All models are wrong[1], some of these might be useful: Options for adjusting school performance indicators for context

Acknowledgments

This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data.

This work uses research datasets which may not exactly reproduce National Statistics aggregates.

This publication includes analysis of the Department for Education National Pupil Database. Inferences or conclusions derived from the NPD in this publication are the responsibility of FFT Education Datalab and not the Department for Education.

Thanks to George Leckie for comments on a draft of this report.

Background

For the past two years, the Northern Powerhouse Partnership (NPP) has worked with Professor George Leckie and the late Professor Harvey Goldstein from Bristol University to publish a Fair Schools Index[2]. Their work[3] extends the Department for Education Progress 8 measure for secondary schools by including seven key pupil demographic characteristics: gender, month of birth, ethnicity, first language, special educational needs, neighbourhood deprivation[4] and pupil disadvantage[5].

We build on this work in a number of ways:

- Including additional variables, including a history of free school meal eligibility
- Including interactions between variables in the model
- Recalculating Attainment 8 to adjust for differences in grade severity between subjects
- Calculating measures for primary schools on the basis of Key Stage 1 (KS1) and Early Years Foundation Stage Profile (EYFSP) results

Extending the Fair Schools Index

We first extend the Fair Schools Index by including a number of additional variables and by calculating measures for primary schools. For secondary schools, we also include interaction effects and calculate a second version of our indicators based on a recalculated Attainment

[1] Box, G. E. P. (1976), Science and statistics, Journal of the American Statistical Association, 71 (356): 791–799.
[2] https://www.arcgis.com/apps/dashboards/1741a670cfcb493eb2cb20f14af8a064
[3] Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, *45*(3), 518-537.

[4] Decile of the income deprivation affecting children index, 2015 (https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015)
[5] Eligible for free school meals in the previous 6 academic years

8 measures in which we have tried to adjust for differences between subjects in grading severity.

Additional Variables

Firstly, and following previous work for NPP[6], we calculate a history of free school meal eligibility for pupils using School Census. This provides a termly snapshot of eligibility from Reception to Year 11 (or Year 6 in the case of pupils assessed at the end of Key Stage 2). Pupils are assigned to one of five bands:

- Never eligible for free school meals
- Eligible for free school meals <25% of terms
- Eligible for free school meals between 25% and 49% of terms
- Eligible for free school meals between 50% and 79% of terms
- Eligible for free school meals for 80% or more of terms (the long-term disadvantaged)

As the chart below shows, Progress 8 scores are biased with respect to pupil-level disadvantage. On average, the P8 score for the long-term disadvantaged is -0.53, over half a grade per subject below pupils with similar prior attainment on average. These biases are substantially improved by switching to the Fair Schools Index but some remains, particularly for the long-term disadvantaged.

These bands do not take account of short spells of eligibility that occur between School Census snapshots. In principle, these calculations could be improved using data on free

[6] http://www.northernpowerhousepartnership.co.uk/publications/educating-the-north-driving-ambition-across-the-powerhouse/

school meal spells collected by the Department for Education although we would not anticipate that doing so would have a material effect on schools' adjusted Progress 8 scores.

Secondly, and for secondary schools only, we include an additional group of pupils whose first language is not English. These are pupils who were classified as having a first language other than English at primary school but not at secondary[7]. For primary schools, we split pupils whose first language is not English into seven groups based on their year of first registration at a state school in England observed in School Census.

Thirdly, we include a flag of pupil mobility. For secondary schools, this indicates whether a pupil joined the school at which they completed Key Stage 4 after the start of Year 10. For primary schools, it indicates whether a pupil joined the school at which they completed Key Stage 2 after the start of Year 5.

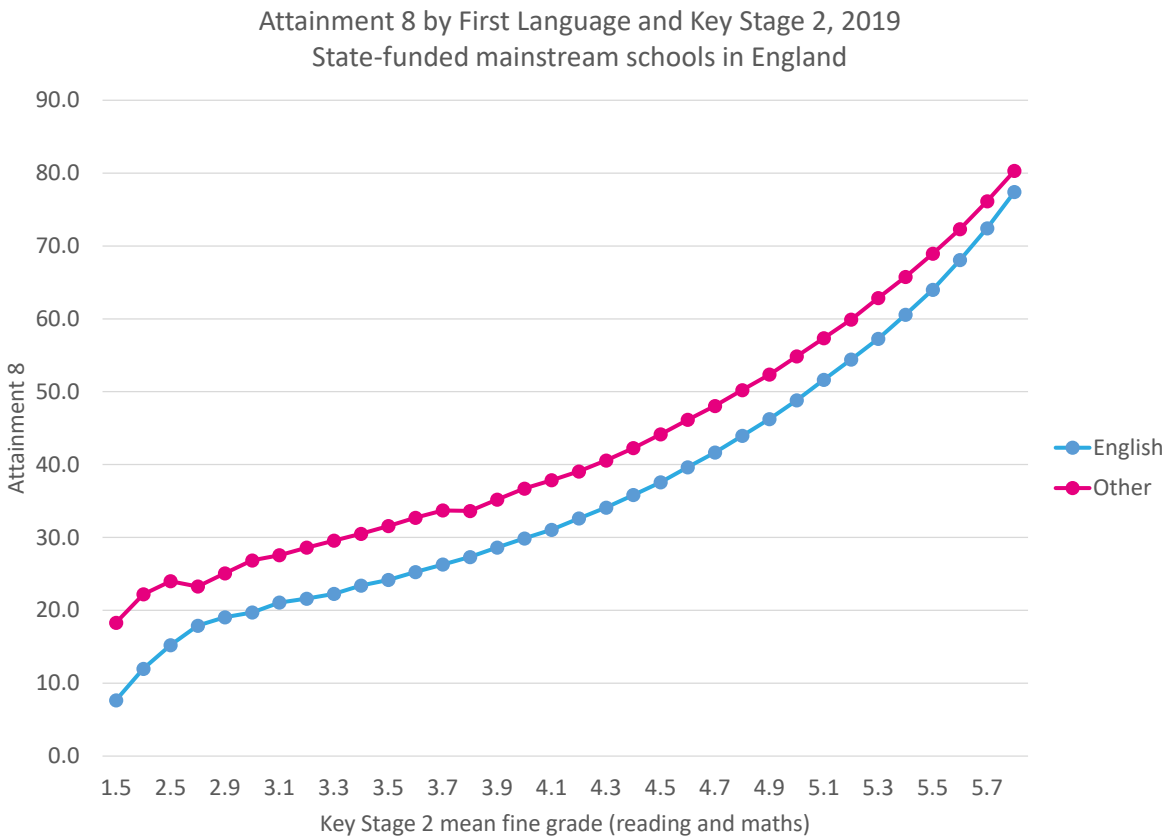
Fourthly, the Fair Schools Index includes pupils' special educational needs (SEN) status at the end of a Key Stage. This is a variable that schools can have a degree of control over. We therefore replace it with a measure of SEN status at the end of the previous Key Stage.

