Smartgrade prediction accuracy: primary

1. Introduction

Smartgrade offers a "mock" KS2 and KS4 exam service. In addition to question- and topic-level analysis, these exams generate predicted grades. Smartgrade commissioned this analysis to evaluate the accuracy of those predictions.

Examination data was shared from one of its large MAT partners, allowing us to analyse predicted scaled scores and performance indicators from Autumn 1, Autumn 2 and Spring 2 mock examinations, compared to actual outcomes. End of Year 5 Reading performance indicator predictions from HeadStart Reading assessments are also included – but scaled score indicators were not supplied for HeadStart because at the time Smartgrade did not calculate these for HeadStart assessments.

This report summarises the findings of this analysis for Smartgrade.

2. Methodology

Data was supplied via Google Drive by Smartgrade. The following data cleaning steps were taken:

1. Standardise terminology between files (e.g. subject naming and year group conventions) to ensure consistency across datasets
2. Match datasets at pupil level
3. Derive additional calculations (performance indicators from scaled scores; variance between predicted and actual grades)
4. Remove duplicate predictions
5. Cross-check manually against source files to ensure no errors were introduced

Overall, match rates between pupils in the predicted grades and final outcome files were strong, producing healthy sample sizes:

Time period	N (scaled scores)	N (matched to outcomes)	Percentage matched
Autumn 1	4206	4007	95.27%
Autumn 2	4204	4064	96.67%
Spring 2	4232	4077	96.34%

3. Analysis findings

3.1) Overall prediction accuracy

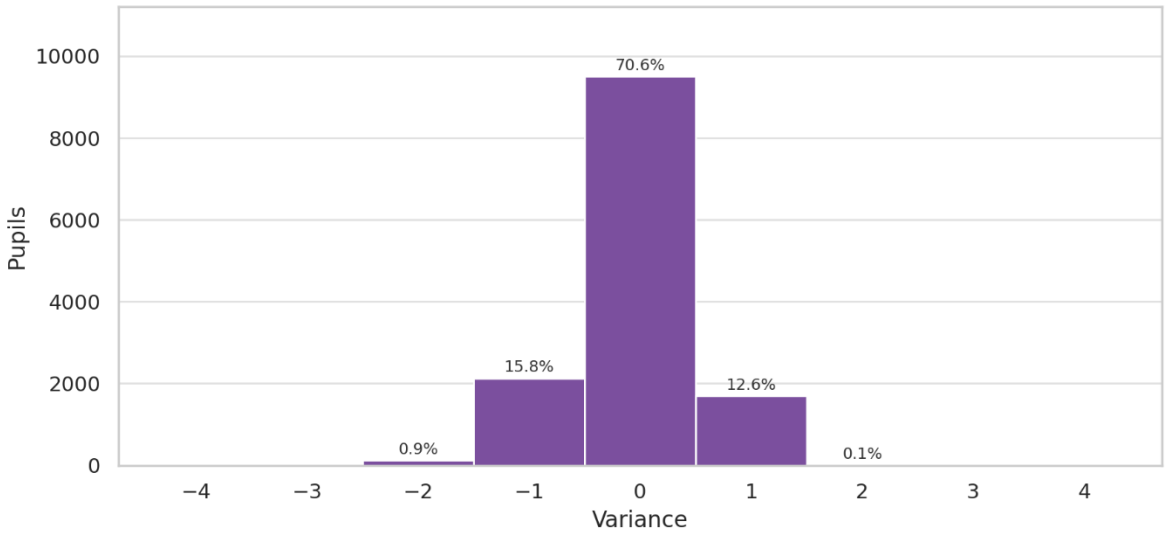
At performance indicator level, prediction accuracy is consistently strong:

- Across the assessment windows analysed, exact match rates – where the predicted indicator matches the achieved indicator – range from 61.4% to 73.6%
- Almost all pupils are placed either exactly correctly or very close: 97.9% to 99.4% are within one band of what was predicted
- Mean absolute error (MAE – the average distance between the predicted and actual band) is around 0.32 of a band, suggesting that when the model misses, it typically does so narrowly

