associations are weaker, sitting between 0.45 and 0.65. Thus, overall, there appears to be a reasonable association in CTT scores across the sub-domains.

It is important to note, however, that the magnitude of these sub-domain correlations will be attenuated (downwardly biased) due to two related factors. First, they do not account for the measurement error present in these test scores – a key problem given the limited number of questions within some of these domains. Second, they do not account for the fact that these scores suffer from ceiling effects (within a sub-domain, many pupils may achieve the maximum score). Consequently, these correlations should be treated as

lower bounds for the true latent correlation in pupils' skills across the different areas of mathematics.

Table 4. Correlations in summative scores across domains

	Calculations	Fractions	Number	Geometry	Measurement	Statistics	Ratio	Algebra
Calculations								
Fractions	0.90							
Number	0.74	0.70						
Geometry	0.63	0.66	0.54					
Measurement	0.75	0.79	0.59	0.59				
Statistics	0.69	0.71	0.56	0.55	0.66			
Ratio	0.79	0.82	0.58	0.57	0.76	0.67		
Algebra	0.61	0.66	0.44	0.55	0.64	0.56	0.63	

Notes: Figures based on 20,905 observations. Figures refer to the correlation in scores across domains based on summative scores.

In Table 5, we turn to CTT measures of item facility/difficulty, as measured by the percentage of correct responses. Number questions tend to be the easiest, with the average item being answered correctly by 83% of pupils. This is consistent with these questions appearing towards the start of the Key Stage 2 test papers. Calculation questions also stand out as typically being easier than the remaining six domains, with the average question being answered correctly by 78% of pupils. These two domains (Number and Calculations) are thus characterised as including easier questions than the other six domains.

There is relatively little difference in the average percentage correct across the remaining six domains, ranging from 67% in Fractions and Statistics to 56% in Algebra. There is, in particular, little difference (when looking at the average percentage correct between 2017 and 2023) between Geometry (60%), Ratio (57%), Measurement (57%) and Algebra (56%). For these domains – which typically contain fewer questions – there is quite a large degree of inter-year variation. Take Geometry, for example. In 2022, the average Geometry question was answered correctly by 47% of pupils, compared to 70% in 2023. In comparison, inter-year variation in the average difficulty of Calculations and Fractions questions is comparatively small (reflecting the larger number of questions asked each year within these domains).

Table 5. Average facility (percentage of items answered correctly) by National Curriculum area

	2017	2018	2019	2022	2023	Average % correct
Number	74	82	88	82	87	83
Calculations	79	80	80	74	75	78
Fractions	64	69	70	68	65	67
Statistics	77	67	67	61	64	67
Geometry	61	66	55	47	70	60
Ratio	57	63	58	53	56	57
Measurement	58	53	60	61	55	57
Algebra	51	55	75	60	40	56

Next, we turn to the percentage of questions that were omitted by pupils (no response provided) or recorded as "not reached" (where pupils do not respond to several questions in a row, and with no further test questions answered). These results are provided in Tables 6 and 7.

Table 6. Average percentage of questions omitted (no response provided) by National Curriculum area

	2017	2018	2019	2022	2023	Average % omitted
Ratio	9	6	9	10	11	9
Algebra	9	5	2	4	10	6
Fractions	7	5	3	6	7	6
Measurement	6	5	4	3	3	4
Geometry	3	1	7	7	2	4
Calculations	2	2	2	4	4	3
Statistics	1	2	3	5	2	3
Number	2	1	0	0	1	1

Table 7. Average percentage of questions not reached by National Curriculum area

	2017	2018	2019	2022	2023	Average % not reached
Ratio	4	2	4	3	6	4
Algebra	4	2	0	1	9	3
Fractions	3	2	1	2	3	2
Measurement	4	3	1	1	1	2
Geometry	0	0	3	4	0	2
Statistics	0	0	2	2	1	1
Calculations	1	1	1	1	1	1
Number	0	0	0	0	0	0

As Table 6 illustrates, in most domains, the omission rate is relatively low. The Ratio domain is the potential exception, with the average question covering this domain being omitted by 9% of pupils. This domain also has the highest "not reached" percentage at 4%, though clearly in absolute terms this is relatively small. Nevertheless, the higher omission and not reached percentages for Ratio questions likely reflect the fact that these questions tend to appear later in the test papers. At the other extreme sits the Number domain, where the omission (1%) and not reached (0%) rates are particularly low, reflecting the early position that these questions occupy on the tests.

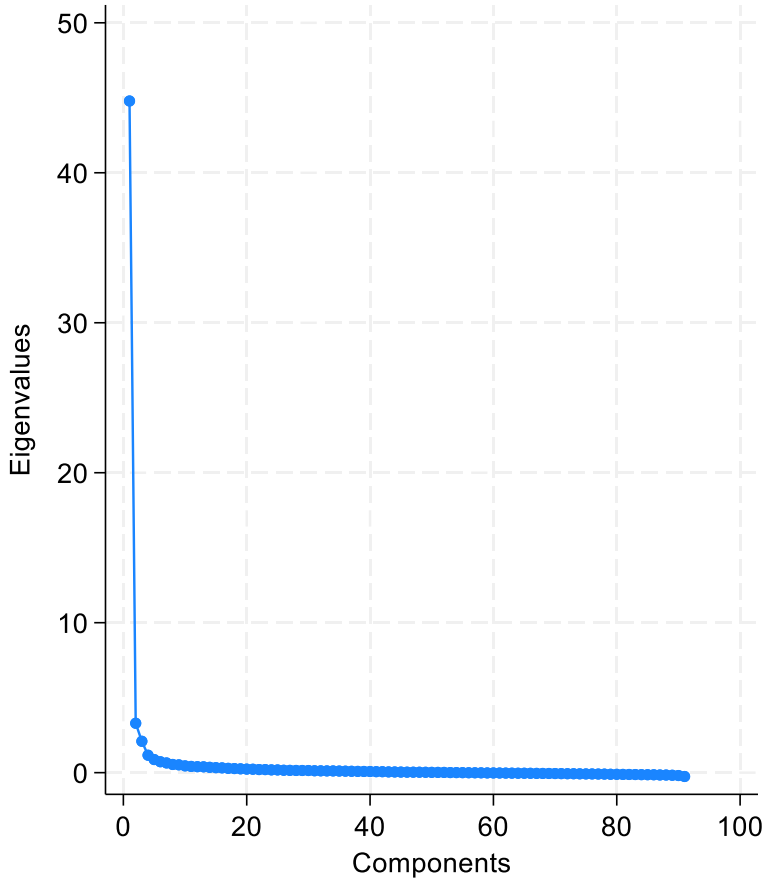
Otherwise, variation in the omission and not reached rates across the remaining domains is relatively limited. This does, however, again vary across years for domains covered by fewer questions within the test. For instance, in 2023, 10% of Algebra questions were omitted, with 9% of these being defined as "not reached". Yet the analogous figures in 2022 were just 4% and 1% respectively. This serves as additional motivation for pooling data and making inferences across more than one year; it will help to smooth such differences out.