Including Interactions

The current Fair Schools Index extends the DfE Progress 8 methodology by entering the seven pupil demographic characteristics alongside the prior attainment bands used in the P8 calculation. As Leckie and Goldstein[8] point out, this ignores important interactions between the variables. As an example, the chart below shows how Attainment 8 scores increase with prior attainment for pupils by first language. Although pupils whose first language is other than English tend to score higher Attainment 8 scores than other pupils with the same level of prior attainment, the difference is narrower among those with higher prior attainment. This is likely to be due to Key Stage 2 results under-estimating the ability of pupils who had yet to reach fluency in English when tested.

[7] https://ffteducationdatalab.org.uk/2020/05/why-does-eal-status-change-between-primary-and-secondary-school-for-some-pupils/

[8] https://www.bristol.ac.uk/media-library/sites/education/documents/FINAL.pdf



We therefore extend the Fair Schools Index for secondary schools by including a number of interactions. These are the interaction between prior attainment and each of the demographic variables, plus the set of two-way interactions between free school meal history, first language, gender and ethnicity.

Including the interactions means fitting prior attainment as a continuous measure rather than as a set of dummy variables. This also requires consideration of the functional form of the relationship between prior attainment and the outcome measure. We find that fitting a cubic relationship meets two basic criteria. Firstly, and unlike with fitting a simple linear effect, it results in positive predictions across the prior attainment range. Secondly, and unlike a quadratic relationship, it produces monotonically increasing predictions across the prior attainment range, that is to say if prior attainment increases then predicted Attainment 8 score increases. It could be improved further (e.g. by fitting of splines) to reduce any remaining bias across the prior attainment distribution although we do not pursue this here.

Note that we do not attempt to interpret the parameter estimates resulting from this model. Our interest here is in producing school-level value added scores rather than precise estimation of any particular effect.

We address the suitability of these measures for measuring school effectiveness later in this report. For now, we will say that all of these models could be improved, for instance including additional interactions or additional compositional (school-level) effects, such as the percentage of disadvantaged pupils in the cohort. Our main aim here is to show how (and by how much) school value added scores can change as a result of changes to models and the choice of variables included.

Adjusting for grade severity

As we set out in a blogpost of ours[9], some Key Stage 4 qualifications appear to be scored more generously than others.

We therefore recalculate the DfE Attainment 8 measure for 2019 to adjust for grading severity in line with the methodology outlined in previous work of ours[10]. We then calculate Progress 8 and adjusted Progress 8 measures based on this measure.

The measurement scale we use is based on pupils' standardised scores in GCSE English and maths. In other words, we try to make grades in different qualifications comparable with GCSE English and maths. We then transform the resulting measures back onto the Attainment 8 scale to aid comparison with the other adjusted Progress 8 measures.

Although this approach removes differences in grading severity between subjects, it has some imperfections. In particular, it does not currently work particularly well for minority subjects taken by relatively few pupils.

Primary schools

The Department for Education currently produces Key Stage 1 to Key Stage 2 value added scores for each of reading, writing and maths. The methodology is similar to Progress 8 in that pupils are allocated to bands based on prior attainment, in this case Key Stage 1 average point score.

However, there is currently no "overall" measure for primary schools analogous to Progress 8. We address this by creating an overall measure of attainment based on pupils' scaled scores in reading, grammar, punctuation and spelling (GPS) and maths tests. The results in maths are double weighted to ensure even balance between literacy and numeracy. We use the versions of these measures used by DfE in their value added calculations, with default values based on teacher assessment allocated to pupils working below the standard of the tests.

We first calculate adjusted progress scores using the same methodology as Leckie and Goldstein. We then extend it by including the additional variables listed above. We also include an additional variable for junior schools given the differences in value-added scores between junior schools and all-through primary schools[11].

We then repeat the process using a different measure of prior attainment, the total score from the Early Years Foundation Stage Profile (EYFSP), usually recorded at the end of the Reception year. A new version of the EYFSP was introduced for the 2012/13 academic year[12]. The vast majority of pupils from this cohort will have taken Key Stage 2 tests in 2018/19. We therefore use scores from this version for the 2018/19 Key Stage 2 cohort and scores from its predecessor for previous cohorts. In general, slightly more pupils will be missing in value added measures based on EYFSP than on Key Stage 1. Most of these pupils will have first registered at a school in England during Key Stage 1 (for instance as a result of migration).

[9] https://ffteducationdatalab.org.uk/2019/09/are-some-qualifications-scored-too-generously/
[10] https://ffteducationdatalab.org.uk/2017/01/another-attempt-at-a-qualification-neutral-progress-8-measure/
[11] https://ffteducationdatalab.org.uk/2018/05/value-added-measures-in-performance-tables-a-recap-of-the-main-issues-for-primary-schools/
[12] https://www.gov.uk/government/publications/the-early-years-foundations-for-life-health-and-learning-an-independent-report-on-the-early-years-foundation-stage-to-her-majestys-government

Models

Below we describe the models we have run. The STATA code and details of variables used are included in the online appendix.

Secondary Schools

For secondary schools we produce adjusted value added indicators for two outcomes:

- Attainment 8
- Attainment 8 adjusted for grading severity

For each outcome we then produce three sets of value added indicators using different model specifications.

