

What Matters in Schools for
children's academic progress?

Why is this important?

UK public expenditure on secondary education is £43bn

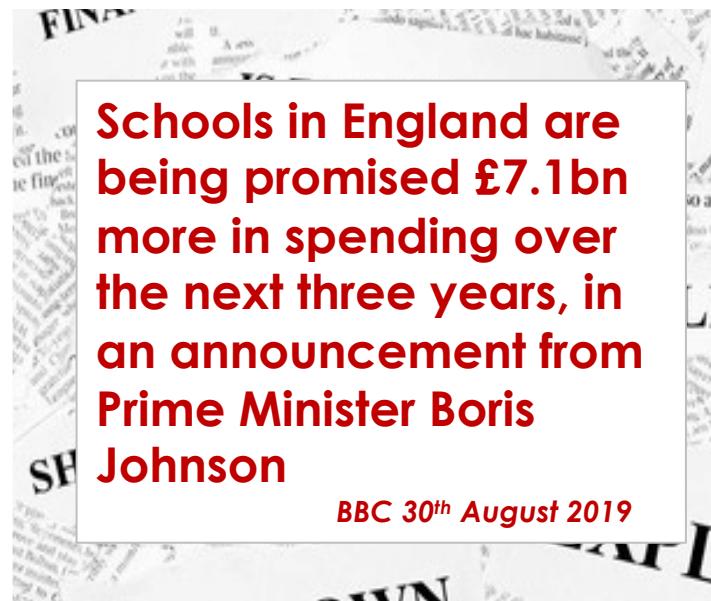
Where did spending go in 2018-19?

	£bn	% of total
Pre-primary	4.2	5%
Primary	26.4	30%
Secondary	43.1	49%
Tertiary	4.5	5%
Subsidiary services	3.5	4%
R&D	2.5	3%
Training	2.2	3%
Other	1.4	2%
Total	87.9	100%

Source: <https://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN01078>

£88bn is 4.2% of GDP; the peak in the last 40 years of 5.5% in 2010

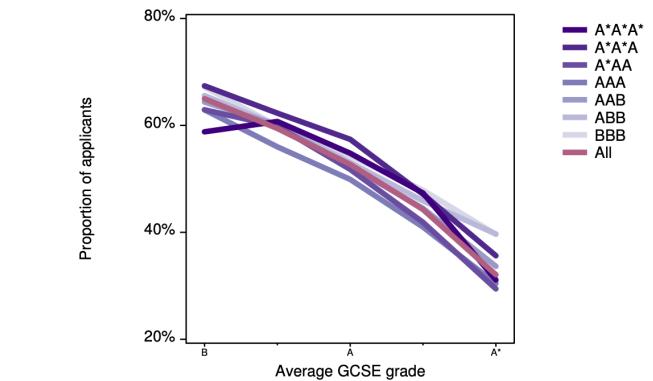
...made more important because it will soon rise...



...and its impact in people's future lives is significant

In 2016 UCAS found that GCSE results and the mix of A levels studied are the key factors determining whether university applicants meet their predicted grades

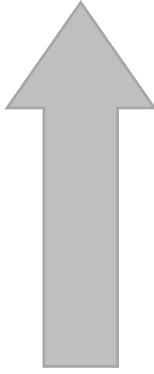
Figure 5 Proportion of applicants who missed their predicted attainment by two or more points by GCSE attainment and best three predicted grade profile