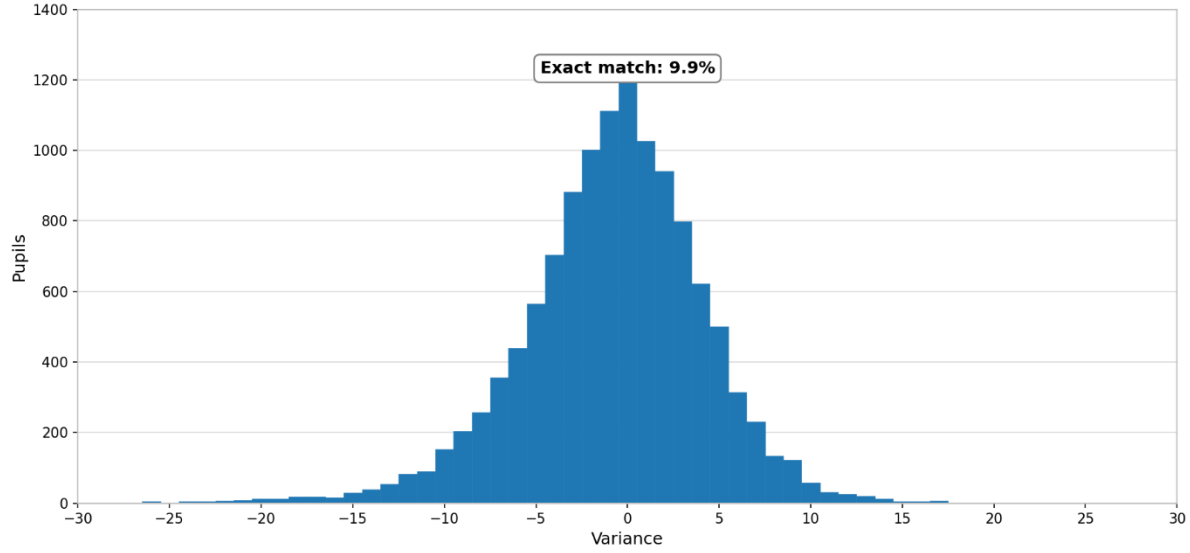
At scaled score level, predictions are naturally less precise. But accuracy is also strong:

- Across assessment windows, MAE is 3.78 points – meaning a typical prediction is around 4 points from the final score – and 57.3% of pupils are predicted within three points.
- Smartgrade predictions are slightly conservative but generally well calibrated: the average predicted score is 0.8 points below the average actual outcome. The higher MAE is driven primarily by individual pupil-level variation, rather than by systematic over- or under-prediction.
- The correlation between predicted and actual scaled scores is strong (r = 0.82). By comparison, EEF analysis reports typical correlations of 0.7–0.8 for mathematics and 0.6–0.7 for reading across most standardised tests.

Performance indicator headline accuracy – all windows, all subjects



Scaled score headline accuracy – all windows, all subjects



3.2) Breakdown by performance indicator

This section focuses on performance indicator accuracy. The key metrics are exact match rate (how often the prediction is the same band), within 1 band rate (how often it falls close) and % achieving their prediction or higher.

Indicator accuracy by assessment window

Timepoint	n	Exact match (%)	Within 1 band (%)	Predicted or higher (%)	MAE
Headstart	1311	61.4	97.9	90.1	0.41
Autumn 1	4007	69.1	99.0	87.4	0.32
Autumn 2	4064	72.2	98.9	84.8	0.29
Spring 2	4077	73.6	99.4	89.1	0.27
Total	13,459	70.6	99.0	87.3	0.30

As might be expected, exact match rates increase significantly the closer to the date of pupils' final exams. The "Within 1 band" rate is consistently high throughout, suggesting that this may be useful as a rough gauge of achievement at least a year ahead.

Notably, this pattern of increasing accuracy closer to the examination date is not replicated in EEF's research, which found no major associations between timing and predictive accuracy for other standardised assessments.

Indicator accuracy by assessment subject

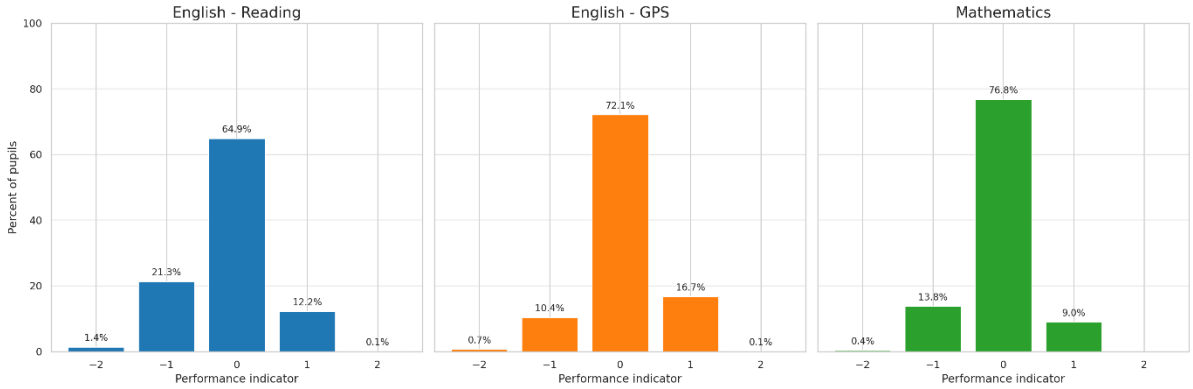
Subject	Exact match (%)	Within 1 band (%)	Predicted or higher (%)	MAE	Mean variance
Mathematics	76.8	99.6	91.1	0.24	-0.06
English - GPS	72.1	99.2	83.2	0.29	0.05
English - Reading	64.9	98.5	87.6	0.37	-0.12