Ver	Description	Variables included	Interactions?
V1	Fair Schools Index (as calculated by Leckie and Goldstein)	Key Stage 2 band, gender, month of birth, disadvantage, ethnic background, first language, IDACI decile, Year 11 SEN status	No
V2	With additional variables	Key Stage 2 band, gender, month of birth, FSM history, ethnic background, first language, IDACI decile, Year 6 SEN status, mobility, first language at primary school	No
V3	With additional variables and interactions	As V2 including some interaction effects and fitting Key Stage 2 as a cubic line rather than as dummy variables	Yes
V4	With additional variables and interactions and mean school prior attainment	As V3 plus school-level mean Key Stage 2 prior attainment	Yes

We also calculate a "Progress 8" measure (based solely on prior attainment) for the Attainment 8 measure adjusted for grading severity.

For Progress 8, DfE caps the scores of pupils with extremely negative scores. This step makes little practical difference[13]. Consequently, we do not apply any adjustments to the additional scores we calculate here.

Our V1 replication of the current Fair Schools Index produces very similar though not identical results. This is likely to be due to a) different approaches to handling missing data and b) changes to the underlying data, for instance as a result of updates to the references that link pupils in different datasets. Scores match to within 0.03 points for 87% of schools. Using the version of the Fair Schools Index prior to capping of extremely negative scores increases this percentage to 93%.

Primary Schools

For primary schools, the outcome measure is average scaled score in reading, GPS and maths tests, with maths double weighted. We produce measures based on both Key Stage 1 and the Early Years Foundation Stage Profile.

[13] https://ffteducationdatalab.org.uk/2018/10/provisional-gcse-and-equivalent-results-2018-the-impact-of-changes-to-progress-8/

Ver	Description	Variables included	Interactions?
V1	KS1-KS2 analogue to the Fair Schools Index (as calculated by Leckie and Goldstein)	Key Stage 1 band, gender, month of birth, disadvantage, ethnic background, first language, IDACI decile, Year 6 SEN status	No
V2	KS1-KS2 with additional variables	Key Stage 1 band, gender, month of birth, FSM history, ethnic background, IDACI decile, Year 2 SEN status, mobility, year of first registration for pupils whose first language is other than English	No
V3	KS1-KS2 additional variables and interactions and mean school prior attainment	As V2 plus school-level mean KS1 prior attainment	Yes
V4	FSP-KS2 analogue to the Fair Schools Index (as calculated by Leckie and Goldstein)	EYFSP band, gender, month of birth, disadvantage, ethnic background, first language, IDACI decile, Year 6 SEN status, junior school flag	No
V5	FSP-KS2 with additional variables	EYFSP band, gender, month of birth, FSM history, ethnic background, IDACI decile, Year 2 SEN status, mobility, year of first registration for pupils whose first language is other than English, junior school flag	No
V6	FSP-KS2 additional variables and interactions and mean school prior attainment	As V5 plus school-level mean EYFSP prior attainment	Yes

Because Department for Education do not calculate an overall measure of value added, we fill this gap by adopting the methodology they employ for their value added calculations in reading, writing and maths.

Results

Percentage of variance explained

Table 1 shows the percentage of variance explained by each model for secondary schools. We also include the Progress 8 measure for comparison. Adding the seven demographic factors (V1) increases the percentage of variance explained from 54.4% to 61.0%. V2, V3 and V4 make small improvements

Table 1: Percentage of variance explained by each model, Secondary Schools 2019

Version	Attainment 8	Attainment 8 adjusted for grading severity
P8	54.2	54.4
V1	61.0	61.4
V2	61.7	62.2
V3	62.1	62.6
V4	62.4	62.9

Number of schools: 3196
Number of pupils: 502510

Similar information is presented for primary schools in Table 2. Here we see a reduction in variance explained as a result of adding additional variables (V1 to V2 and V4 to V5). This is due to the inclusion of Year 2 SEN status in place of Year 6 SEN status. The latter is more predictive of Key Stage 2 outcomes. Almost certainly some pupils would have had yet to have SEN diagnosed in Year 2. The V2 percentage of variance explained would increase to 65.6% if Year 6 SEN status was used. Similarly, the percentage for V5 would increase to 44.8%.

Models based on Foundation Stage attainment are much less correlated with Key Stage 2 outcomes than models based on Key Stage 1. This is to be expected given the 6-year gap between the end of the Foundation Stage and Key Stage 2. In addition, the "new" EYFS profile data is less strongly correlated with Key Stage 2 than the previous version. The percentage of variance explained in the EYFSP-KS2 VA model for 2018 was 34.8%, for example.

Table 2a: Percentage of variance explained by each KS1-KS2 model, Primary Schools 2019

Version	% of variance explained
VA	61.0
V1	65.0
V2	64.6
V3	64.7

Number of schools: 15149
Number of pupils: 604293

Table 2b: Percentage of variance explained by each EYFSP-KS2 model, Primary Schools 2019

Version	% of variance explained
VA	30.8
V4	44.6
V5	39.9
V6	41.0

Number of schools: 15148
Number of pupils: 590573

The unexplained variation gives rise to schools' value added scores. If all variance were explained, all schools would have a value added score of zero. The remaining variation is assumed to be attributable to schools. In fact, it contains any school-level effects plus the influence of unmeasured variables (e.g. tutoring or parental support). Some of the unmeasured variables may be correlated with variables included in the model (e.g. disadvantage).

Correlations between school-level value added scores

We now examine the correlation in school-level scores between the different versions of the models and the published Progress 8 measure (Table 3).

Secondary schools

In general, all of the school-level contextual value added measures (V1-V4) are highly correlated, with correlations of 0.949 or higher. Adding in further variables tends to reduce the correlation to the published DfE P8 measure although they still exceed 0.8.

Table 3: Correlation between 2019 KS2-KS4 measures, state-funded mainstream schools and colleges 2019

Measure	P8	V1	V2	V3	Schools
V1	0.903				3196
V2	0.892	0.994			3196
V3	0.883	0.990	0.998		3196
V4	0.816	0.949	0.962	0.965	3196

Table 4 shows the correlation between 2019 school-level CVA measures and those from 2017 and 2018. Only schools with three years of data are included. In general, the measures reduce in correlation from year to year. In other words, schools value added scores exhibit a degree of change from year to year. This is to be expected: schools change. The more variables that are included in the measures, the lower the year-on-year correlation tends to be. Once systematic differences in intakes between schools are removed, the resulting measures tend to exhibit less stability.