Dimensionality of the Key Stage 2 mathematics test

Figure 1 presents the results following a factor analysis of the polychoric correlation between the test items. It provides a screeplot of the eigenvalues, which can be used to infer the "dimensionality" of the data. In other words, how many separate skills/attributes does the Key Stage 2 test seem to measure?

The key finding is that the Key Stage 2 mathematics test essentially appears to be unidimensional. This is illustrated by the very sharp decline between the first and second eigenvalue; the first eigenvalue clearly dominates all others and explains the vast amount of variation in the data. There is some weak evidence that a second and possibly third dimension may be captured within the data, though the evidence even for these is rather tentative. This provides a first indication that it may not be possible to measure eight separate latent constructs (i.e. the eight National Curriculum areas) within the Key Stage 2 mathematics test data.

Figure 1. Screeplot following an Exploratory Factor Analysis of the 2023 SATs item responses



Notes: Screeplot generated following factor analysis of the polychoric correlation between items.

Exploratory factor analysis

To conclude our initial investigations, we use exploratory factor analysis to create eight factors based on question responses for each year. Each question is then assigned to the factor it most strongly loads upon. Table 8 then illustrates the correspondence between the primary National Curriculum area each item is intended to measure and the factor to which it has been assigned. For brevity, we present results based on the 2023 assessment. Different colours in this table refer to the different National Curriculum areas. If there is a close correspondence between the National Curriculum domain and the results of the factor analysis, then the colours would sit close together.

The key conclusion from this table is that questions within the same National Curriculum area do not clearly cluster together within the same latent factor. This is illustrated by the

colours in the table (representing the different National Curriculum areas) being spread across the rows, with little evidence that they sit closely together. Take the Measure domain, for instance. The seven questions supposed to measure this area of the National Curriculum are spread across four separate factors (factors 1, 2, 4 and 7). Similarly, the Fractions questions (green shading) do not sit together and are widely spread across the eight factors (it is only factor 6, with only two questions, where none of the Fractions questions appear). These results thus indicate that questions assigned to the same primary National Curriculum domain do not empirically cluster closely together.

Table 8. Questions from 2023 by factor assigned

Factor	No. questions	Questions assigned to factor
Factor 1	18	M1_Q19_2023, M1_Q20_2023, M1_Q23_2023, M1_Q24_2023, M1_Q25_2023, M1_Q26_2023, M1_Q27_2023, M1_Q29_2023, M1_Q30_2023, M1_Q31_2023, M1_Q32_2023, M1_Q33_2023, M1_Q34_2023, M1_Q35_2023, M1_Q36_2023, M2_Q20_2023, M2_Q21_2023, M3_Q21_2023
Factor 2	27	M1_Q1_2023, M1_Q2_2023, M1_Q4_2023, M1_Q6_2023, M1_Q12_2023, M1_Q17_2023, M2_Q1_2023, M2_Q2_2023, M2_Q4_2023, M2_Q5_2023, M2_Q6_2023, M2_Q7_2023, M2_Q8_2023, M2_Q10_2023, M2_Q11A_2023, M2_Q11B_2023, M2_Q25A_2023, M3_Q1_2023, M3_Q2_2023, M3_Q3_2023, M3_Q4_2023, M3_Q5_2023, M3_Q6_2023, M3_Q10_2023, M3_Q11_2023, M3_Q14A_2023, M3_Q18A_2023
Factor 3	9	M1_Q3_2023, M1_Q5_2023, M1_Q7_2023, M1_Q8_2023, M1_Q9_2023, M1_Q10_2023, M1_Q11_2023, M1_Q13_2023, M1_Q15_2023
Factor 4	19	M2_Q3_2023, M2_Q9_2023, M2_Q12_2023, M2_Q14_2023, M2_Q15_2023, M2_Q17_2023, M2_Q22_2023, M2_Q25B_2023, M3_Q7_2023, M3_Q8_2023, M3_Q9_2023, M3_Q12_2023, M3_Q14B_2023, M3_Q16_2023, M3_Q17_2023, M3_Q18B_2023, M3_Q20_2023, M3_Q22_2023, M3_Q23_2023
Factor 5	3	M1_Q14_2023, M1_Q21_2023, M1_Q28_2023
Factor 6	2	M2_Q26A_2023, M2_Q26B_2023
Factor 7	9	M2_Q13_2023, M2_Q16_2023, M2_Q19_2023, M2_Q23_2023, M2_Q24_2023, M3_Q13_2023, M3_Q15A_2023, M3_Q15B_2023, M3_Q19_2023
Factor 8	4	M1_Q16_2023, M1_Q18_2023, M1_Q22_2023, M2_Q18_2023

Red = Algebra; Dark Blue = Calculations; Green = Fractions; Purple = Geometry; Orange = Measure; Black = Number; Grey = Ratio; Light Blue = Statistics.