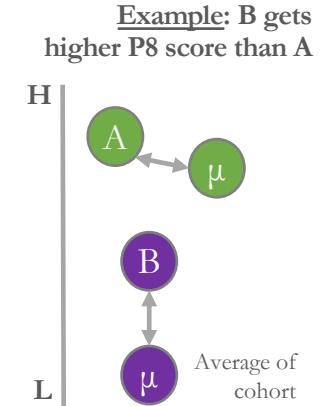
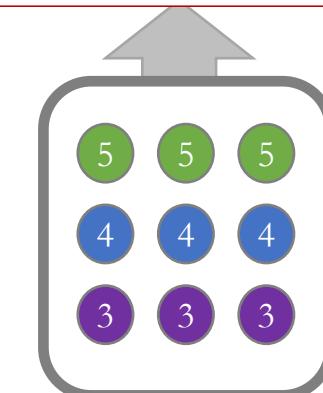
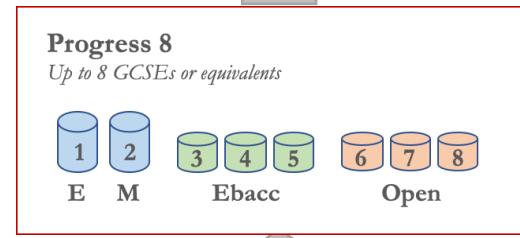
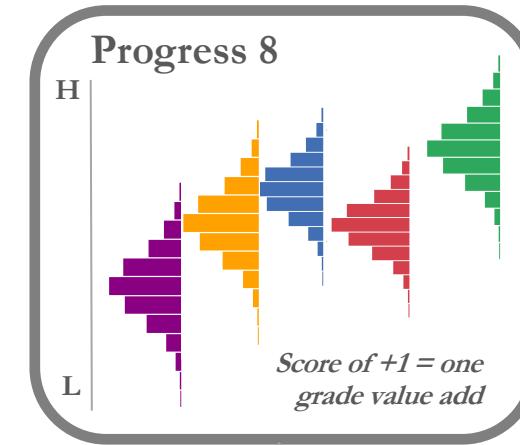
Source: <https://www.ucas.com/file/71796/download?token=D4uuSzur>

How are English secondary schools measured?

A: on the progress that children in their schools make relative to their academic peers when they were 11



5 Score in Maths and English tests at 11 years old



Goal

What explains English children's academic progress between 11 and 16 years old, relative to their peers?

...as measured by R^2 , and looking at data for GCSEs taken in 2018

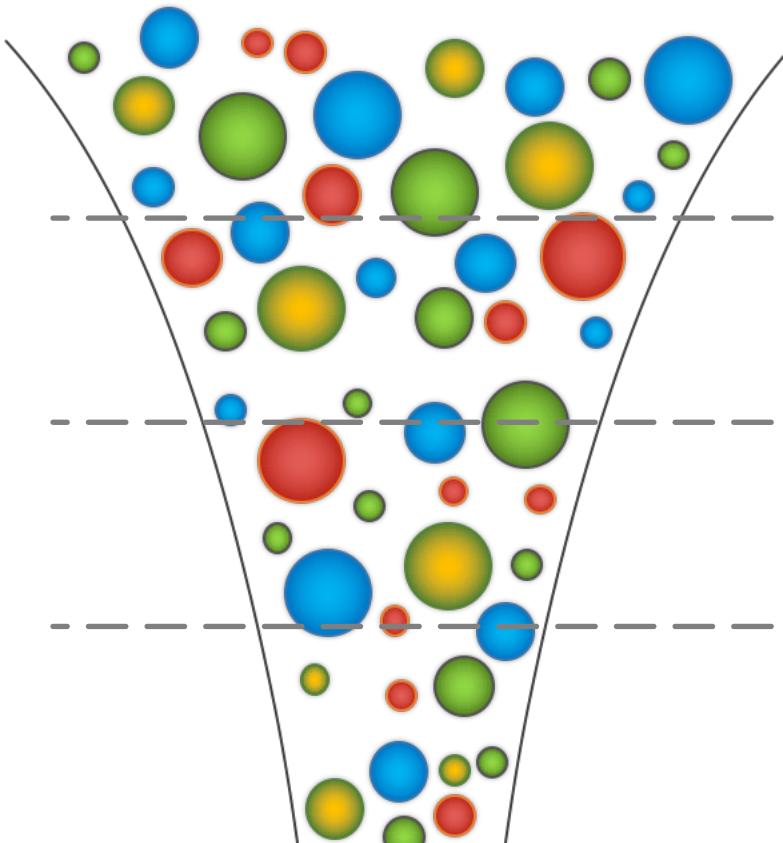
Data

Exams taken in 2018

5,673 ‘observations’

496 features

- Target y
- Not relevant data
- Fix NaNs
- Additional data



	# Features	# Schools
• 1,980 rows of data with non-numeric target y	496	3,541
• 152 rows of summary data • 376 Special Schools *	279	3,165
• Duplicate features		
• Imputed x5 non-numeric categories used by govt		
• NaNs imputed or features removed	27	3,165
• Added 4 new features incl Teacher / Pupil Ratio	31	<u>3,067</u>

71% of all target schools

95% of children in target schools

Adding more features would have continued to reduce the dataset

* There are 1,191 Special Schools in total (some were excluded in the first data cleaning step), all excluded from the model as they provide a very specific learning environment for students

Final Features list

31 Features

10 one-hot encoded

(marked with *)

3 dropped (IDs)