Table 4: Correlations between 2017, 2018 and 2019 KS2-KS4 measures, state-funded mainstream schools and colleges

Measure	2017	2018	Schools
P8	0.76	0.86	3090
V1	0.65	0.79	3090
V2	0.64	0.79	3090
V3	0.64	0.78	3090
V4	0.60	0.75	3090

Adjusting for grade severity in Key Stage 4 qualifications makes little practical difference. Compared to analogous models based on unadjusted Attainment 8, those which adjust for grading severity are highly correlated (Table 5). However, it does make a difference in exceptional circumstances. Comparing model V4 for both measures shows that 18 schools (just half of one percent) see their score fall by 0.25 or more when Attainment 8 is adjusted for grading severity.

Table 5: Correlations between value added scores based on Attainment 8 and value added scores based on Attainment 8 adjusted for grading severity 2019, state-funded mainstream schools and colleges

Measure	Corr	Schools
P8	0.99	3196
V1	0.99	3196
V2	0.99	3196
V3	0.99	3196
V4	0.99	3196

Primary schools

Correlations for the various school-level value added measures for primary schools are presented in Tables 6a and 6b[14]. The contextual value added measures are strongly correlated with the un-contextualised value added measures. These correlations are slightly stronger than those for secondary schools, particularly those based on Key Stage 1.

Table 6a: Correlation between 2019 KS1-KS2 measures, state-funded mainstream schools 2019

Measure	KS1 VA	V1	V2	Schools
V1	0.946			14883
V2	0.942	0.987		14883
V3	0.940	0.979	0.987	14883

Table 6b: Correlation between 2019 FSP-KS2 measures, state-funded mainstream schools 2019

Measure	FSP VA	V4	V5	Schools
V4	0.917			14883
V5	0.924	0.973		14883
V6	0.888	0.932	0.951	14883

Year-on-year stability in school-level value added scores for primary schools is shown in Tables 7a and 7b. Compared to secondary schools, these correlations are lower. This indicates more year-on-year variability in scores, arising at least in part from smaller cohort sizes.

Table 7a: Correlations between 2017, 2018 and 2019 KS1-KS2 measures, state-funded mainstream schools

Measure	2017	2018	Schools
VA	0.485	0.610	14199
V1	0.465	0.601	14199
V2	0.447	0.589	14199
V3	0.474	0.602	14199

Table 7b: Correlations between 2017, 2018 and 2019 EYFSP-KS2 measures, state-funded mainstream schools

Measure	2017	2018	Schools
VA	0.462	0.554	14199
V4	0.444	0.547	14199
V5	0.430	0.534	14199
V6	0.494	0.582	14199

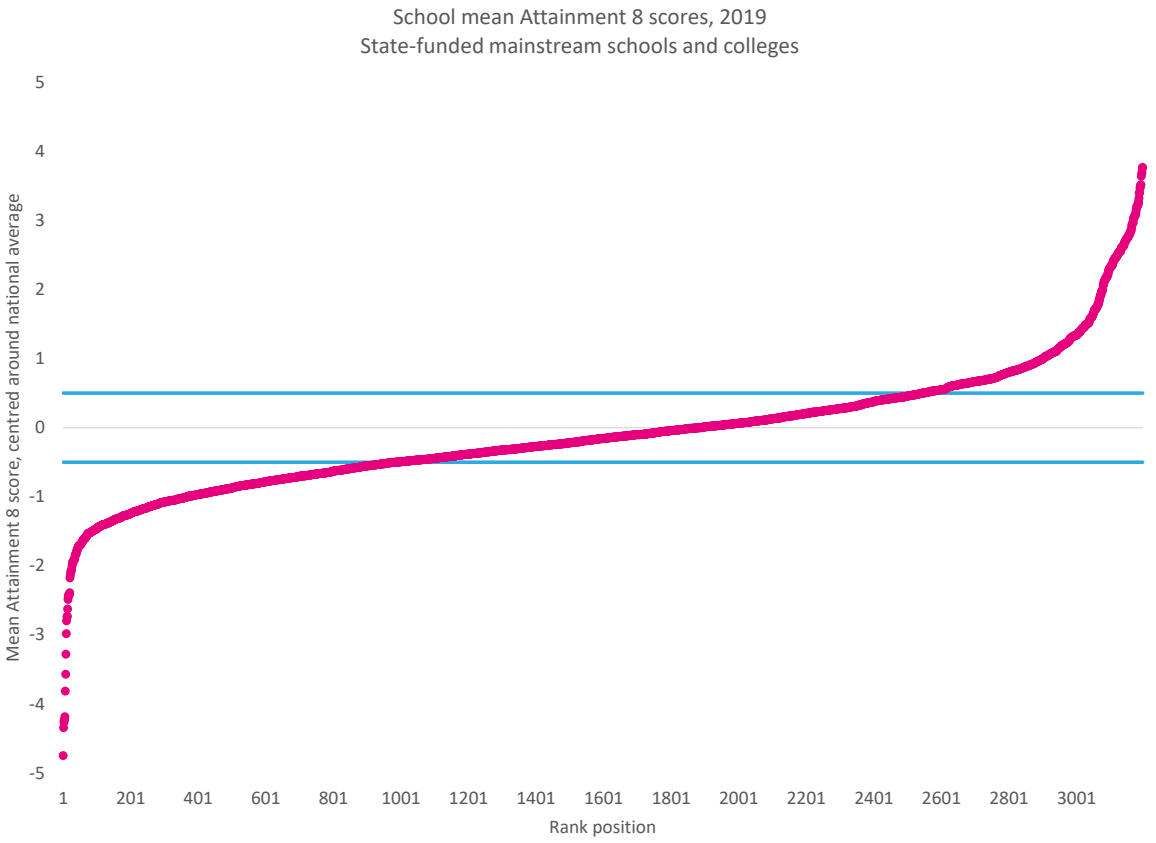
[14] These tables are based on schools with at least six pupils with value added scores in 2019

In short then, different contextualised value added models produce broadly similar results. As we will see in the following sections, this can still result in large changes in scores for a small number of schools and a large change in rank position for a larger number, particularly those in the middle of the distribution. Contextualised value added scores exhibit a degree of volatility from year to year, particularly for primary schools. In other words, a score for one year may not be predictive of the score next year.