4. Key Stage 2 sub-domain scores. School-level results.

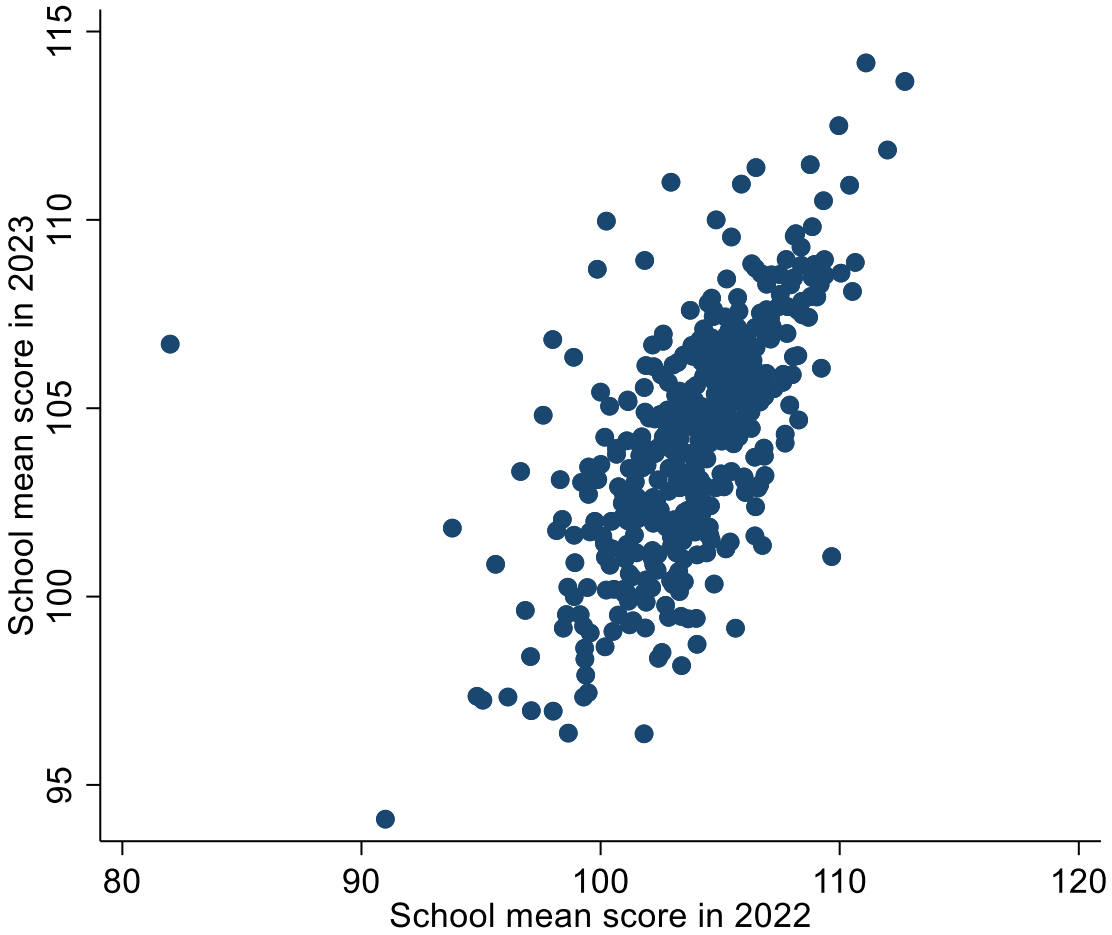
Our main approach to creating Key Stage 2 mathematics scores uses a MIRT latent regression approach. Fu and Qu (2018) review sub-score estimation methods and note how the use of MIRT is one of the most reliable – if computationally most challenging – approaches. This approach has broad similarities to the methodology underpinning International Large-Scale Assessments such as PISA and TIMSS, along with national studies such as the National Reference Test in England.

In this model, each National Curriculum area is treated as a separate latent construct, though each is correlated with another. The IRT structure is simple, with each item only loading on a single domain, with no cross-loadings of items (e.g. only Number items load onto the Number trait). In other words, from an IRT perspective, the measurement model is effectively formed of eight unidimensional sub-tests. From each model specification, we then produce estimates of pupils' abilities in each National Curriculum area.

4.1 How stable are school-level estimates of SATs sub-domain mathematics scores across years?

To begin, we consider the consistency of school-level sub-domain scores across academic years. In other words, to what extent do schools that get higher average scores in a given National Curriculum area (e.g. Statistics) in one year also get higher scores in that area the following year? For brevity, we also focus on the results comparing 2022 to 2023. For context, Figure 2 illustrates the school-level correlation in 2022 and 2023 official SATs scores for the 460 schools in our sample with more than 10 pupils. There is a moderate association, with a Pearson correlation of 0.64 and a Spearman correlation of 0.66. This indicates that some schools experience quite large changes in their official SATs mathematics scores between academic years.