Prediction accuracy is highest in Maths and lowest in Reading. In Maths, about three in four predictions hit the exact band; in Reading it's closer to two in three. When Reading misses, it also tends to miss by a slightly larger amount (reflected in the MAE).

A full breakdown by assessment window and subject is shown below:

Timepoint	Subject	n	Exact match	Predicted or higher	Within 1 band	MAE
Headstart	Reading	1311	61.4	90.1	97.9	0.41
Autumn 1	Mathematics	1348	72.7	92.3	99.2	0.28
Autumn 1	GPS	1311	70.6	82.5	99.5	0.3
Autumn 1	Reading	1348	63.9	87.1	98.4	0.38
Autumn 2	Mathematics	1340	77.4	87.5	99.5	0.23
Autumn 2	GPS	1358	71.0	80.1	98.6	0.3
Autumn 2	Reading	1366	68.3	86.7	98.8	0.33
Spring 2	Mathematics	1371	80.2	93.3	100	0.2
Spring 2	GPS	1335	74.6	87	99.5	0.26
Spring 2	Reading	1371	66.0	86.9	98.8	0.35

All time distributions by subject



3.3) Breakdown by scaled score

This section focuses on scaled score accuracy. As above, we report MAE – the typical size of error in scaled-score points, regardless of direction. Pearson's R complements MAE by measuring whether predicted and actual scores move together across pupils. The other key metric is the percentage of predictions falling within 3 scaled score points, and the percentage of results that achieved predicted grade or higher.

Scaled score accuracy by assessment window

Timepoint	n	MAE	Mean variance	Predicted or higher (%)	Within 3 points (%)	Pearson's R
Autumn 1	4007	4.05	-0.87	59.9	54.5	0.8
Autumn 2	4064	3.68	-0.37	55.1	58.2	0.83
Spring 2	4077	3.61	-1.18	64.9	59.2	0.85
Total	12,148	3.78	-0.81	60.0	57.3	0.82

This is consistent with the performance indicator analysis above: correlation between predicted grade and the outcome is strong, with accuracy increasing nearer to the point of final assessment. Pearson's R compares favourably to external benchmarks, with the average of 0.82 being above EEF's analysis of standardised assessments.

Scaled score accuracy by subject

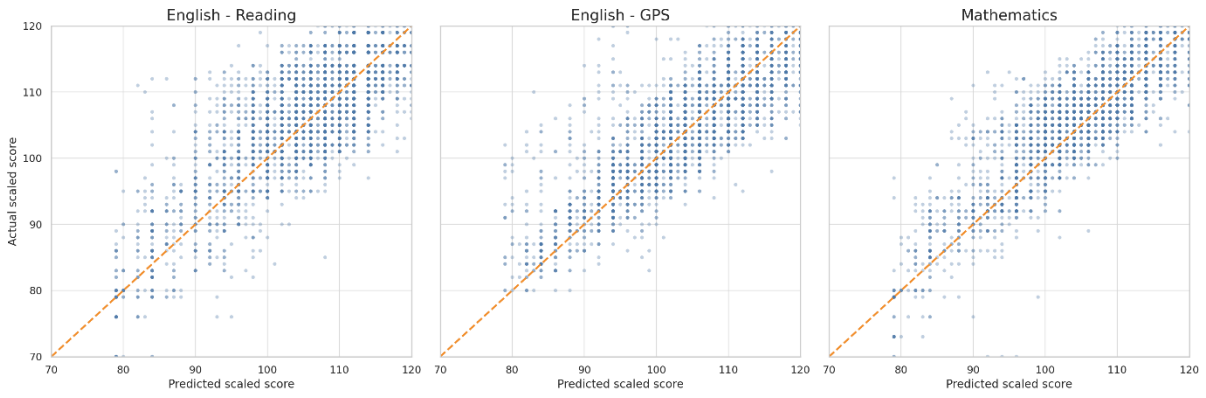
Subject	MAE	Mean variance	Predicted or higher (%)	Within 3 points (%)	Pearson's R
Mathematics	3.30	-1.15	65.2	63.5%	0.86
English - GPS	3.63	-0.16	54.1	60.1%	0.84
English - Reading	4.39	-1.09	60.6	48.4%	0.78