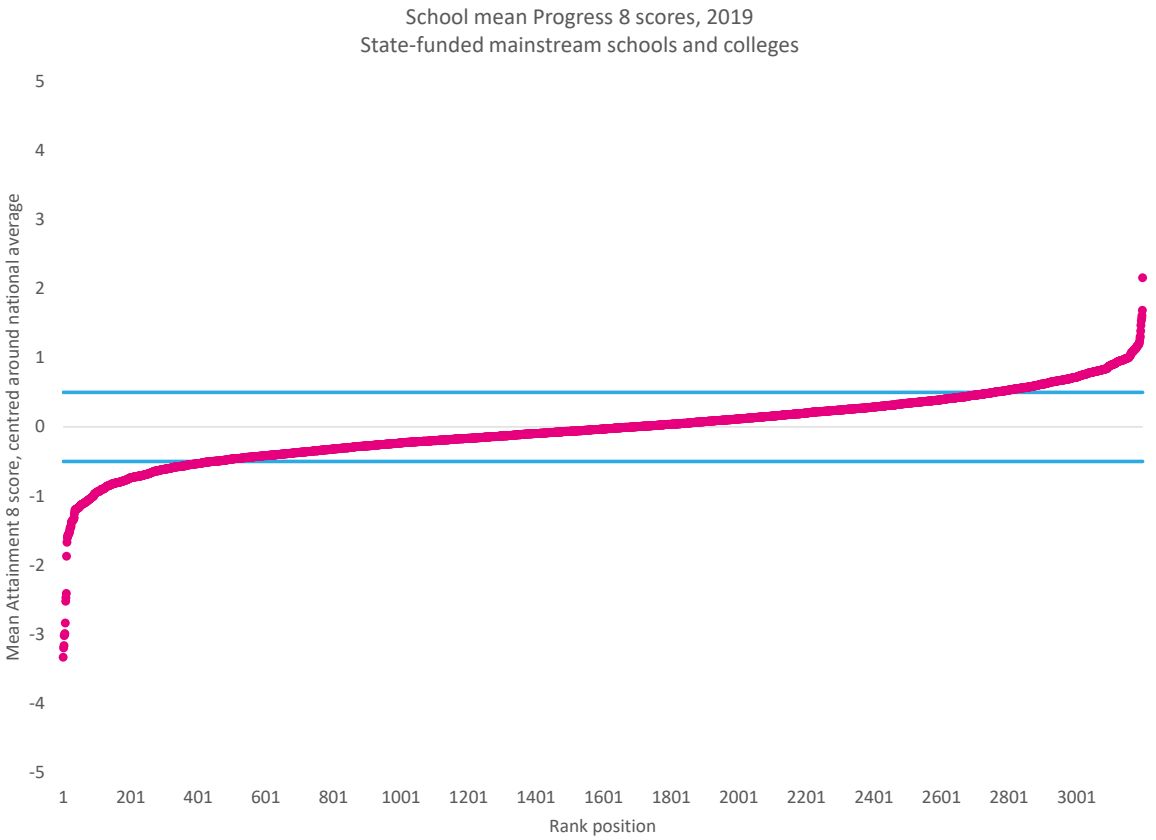
How different are schools' value added scores?

In the following charts, we show the distribution of scores for secondary schools ranked from lowest to highest.

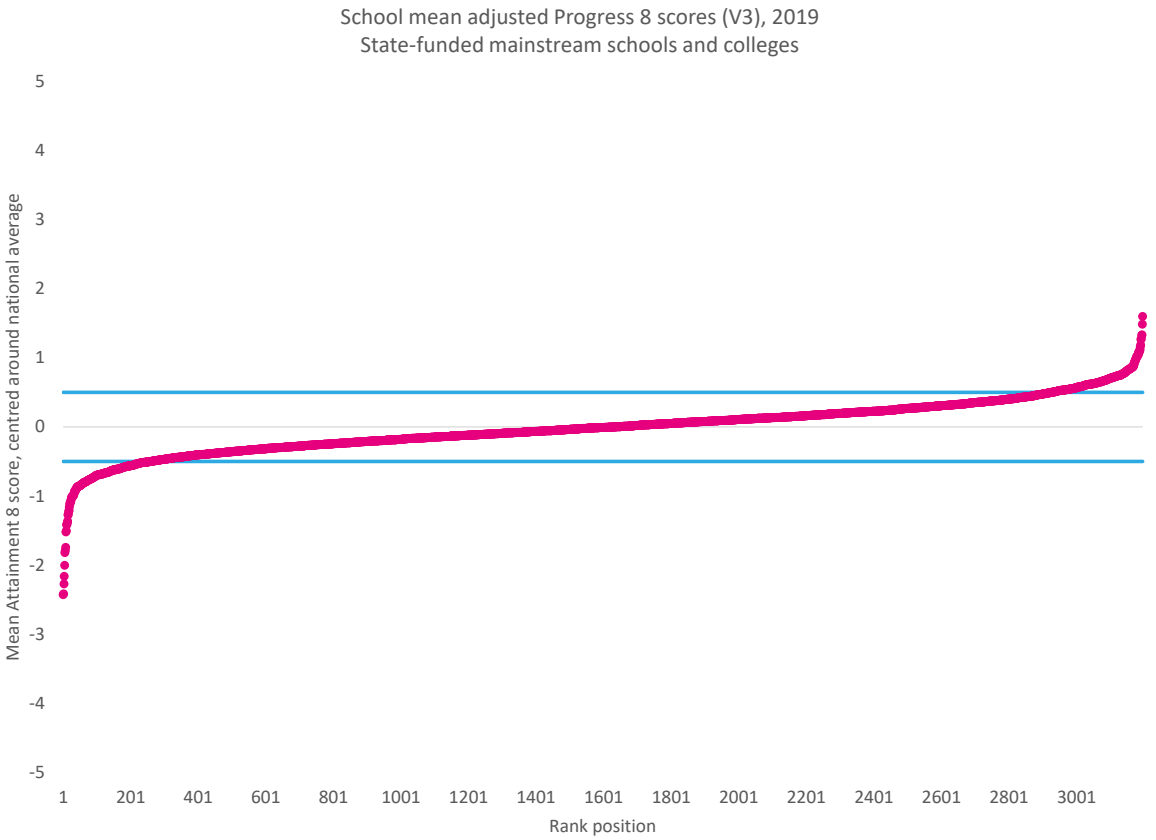
To start with, we plot the distribution for Attainment 8 before calculating any value added measures. To aid comparison with subsequent charts, we have put Attainment 8 onto the same scale as Progress 8 by centring around the national average for state-funded mainstream schools (47.4) and dividing by 10. For instance, a value of 1 on the chart equates to an Attainment 8 score of 1*10 + 47.4 = 57.4. The blue horizontal lines are plotted at 0.5 and -0.5. Pupils at schools with scores of 0.5 (or -0.5) achieve, on average, half a grade above (or below) the national average in each subject counted in Attainment 8. 40% of schools are plotted between these lines.