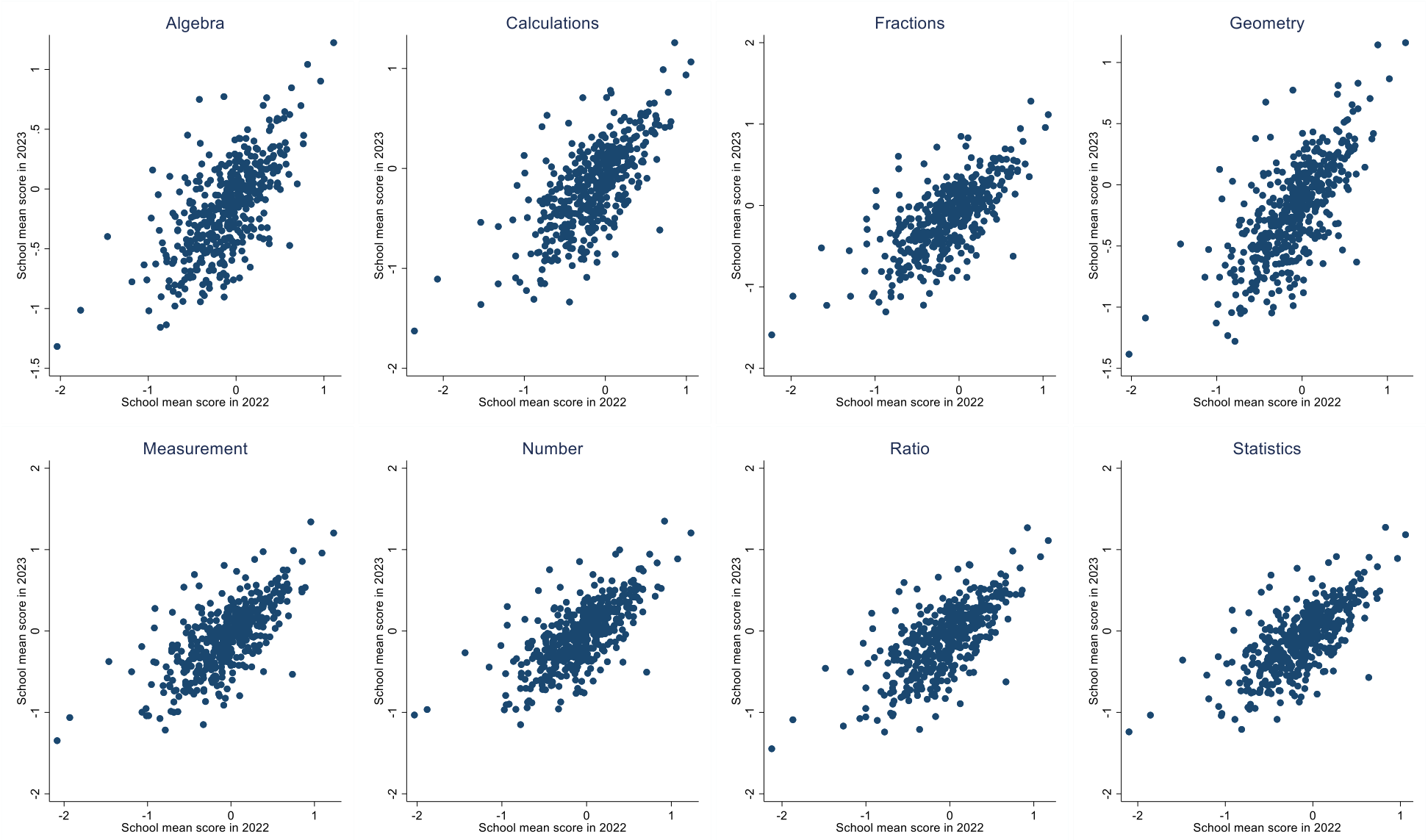
Figure 2. School average Key Stage 2 mathematics official scaled scores



Notes: Sample restricted to schools with at least 10 pupils with data available. Each dot in the graph refers to a single school. Figures on the horizontal axis capture schools' published SATs mathematics scores in 2022, with those on the vertical axis the scores in 2023. Pearson correlation is 0.64.

Figure 3 presents analogous scatterplots for each of the eight National Curriculum areas. Visually, each looks very similar to the pattern displayed when using the official SATs scores in Figure 1. All exhibit a moderate positive association, and do not appear to be any more or less noisy than the inter-year comparison of official SATs scores. This finding is supported by Table 9; the school-level correlation between the 2022 and 2023 official SATs results (0.64) is slightly lower than for our SATs scores in each of the eight National Curriculum areas (which range from 0.66 for Algebra to 0.69 for Number). This implies that the between-year stability in our school-level sub-domain mathematics scores is at least equal to the between-year stability in official school-level overall mathematics scores.

Figure 3. School-level correlations across National Curriculum areas over the years (2022 versus 2023).



Notes: Figures on the horizontal axis capture school-average sub-domain mathematics scores in 2022, with those on the vertical axis the analogous scores from 2023. See Table 3 for the associated correlation coefficients.

Table 9. The correlation in school-average Key Stage 2 mathematics scores across years (2022/2023)

Domain	Pearson	Spearman
Number	0.69	0.68
Statistics	0.69	0.68
Measurements	0.69	0.68
Geometry	0.69	0.68
Fractions	0.69	0.67
Calculations	0.68	0.66
Ratio	0.68	0.66
Algebra	0.66	0.64
Overall	0.69	0.67
Official scaled scores	0.64	0.66

Notes: Overall refers to estimates from a unidimensional IRT model that we have estimated. Official scaled scores refer to the school averages using the final scaled scores reported by the Department for Education. Analysis based on data from 500 schools.