Prediction accuracy is highest in Maths and lowest in Reading. Maths typical error is 3.3 points and Reading typical error is 4.4 points. The "close prediction" rate is much lower in Reading: 64% within 3 points for Maths; 48% for Reading. The correlation is also lower in Reading. Maths and Reading were slightly more likely to underpredict than GPS.

A full breakdown by assessment window and subject is shown below:

Timepoint	Subject	n	MAE	Mean variance	Predicted or higher (%)	Within 3 points (%)	Pearson's R
Autumn 1	Mathematics	1348	3.76	-1.62	68.4	58.8	0.83
Autumn 1	GPS	1311	3.79	+0.11	51.2	57.4	0.83
Autumn 1	Reading	1348	4.58	-1.07	59.8	47.5	0.75
Autumn 2	Mathematics	1340	3.15	-0.49	56.7	65.2	0.87
Autumn 2	GPS	1358	3.63	+0.36	48.6	59.9	0.84
Autumn 2	Reading	1366	4.24	-0.98	60	49.5	0.80
Spring 2	Mathematics	1371	3.0	-1.34	70.4	66.4	0.89
Spring 2	GPS	1335	3.47	-0.96	62.4	62.8	0.86
Spring 2	Reading	1371	4.36	-1.22	61.9	48.4	0.79

All time distributions by subject



3.4) Other trends of note

This section briefly explores other areas of interest.

Demographic analysis

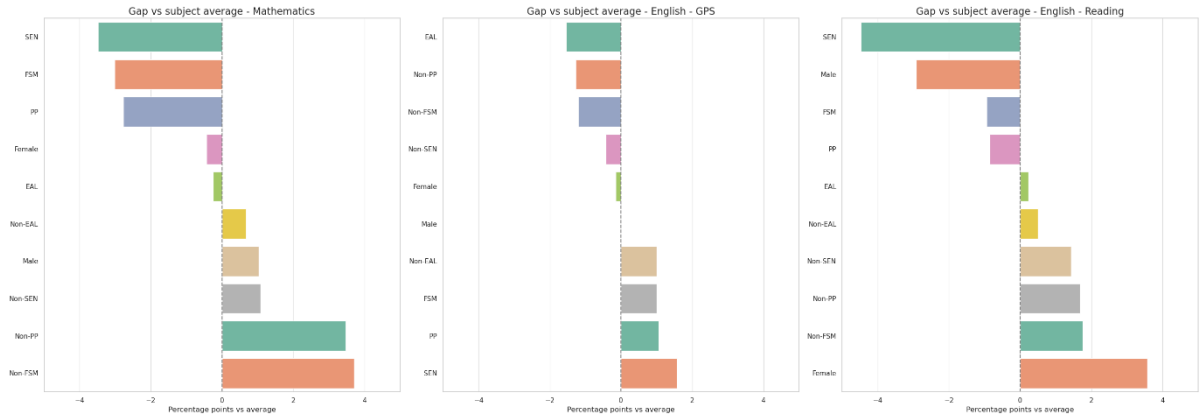
Demographic data was supplied for each pupil, enabling analysis of whether prediction accuracy varies by pupil group. Given limitations with sample size:

- LAC is removed entirely from this analysis
- This represents pooled results across all assessment windows

This analysis uses scaled scores as the primary means of reporting variation, given these are more sensitive to small differences. The percentage of predictions within 3 scaled score points is used as the primary measure; other measures could be added with further analysis.