Next, we plot the distribution of Progress 8 scores. This removes the effect of prior attainment from Attainment 8. The distribution is flatter: variation in school attainment is reduced. This emphasises that for all its faults, Progress 8 offers a fairer basis for comparing schools than Attainment 8. The percentage of schools plotted between the horizontal blue lines increases to 59%.



Finally, we plot one of the contextualised value added measures (V3). This plot is flatter still, with 68% of schools plotted between the horizontal blue lines.



Broadly speaking, the final chart suggests that differences in attainment between the majority of schools are relatively small once prior attainment and pupil context is taken into account. Over 700 (22%) schools have a CVA score between -0.1 and 0.1. These schools differ by, on average, at most two grades across the 10 slots included in Attainment 8[15]. This begs the question of whether there are any substantive differences in attainment between these schools.

There does not appear to be a consensus on the size of a school value added score that is substantively important. The nearest thing to this was the now abandoned floor standard of -0.5. This indicated that pupils' results at a school were, on average, half a grade per subject lower in each of the 10 slots of Attainment 8 compared to pupils with similar prior attainment nationally. Although this related to A*-G grades when first introduced, the same principle would apply with GCSEs graded 9-1. A difference of -0.5 would be equivalent to the difference between one pupil who achieved grade 5 in every slot of Attainment 8 and another pupil who achieved grade 5 in five slots and grade 4 in the remaining five slots.

In Performance Tables, schools are banded into five groups:

Well above average	Score is 0.5 or higher and lower confidence interval is above 0
Above average	Score is above 0 but below 0.5 and the lower confidence interval is above 0
Average	Lower confidence interval is below 0 and upper confidence interval is above 0

[15] Attainment 8 is composed of 2 slots in English, 2 slots in Maths, 3 slots in other EBacc subjects and 3 "open" slots.

Below average	Score is below 0 but above -0.5 and the upper confidence interval is below 0
Well below average	Score is -0.5 or lower and the upper confidence interval is below 0

We can think of the confidence intervals used in the calculations like this. Let's take an average sized secondary school with 160 pupils in Year 11 as an example. If we were to take random samples of pupil-level Progress 8 scores from the national population, 95% of the samples would have an score between -0.2 and +0.2[16]. Any schools of size 160 with Progress 8 scores between these values would be classified as "average". The size of the range varies with respect to number of pupils, the greater the number of pupils, the smaller the P8 range classified as "average".

We can show how schools' ratings would change under each value added model. For comparison purposes, we use the same standard errors as the Progress 8 calculation[17]. In other words, we use the same width for the confidence interval of each model. Results for secondary schools are shown in Table 8. The number of schools classified as average increases from 1145 (36%) under P8 to 1517 (47%) under V4 as more factors are included.

Table 8: Number of secondary schools in each Performance Tables rating band for Key Stage 2 to Key Stage 4 value added, 2019

	P8	V1	V2	V3	V4
Well above average	445	289	277	272	244
Above average	543	599	590	589	590
Average	1145	1384	1429	1439	1517
Below average	621	640	636	638	622
Well below average	442	284	264	258	223
Total	3196	3196	3196	3196	3196

For primary schools, the value added measure we use has never been published and therefore never used in floor standards. We set a score of +3.0 to define the well-above average group and -3.0 for the below average group. In Table 9, we show the number of schools in each band based on the Key Stage 1 to Key Stage 2 measures.

Table 9: Number of primary schools in each Performance Tables rating band for Key Stage 1 to Key Stage 2 value added, 2019

	VA	V1	V2	V3
Well above average	1359	1277	1214	1158
Above average	1817	1832	1796	1836
Average	8429	8818	8812	8867
Below average	1779	1733	1763	1745
Well below average	1499	1223	1298	1277
Total	14883	14883	14883	14883

[16] Given the 2019 national standard deviation in Progress 8 scores of 1.2825 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872997/Secondary_accountability_measures_guidance_February_2020_3.pdf)