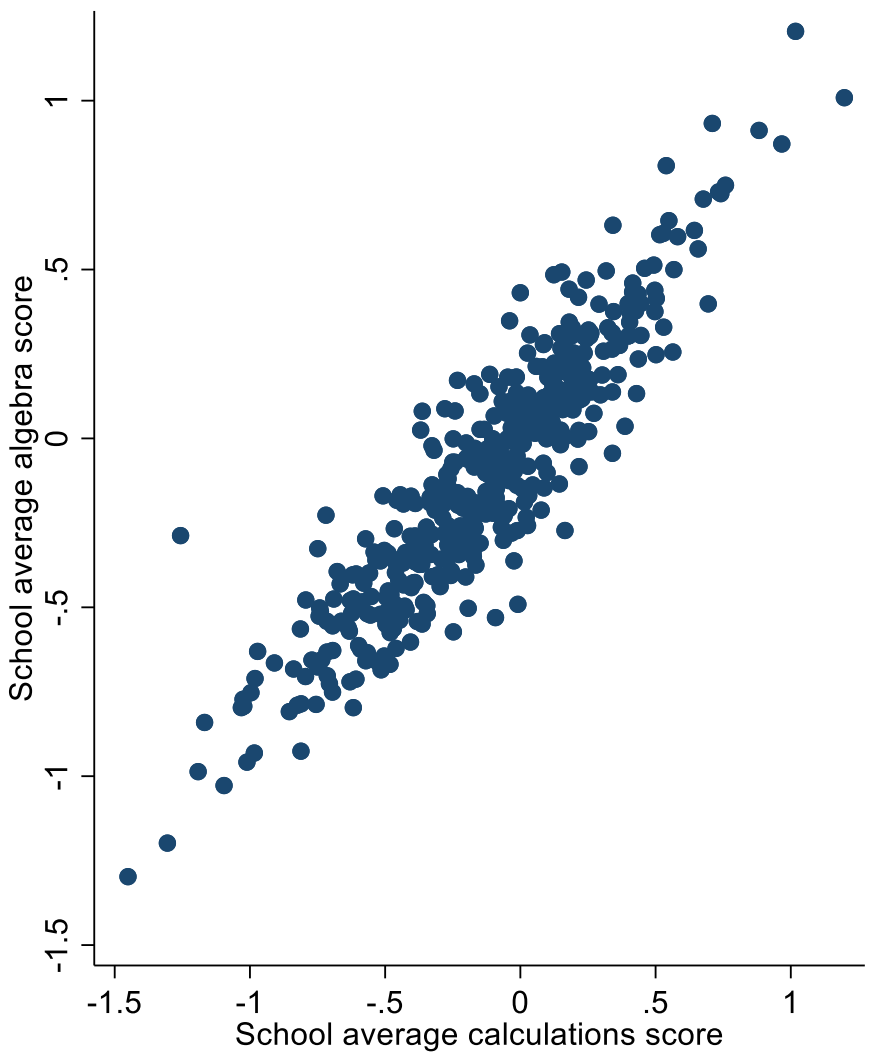
To what extent do schools perform differently across the eight mathematics sub-domains?

In the next stage of our analysis, we consider the school-level correlation across the different National Curriculum areas. In other words, to what extent do schools that perform well in one National Curriculum area (e.g. Algebra) also perform well in other areas (e.g. Statistics, Ratio, Number)? This is important as, if these correlations are very strong, it indicates that schools perform very similarly across different areas of the mathematics curriculum (meaning sub-domain scores may be of limited value) or that the test cannot differentiate sufficiently well across different areas. The strength of the correlation is reported in Table 10, based on the Key Stage 2 SATs data from 2023.

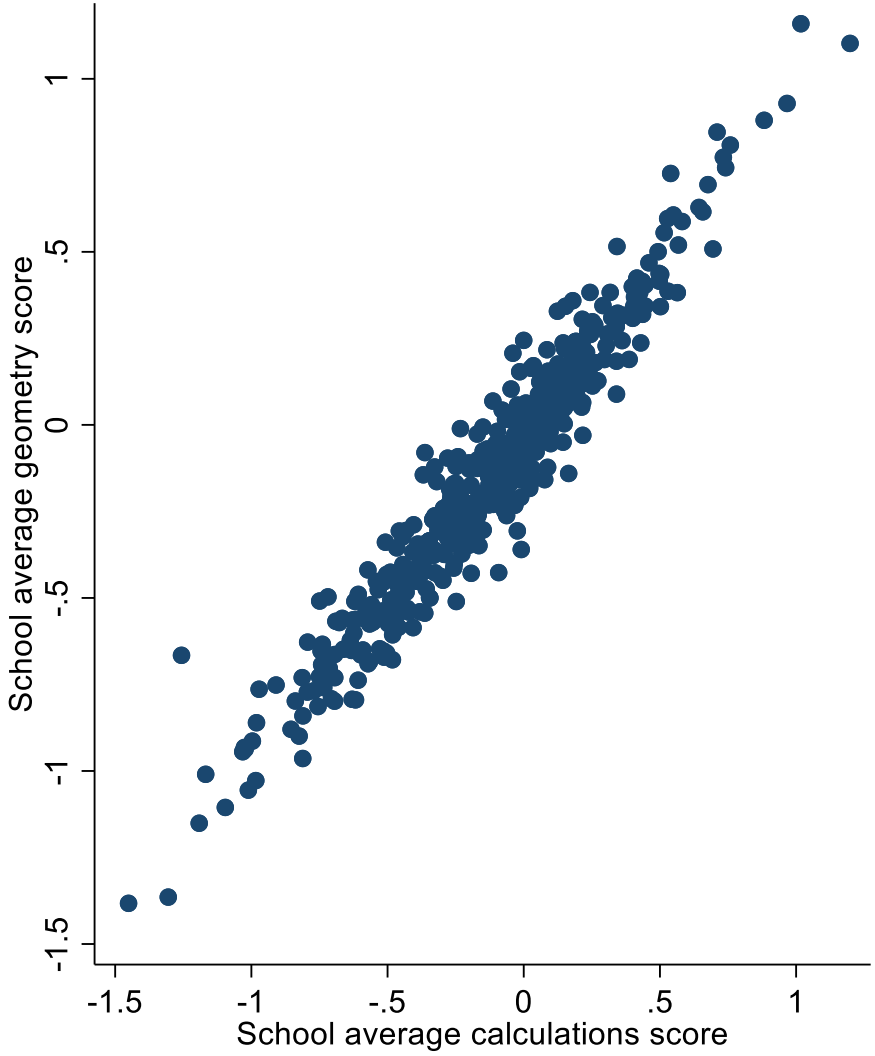
The school-level correlations for scores across the different National Curriculum areas are very strong. With the exception of the Algebra domain, the correlations stand at 0.97 or above. A visual representation of two of the weakest correlations is presented in Figure 4. This demonstrates just how similarly schools perform across the different parts of the mathematics curriculum in the SATs (particularly given that these represent some of the weakest correlations observed). Consequently, while the analysis presented above suggested that our approach to estimating sub-domain SATs scores for each school is feasible, the results presented in Table 10 and Figure 4 suggest that such information may be of only limited value to schools. Specifically, the variation in schools' results across the different sub-domains is likely to be too small to be of much substantive use.