(marked with +)

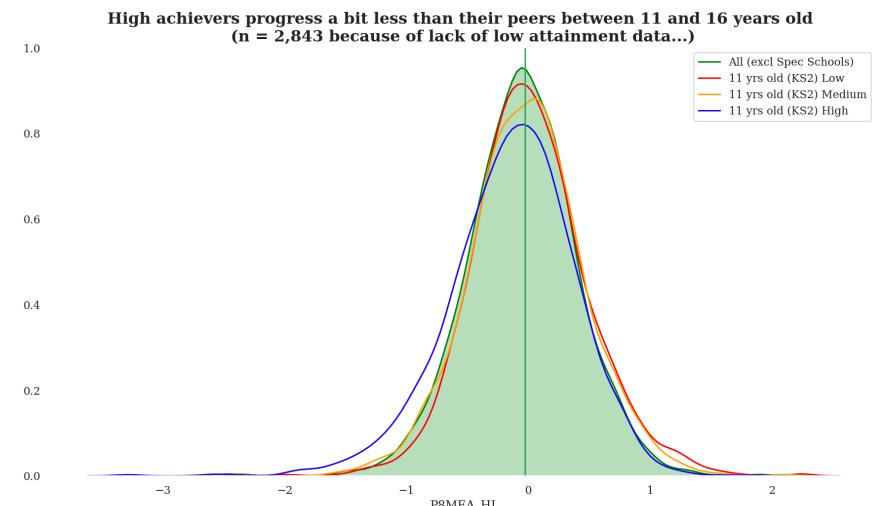
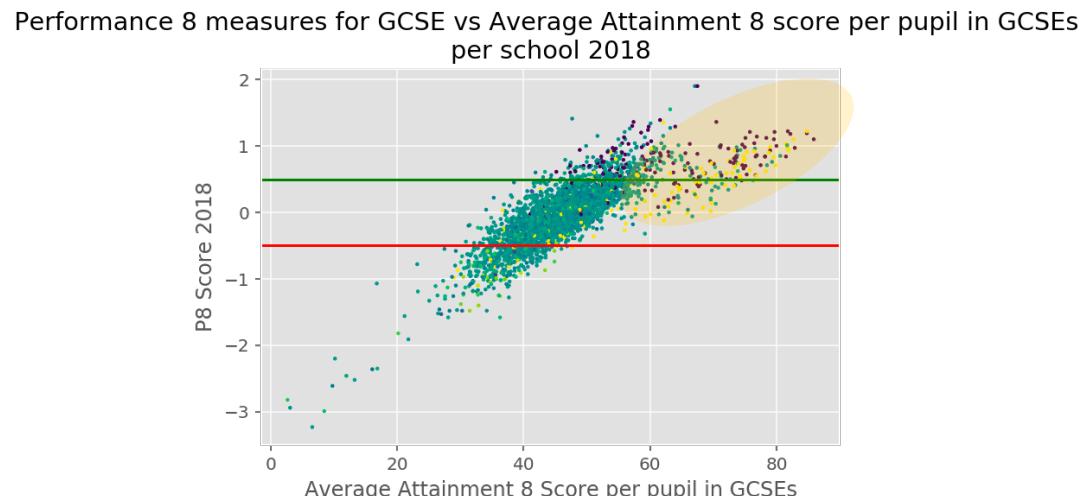
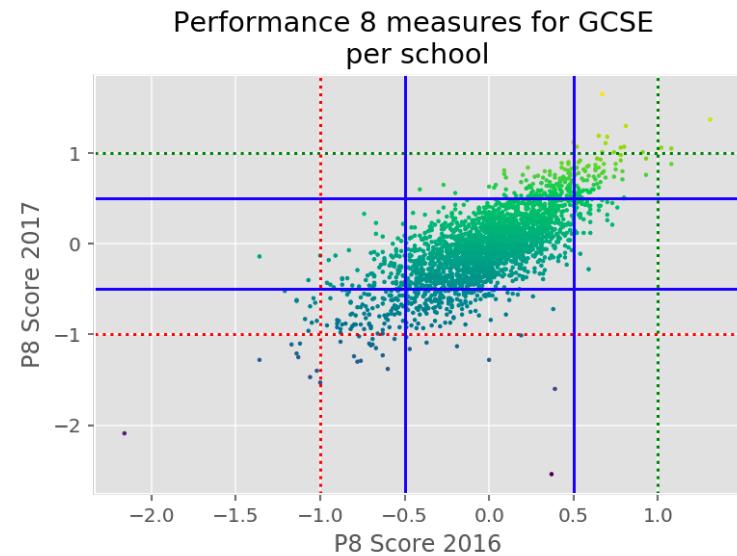
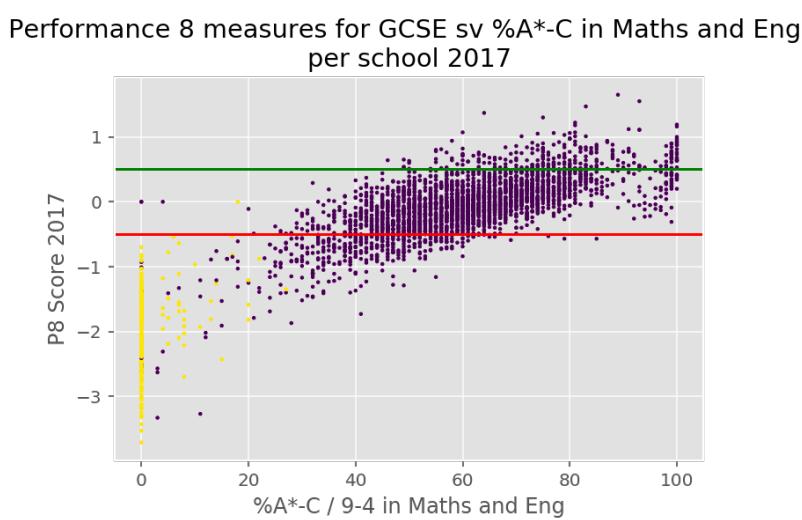
Feature_name	Metafile description
+ School_type	Record type (1=mainstream school; 2=special school; 4=local authority; 5=National (all schools); 7=National (maintained schools))
* Local_Ed_Auth	Local authority code (see separate list of local authorities and their codes)
* Estab_group	Establishment number
+ School_ID	School Unique Reference Number
* Is_Closed	Closed school flag (0=open; 1=closed)
* School_type_2	School type (see separate list of abbreviations used in the tables)
* Rel_Denom	School religious character
* Selective_School	School admissions policy as in KS4 SFR from 2017 onwards See 'abbreviations' document for more information.
* School_gender	School gender of entry
* School_feeder_sixth_form	Indicates whether school is a feeder school for sixth form centre/consortia (0=No; 1=Yes)
* School_primary	Indicates whether school is published in the primary school (key stage 2) performance tables (0=No; 1=Yes)
* School_sixth_form	Indicates whether school is published in the school and college (16-18) performance tables (0=No; 1=Yes)
Num_pupils_GCSEs	Number of pupils at the end of key stage 4
Pct_Boys	Percentage of pupils at the end of key stage 4 who are boys
DISAD_pct_GCSEpup	Percentage of pupils at the end of key stage 4 who are disadvantaged