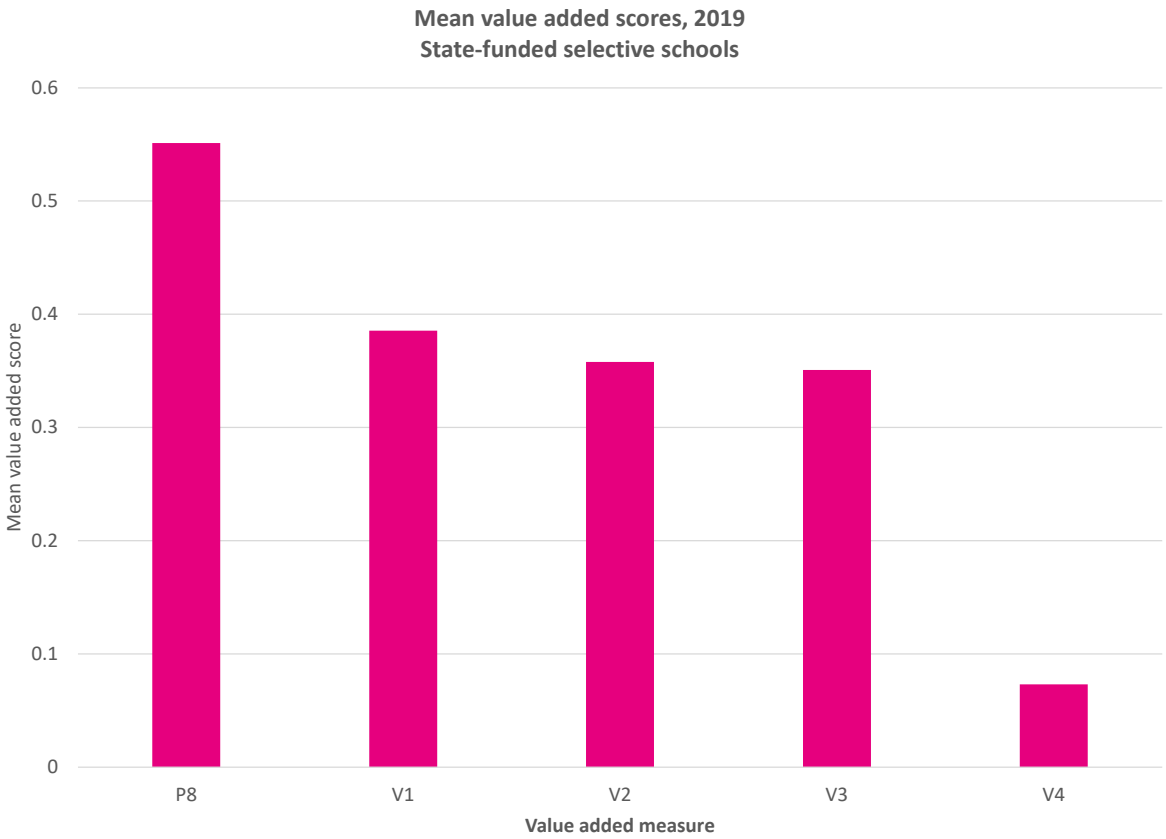
[17] In reality, each value added measure has its own standard error

As with secondary, the number of schools classified as average increases when contextual factors are included. Partly as a function of smaller cohort sizes, more primary schools tend to be classified as average. 8429 (57%) are classified as average for uncontextualised value added compared to 36% of secondary schools.

Perhaps as a consequence of the lack of consensus on the size of value added score that is educationally important, there has been a focus on the rank position of schools and their 5-point summary ratings. However, in the middle of the distribution small changes in score can lead to large changes in rank. More attention should be paid to schools' scores and what they mean in practice. Although a school with a P8 score of +0.2, for example, might be classified as "above average", we should ask what it means in real terms (i.e. pupils have achieved a single grade higher in two of the 10 slots of Attainment 8 compared to the average of similar pupils nationally).

Selective Schools

State-funded selective (grammar) schools tend to achieve above average Progress 8 scores. In 2019, this average stood at 0.55 across the 162 selective schools with published P8 measures.



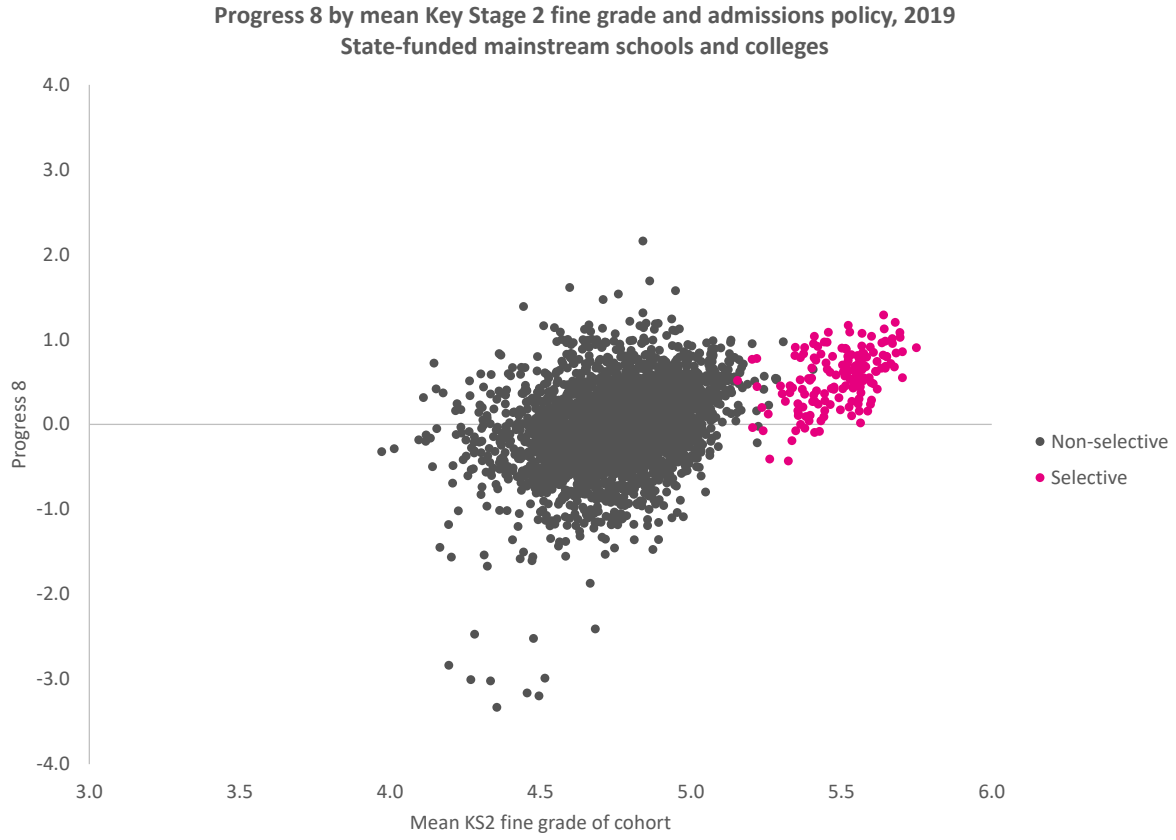
The reasons for this are not clear[18]. It could be the case that selective schools are able to recruit more effective teachers or that the curriculum can be delivered at greater pace when higher-attaining pupils are taught together.