Figure 4. Scatterplots for the school-level correlations across domains

(a) Algebra versus Calculations



(b) Geometry versus Calculations



Notes: School-level averages based on the model without latent regressors. Correlation is 0.92 in the left-hand panel and 0.97 in the right-hand panel.

Table 10. School-level correlation in average scores across the various National Curriculum areas

	Algebra	Calculations	Fractions	Geometry	Measurement	Number	Ratio	Statistics
Algebra								
Calculations	0.92							
Fractions	0.93	1.00						
Geometry	0.99	0.97	0.97					
Measurement	0.94	0.99	0.99	0.98				
Number	0.95	0.98	0.98	0.99	1.00			
Ratio	0.92	1.00	1.00	0.97	1.00	0.99		
Statistics	0.95	0.99	0.99	0.99	1.00	0.99	1.00	

Notes: Scatterplots for the cells in green can be found below. Based on a multi-dimensional IRT latent regression model that does not include school fixed effects. Analysis based on our sample of 500 schools where data was available for at least 10 pupils each year.

To provide further context to this result, Table 11 draws on data from the TIMSS 2023 4th grade (Year 5) data for England. Do we also observe strong school-level correlations across different areas of mathematics in this test of similarly aged primary school pupils?

Table 11. School-level correlations across TIMSS mathematics domains

	Overall	Number	Geometry	Data
Overall	-	-	-	-
Number	0.993	-	-	-
Geometry	0.989	0.986	-	-
Data	0.988	0.984	0.983	-

Notes: Analysis based on TIMSS grade 4 (Year 5) data for England from 2023.

The answer is yes. In TIMSS, the correlation between schools' results in Geometry and Number domains is 0.986, very similar to the correlation of 0.99 we observe in the SATs. Likewise, the correlation between Data and Geometry domains in TIMSS is 0.983, very close to the 0.99 correlation between Geometry and Statistics domains in the SATs. This provides a degree of reassurance that the results we have obtained using our approach produce broadly similar school-level associations in another independent dataset.

4.2 What are the benefits of pooling together schools' results from across more than one year?

Pooling data together across years offers two important advantages. First, sample sizes for individual schools will roughly increase. This will lead to smaller standard errors for each school, thus enhancing the precision of school-level results, as observed in Table 12. After pooling the data across two years, the average standard error surrounding schools' results falls by around a quarter, from around 0.17 to 0.13. Equivalently, the

standard error falls from around 40-45% of the between-school standard deviation down to around 30%. There are hence tangible benefits to presenting school-level Key Stage 2 mathematics scores by pooling results from more than one year. This is, however, true for both our sub-domain mathematics SATs scores and the official SATs scores (i.e. those produced by the Department for Education).

Table 12. The standard error surrounding school average Key Stage 2 mathematics scores. Results combining data from 2022 and 2023

Outcome	SD of scores across schools	Average school standard error	standard error as a % of the standard deviation across schools
Algebra	0.38	0.13	34%
Calculations	0.44	0.13	30%
Fractions	0.44	0.13	30%
Geometry	0.41	0.13	31%
Measurement	0.42	0.13	31%
Number	0.41	0.13	32%
Ratio	0.43	0.13	31%
Statistics	0.40	0.13	31%
Overall scores	0.44	0.13	30%

Second, given the limited number of test questions in some domains within a single year, pooling data across years helps to broaden construct coverage and reduce random differences that occur due to, for instance, the percentage of omitted or not reached questions that occur within a single year. Pooling data from across multiple years can thus help to smooth such fluctuations.

5. Differences in mathematics sub-domain scores across demographic groups

We now turn to our analysis of gaps in achievement between demographic groups across different parts of the Key Stage 2 mathematics curriculum. These results are presented in Table 13. As the standard deviation of scores in each National Curriculum area is around one, figures can be broadly interpreted in terms of effect size differences. To aid interpretation of results, one can multiply these values by 37 to convert the differences into percentile ranks (i.e. an effect size difference of 0.1 is roughly equivalent to a difference of four places in a ranking of 100 pupils) – see von Hippel (2024) for further details. These results are based on data pooled across 2018 to 2023.

Table 13. Gaps between demographic groups in performance on different components of the Key Stage 2 mathematics curriculum. Pooled data from 2018 to 2023.