Percentage correct to within 3 scaled score points by demographic group

The first thing to note is that demographic variance is generally relatively small. Accuracy tends to be higher for less disadvantaged groups (e.g. non-PP accuracy is higher than PP). Disparities in demographic prediction are also wider in Reading – this is particularly notable for SEN (43.6% vs 49.5% non-SEN) and gender (lower accuracy for boys). Below we show the same data plotted by extent of variation against the average:

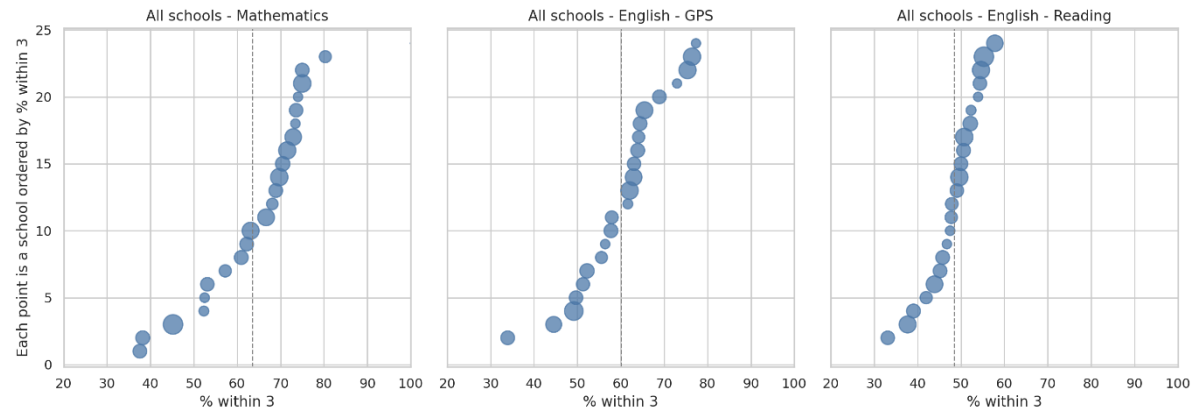


The practical takeaway is that some work could be done to reduce variation by demographic group, but these variations are relatively small.

School-level analysis

We also examined variation in prediction accuracy at school level. This analysis is exploratory, as small cohort sizes make it difficult to draw firm inferences. However, if consistent trends emerged whereby predictions in some schools were more accurate than others, this could point to areas such as improving the consistency of assessment conditions. Here is what the data looks like:

Percentage correct to within 3 scaled score points by school



In practical terms, there is considerable variation in the proportion of predictions falling within 3 scaled score points – to be expected given sample size. Across the subjects, no school appears in the top quintile across all three subjects, and none in the bottom quintile across all three – see the highest variation schools below. This largely rules out edge cases of a given school being consistently low or high in terms of predictive accuracy.

Initial statistical analysis does however suggest that school-level variation may not be entirely noise – for example, correlation analysis shows that there is some relationship between Maths and GPS in terms of a school being likely to show over or underprediction. This would require more in-depth analysis to dig into further.

4. Suggested recommendations and next steps

The analysis presented in this report indicates that Smartgrade's prediction model is performing well at primary level. The following recommendations are intended to help Smartgrade build on these strengths and address the areas where accuracy could be improved further.

4.1) Build on the strength of prediction accuracy

General prediction accuracy is strong: 70.6% of performance indicator predictions are an exact match, 99.0% fall within one band, and Pearson's R of 0.82 compares favourably to EEF benchmarks of 0.6–0.8 for standardised tests. Accuracy improves meaningfully across assessment windows, from 61.4% exact match at HeadStart to 73.6% at Spring 2. Clearly the existing prediction model is working well in a number of aspects, and can be used as a basis for further iterations and improvements.

4.2) Investigate the Reading accuracy gap

Across all primary assessment windows, Reading predictions are consistently less accurate than Mathematics and GPS, with a lower exact match rate (64.9% vs 76.8% for Maths), a higher MAE, and a weaker correlation. This gap persists even as predictions improve over time for all subjects. It would be worth investigating whether this reflects something inherent to the variability of reading performance at KS2, or whether there are model-level adjustments that could narrow the gap.

4.3) Review demographic and school-level prediction patterns

While demographic variation in prediction accuracy is generally small, it does indicate a generally lower level of accuracy for disadvantaged students or those with additional needs. Refinements to the prediction model could potentially be made to address these.

Similarly, while school-level variation in prediction accuracy is largely explained by small cohort sizes, the analysis does suggest some relationship that may be worth exploring. Local assessment conditions (such as invigilation practice, paper security, or timing of mocks) can of course always affect the accuracy of predicted grades.

4.4) Expand the evidence base in future cycles

This analysis has established a strong baseline for evaluating prediction accuracy. To build on it, Smartgrade could consider tracking year-on-year trends to assess whether model refinements are improving accuracy over time. Where sample sizes allow, deeper demographic analysis – including intersectional breakdowns – would also strengthen the evidence base.