[18] https://ffteducationdatalab.org.uk/2015/05/why-do-pupils-at-schools-with-the-most-able-intakes-tend-to-make-the-most-progress/

Tom Perry[19] has suggested that some of the difference may be due to measurement error in Key Stage 2 results tending to affect selective schools more than others. This supports previous work of ours which showed that the P8 advantage for selective schools disproportionately accrued from pupils with lower Key Stage 2 results[20]. Perry suggests correcting for this by including the mean Key Stage 2 score of the cohort in the value added model as we have done in measure V4.

As the chart above, the value added advantage for selective schools diminishes but persists when pupil context is taken into account (V1-V3). However, it reduces markedly when cohort mean Key Stage 2 is included (V4).

However, including the mean Key Stage 2 score of the cohort changes the interpretation of the value added score. This can be illustrated by plotting Progress 8 scores and mean cohort Key Stage 2 score for all state-funded schools as in the chart below. Although there are a handful of non-selective schools with similar mean Key Stage 2 scores to some selective schools there is not much in the way of overlap.



By including the mean Key Stage 2 score of the cohort in value-added calculations, the comparison being made changes from that of similar pupils to that of similar pupils in similar schools.

[19] http://pure-oai.bham.ac.uk/ws/files/48300038/Perry_phantom_compositional_effects_Research_Papers_in_Education_2018.pdf

[20] https://ffteducationdatalab.org.uk/2016/10/provisional-ks4-data-2016-grammar-schools-reporting-fantastic-progress-8-scores-not-so-fast/

Limitations

Many of the problems associated with Progress 8 and primary value added measures stem from them being used as measures of school effectiveness. They are not. They are measures of attainment net of the effect of prior attainment.

None of the models presented here are measures of school effectiveness either. We can get a bit closer to this goal by adjusting for factors besides prior attainment that are associated with attainment and yet are beyond the control of schools such as disadvantage and ethnic background. These often act as proxies of factors we do not observe directly, such as parental support, access to resources and tutoring. In an ideal world, we would have data about them.

In addition, schools (secondary schools in particular) can exercise some degree of control of which pupils are included in Performance Tables calculations through formal and informal exclusion prior to the January of Year 11[21]. None of the measures presented here adjust for pupil exits.

None of the models adjust for measurement error in either outcomes or the factors that are used as controls, including prior attainment. GCSEs and other qualifications are not measured with perfect reliability[22] although the aggregation of them into Attainment 8 reduces some of the error in individual subject results. Similarly, Key Stage 2 results are highly reliable but not perfect tests[23].

This matters because the value added scores presented here assume perfect measurement. It may not be reasonable to assume that errors affect schools equally. Tom Perry[19] shows how errors in measuring prior attainment may favour grammar schools, for example.

And because neither attainment nor prior attainment is measured perfectly, measurement errors are amplified when controlling for prior attainment to calculate value added measures as Stephen Gorard showed[24]. This means there may be a fair amount of statistical noise in value added scores, particularly from year to year.

Summary

Value added models of attainment such as Progress 8 tend to show that differences in attainment between schools tend to be small. They become smaller still when other pupil background factors such as disadvantage and ethnic background are included.

Value added models are sensitive to which factors are included, how they are included (e.g. whether there are interactions) and how missing data is treated. Different models will produce different results. For many schools these differences will be small but there may be exceptional cases where differences are larger.

Including other pupil background factors in value added models offers a fairer basis for comparing schools. However, they fall short of being measures of school effectiveness.

Ultimately, the value added scores for the majority of schools are not that different from each other. They also exhibit a fair degree of volatility from year to year. Given some of the

[21] https://ffteducationdatalab.org.uk/2019/12/whos-left-2019-part-two-how-do-you-lose-6700-pupils/

[22] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578868/2011-03-16-estimates-of-reliability-of-qualifications.pdf

[23] https://www.gov.uk/government/publications/2019-national-curriculum-test-handbook

[24] https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1080/01411920903144251

uncertainty in the measurements on which they are based, they are unsuitable for putting schools in order from best to worst or at detecting changes in attainment from year to year.

School-level value added measures tell us nothing about the variation in attainment within schools. Previous research tends to show that only around 14% of the variation in Progress 8 schools is between schools, the rest between pupils[25]. In other words, there is substantial variation in pupil performance in all schools. Moreover, there is more variation between departments in schools than between schools[26].

However, value added measures still have a role to play. When there are consistent results over a number of years, they can detect a small number of schools where support may be needed to improve attainment and a similar small number of schools which appear to be achieving results beyond expectation.

Value added measures also show how much variation in attainment remains within the school system once prior attainment and pupil backgrounds are taken into account. Not all schools can achieve a value added score above zero. But a system in which all schools' results were close to zero would represent a degree of success as it would suggest that there were no systematic differences in attainment between schools. Attainment gaps would likely remain (e.g. between disadvantaged pupils and their peers) but these would require policies that influenced the work of all schools rather than just those judged to be under-performing.

Appendices

We provide school-level scores for each of the models described above in two Excel workbooks, one for primary and one for secondary.

The appendices contain data for three years. We have linked schools to predecessor institutions where they have merged, changed academy trust or seen any other change in governance.

The secondary workbook contains two tabs:

- Original
- Qualification Adjusted

The "original" tab shows value added scores based on the Department for Education Attainment 8 measure. The "qualification adjusted" tab shows value added scores based on the version of this adjusted for grading severity. These scores have only been calculated for 2019.

The primary workbook contains two tabs:

- KS1-KS2
- EYFS-KS2

The "KS1-KS2" tab contains Key Stage 1 to Key Stage 2 value added measures and the "EYFS-KS2" tab contains Early Years Foundation Stage to Key Stage 2 value added measures.

[25] http://ftp.iza.org/dp11372.pdf
[26] https://ffteducationdatalab.org.uk/2020/01/are-we-looking-in-the-wrong-place-to-improve-attainment/

We do not include results for further education colleges which admit pupils at age 14 in the secondary workbook. This is because they do not complete School Census and therefore we cannot currently include some of the demographic variables used in the models such as ethnic background.

We have not applied capping of extremely negative values to the contextualised value added scores of secondary schools although they are applied to the Progress 8 scores so that they match published statistics. Capping only has a negligible effect on school-level scores[13].