	FSM		Gender		SEN support		EHCP		EAL		Summer	
	Gap	SE	Gap	SE	Gap	SE	Gap	SE	Gap	SE	Gap	SE
Algebra	-0.50	0.01	-0.17	0.01	-0.85	0.01	-0.99	0.02	0.17	0.01	-0.14	0.01
Calculations	-0.55	0.01	-0.15	0.01	-0.99	0.01	-1.20	0.02	0.23	0.01	-0.15	0.01
Fractions	-0.55	0.01	-0.15	0.01	-0.98	0.01	-1.19	0.02	0.24	0.01	-0.15	0.01
Geometry	-0.53	0.01	-0.17	0.01	-0.91	0.01	-1.07	0.02	0.19	0.01	-0.15	0.01
Measurement	-0.53	0.01	-0.18	0.01	-0.93	0.01	-1.10	0.02	0.19	0.01	-0.15	0.01
Number	-0.54	0.01	-0.18	0.01	-0.95	0.01	-1.12	0.02	0.19	0.01	-0.15	0.01
Ratio	-0.54	0.01	-0.17	0.01	-0.97	0.01	-1.16	0.02	0.22	0.01	-0.15	0.01
Statistics	-0.53	0.01	-0.17	0.01	-0.93	0.01	-1.10	0.02	0.20	0.01	-0.15	0.01
Overall	-0.57	0.01	-0.17	0.01	-1.01	0.01	-1.21	0.02	0.23	0.01	-0.26	0.01

Notes: Overall refers to estimates from a unidimensional model. EAL = English as an Additional Language. FSM = Free School Meals. SEN = Special Educational Needs. EHCP = Education, Health and Care Plan.

Starting with the results by FSM eligibility, there is a sizeable gap in each of the mathematics sub-domains as well as overall mathematics scores. However, on the whole, the magnitude of the gap is very similar across the different areas of mathematics. With the exception of Statistics and Algebra domains, we estimate the FSM gap in all other domains to be between 0.53 and 0.57 standard deviations – very similar to the FSM gap in overall scores. Moreover, most pairwise comparisons of FSM gaps across sub-domains are not statistically significant at conventional thresholds. The only partial exception is the Algebra domain, where the FSM gap is 0.49, which is slightly smaller than in other areas such as Calculations, Fractions and Measurement. Nevertheless, overall, the magnitudes of FSM gaps are broadly similar across different parts of the Key Stage 2 mathematics curriculum.

The next column turns to gender differences in performance across the mathematics sub-domains. On average, boys tend to outperform girls in each National Curriculum area of mathematics, as well as overall scores. The magnitudes of the differences are, however, relatively small, standing between 0.15 and 0.2 standard deviations. This is consistent with Coates (2025), who also reports a small gain in favour of boys in Key Stage 2 mathematics performance. Again, in terms of individual mathematics domains, most differences are relatively small.

For gender, it is also possible to compare results to analogous results from TIMSS. Therefore, in Table 14, we present the gender gap in grade 4 (year 5) TIMSS mathematics

overall and by sub-domain. Note that 100 points on the TIMSS scale is roughly equivalent to a one international standard deviation change.

Table 14. Gender differences in TIMSS mathematics domain scores

	Gender gap	Standard error
Number	17.6	3.86
Geometry	17.9	4.47
Data	13.4	4.16
Overall	18.3	3.51

Notes: Author's calculations based on analysis of the TIMSS 2023 grade 4 (Year 5) data for England.

The results from TIMSS also point towards a gender gap, with higher scores for boys. For instance, based on our overall Key Stage 2 mathematics scores, we find that boys achieve scores around 0.17 standard deviations higher than girls. The equivalent difference in TIMSS is around 18 test score points – equivalent to roughly 0.18 standard deviations. Moreover, from TIMSS, we see the gender gap to be very similar across the Number and Geometry domains, similar to our results from the SATs test. The gender gap is, however, smaller in the Data domain in TIMSS than for Geometry and Number domains, which is not the case with respect to the Statistics domain in Key Stage 2. Hence, while TIMSS and Key Stage 2 SATs are consistent in showing boys outperforming girls in each mathematics domain, there are some modest differences in the nuances lying behind these broad patterns.

Returning to Table 13, the next two columns focus on comparisons between pupils with and without special educational needs. EHCP refers to the difference between those with Education, Health and Care plans and other pupils[2]. The final column presents analogous differences for those with and without special educational needs (SEN) support.

[2] Only a subset of pupils with EHCPs take the Key Stage 2 tests; those who do not are typically the pupils with the most severe special educational needs. Our estimates are consequently likely to underestimate the true EHCP gap in achievement.

As expected, both sets of results illustrate how pupils with special needs obtain substantially lower scores across all National Curriculum areas than their peers without such needs. Gaps are around one standard deviation between those with and without SEN support, and 1.1 to 1.2 standard deviations for those with EHCPs compared to those without a special educational need. Moreover, for both groups, the gap in scores appears to be smaller in Algebra, Geometry and Statistics domains than in most other areas.

Finally, the last column of Table 13 presents the gap between summer-born (i.e. the youngest) and older pupils. The youngest pupils tend to achieve lower scores on the Key Stage 2 SATs mathematics test. There is, however, little evidence of any meaningful difference across the various sub-domains.

6. Recommendations for policy and practice

The findings of this report have important implications for the way that performance data from Key Stage 2 assessments are reported back to schools. At present, the quantity of information provided risks schools reaching erroneous conclusions regarding their strengths and weaknesses across different areas of the mathematics curriculum. The provision of such information without adequate articulation of uncertainty places an unreasonable burden on schools and increases the likelihood of misinterpretation of the data provided.

To address these issues, we set out five recommendations for the Department for Education regarding the reporting of Key Stage 2 sub-domain scores back to schools.

1. Discontinue the provision of Key Stage 2 sub-domain raw scores.

The Department for Education should stop providing schools with sub-domain raw scores within its Analyse School Performance tool. These figures are presented without any accompanying indication of uncertainty, making them prone to misinterpretation. School leaders and staff are not experts in the statistical nuances of such data, and the provision of raw, unqualified results carries a serious risk of schools reaching erroneous and potentially harmful conclusions.

2. Prioritise fewer, higher-quality indicators.

The Department for Education should more broadly review the type and quantity of information that is fed back to schools. In this context, less is often more. Making many pieces of information available – such as sub-domain scores – risks diverting schools' attention and resources. Schools would be better served by the provision of a smaller number of carefully selected, robust pieces of information that they can act upon with confidence. The sub-domain scores provide a clear example where the availability of more data is counterproductive, creating noise rather than clarity.

3. Reform sub-domain reporting where demand exists.

If, due to user demand, the Department for Education chooses to continue reporting sub-domain scores within the Analyse School Performance tool, these must be constructed using a more robust methodology. The current practice of reporting raw scores without measures of uncertainty is inadequate and misleading. The methodology set out in this report offers a more principled approach. Explicitly accounting for measurement error enables schools to understand their comparative position across curriculum areas in a way that is both reliable and informative. It would thereby reduce the risk of schools making misinformed decisions.

4. Redesign Key Stage 2 tests if they are to provide diagnostic information.

If the Department for Education wishes Key Stage 2 assessments to serve a diagnostic purpose, then the tests themselves need to be fundamentally redesigned. In their current form, the assessments are not fit to inform schools of their (and their pupils') relative strengths and weaknesses. Attempting to repurpose them as diagnostic tools without structural change is unwise and risks generating misleading conclusions about pupil attainment in specific curriculum areas. A redesigned assessment would require a stronger balance of questions across domains, as well as items that more effectively discriminate between pupils' skills across different areas of the National Curriculum.

5. Base Key Stage 2 accountability on multi-year averages.

Beyond the issue of sub-domain reporting, our analysis also highlights the wider problem of volatility in Key Stage 2 results at the school level. Year-on-year variation is substantial, driven largely by the small size of primary school cohorts (e.g. on average, there were 42

Year 6 pupils per school in 2024). Therefore, reliance on single-cohort results is inappropriate for accountability purposes. The Department should reform the accountability framework so that school-level performance is assessed using multi-year averages. This would provide a more stable and accurate measure of school performance and reduce distortions arising from school cohort sizes. For further discussion of this issue, see Menzies and Jerrim (2021).

References

Allen, R., & Burgess, S. (2011). Can school league tables help parents choose schools? *Fiscal Studies*, *32*(2), 245–261. http://www.jstor.org/stable/24440204

Allen, R. (2021). The limited uses of question level analysis. Retrieved from https://rebeccaallen.co.uk/2021/12/29/the-limited-uses-of-question-level-analysis/

Benton, T. (2012). *Calculating the number of marks needed in a subtest to*

make reporting subscores worthwhile. Cambridge Assessment Research Report.

Cambridge, UK: Cambridge Assessment.

Coates, A. (2025). Tracking mathematics achievement gaps in England: Gender, socioeconomic status and ethnicity. *British Educational Research Journal*.

Department for Education. (2025). National curriculum test development handbook. https://www.gov.uk/government/publications/national-curriculum-test-development-handbook

Fu, J. and Qu, Y. (2018). A Review of Subscore Estimation Methods. *ETS Research Report Series*: 1-15. https://doi.org/10.1002/ets2.12203

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229. https://www.jstor.org/stable/20172113

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79-95. https://doi.org/10.1348/000711007X248875

Jewsbury, P. A., Jia, Y., & Gonzalez, E. J. (2024). Considerations for the use of plausible values in large-scale assessments. *Large-scale Assessments in Education*, 12, 24. https://doi.org/10.1186/s40536-024-00213-y

Menzies, L & Jerrim, J. (2021). Improving headline school performance measures. Retrieved from https://cfey.org/wp-content/uploads/2020/07/Improving-headline-school-performance-measures.-Multi-year-averages-should-be-used-in-school-league-tables.-Menzies-and-Jerrim-2020-1.pdf

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

OECD. (2023). *Education at a Glance 2023: OECD Indicators*, OECD Publishing, Paris. https://doi.org/10.1787/e13bef63-en.

Ofqual. (2024). National assessments regulation annual report 2024. Retrieved from https://dera.ioe.ac.uk/id/eprint/40972/1/National%20assessments%20regulation%20annual%20report%202024%20-%20GOV.pdf

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008). Comparison of subscores based on classical test theory. ETS Research Report ETS RR-08-54. Retrieved from https://files.eric.ed.gov/fulltext/EJ1111369.pdf

Robitzsch, A.; Kiefer, T. & Wu, M. (2024). *TAM: Test Analysis Modules*. R package version 4.2-21, https://CRAN.R-project.org/package=TAM.

Royal, K. D., & Hedgpeth, M. W. (2018). Think Subscores Are a Helpful Form of Feedback? Think Again. *Journal of veterinary medical education*, *45*(4), 567–570. https://doi.org/10.3138/jvme.0117-014r1

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, *80*(3), 727–747. https://doi.org/10.1007/s11336-014-9415-z

Sibieta, L. (2024). The latest picture on school funding and costs in England [Comment]. *Institute for Fiscal Studies*. Available at: https://ifs.org.uk/articles/latest-picture-school-funding-and-costs-england

Standards and Testing Agency. (2025). National curriculum test development handbook. Retrieved from https://www.gov.uk/government/publications/national-curriculum-test-development-handbook

Teacher Tapp. (2024). Low-level disruption, parent - school relationship and 'no go' admin tasks. Retrieved from https://teachertapp.com/uk/articles/low-level-disruption-parent-school-relationship-and-how-everyone-got-to-school-on-friday/

von Hippel, P. T. (2024). Multiply by 37 (or divide by 0.027): A surprisingly accurate rule of thumb for converting effect sizes from standard deviations to percentile points. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/01623737241239677

Ward, H. (2017). Sats costs revealed: £44m in first year of new system. Retrieved from https://www.tes.com/magazine/archive/sats-costs-revealed-ps44m-first-year-new-system

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114-128. https://doi.org/10.1016/j.stueduc.2005.05.0

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer Singapore

Wu, M.; Ping Tam, H. & Jen, T-H. (2016). *Educational Measurement for Applied Researchers. Theory into Practice*.