NONMOB_Pct_GCSEpup	Percentage of pupils at the end of key stage 4 who are non-mobile
Pct_GCSEpup_SEN	Percentage of pupils at the end of key stage 4 with special educational needs (SEN) with a statement or Education, health and care (EHC) plan
Prog8_overall_measure	Progress 8 measure after adjustment for extreme scores
Num_GCSEpup_all_Ebacc	Number of key stage 4 pupils with entries in all English Baccalaureate subject areas
Pct_GCSEpup_enter_Ebacc_Maths	Percentage of pupils entering the English Baccalaureate Maths subject area
Pct_GCSEpup_enter_Ebacc_Science	% of pupils entering the Ebacc Science subject area
Pct_GCSEpup_enter_Ebacc_Humanities	% of pupils entering the Ebacc Humanities subject area
Pct_GCSEpup_enter_Ebacc_Lang	% of pupils entering the Ebacc Language subject area
Pct_GCSEpup_any_qual	Percentage of pupils achieving any qualifications
Pct_GCSEpup_more_onelang	Percentage of pupils entering more than one language
Pct_GCSEpup_3_sciences	Percentage of pupils entering biology, chemistry and physics
Avg_num_GCSEs_equiv_pup	Average number of GCSE and equivalents entries per pupil
+ Prog8_band	Progress 8 banding shown on school performance tables website
Teachers_FTE	Total number of teachers (in school)
TAs_FTE	Total number of Teaching Assistants (in school)
Staff_non_classroom	Non-classroom staff (in school)
Pupil_Teacher_ratio	Pupil Teacher ratio for the school (Pupils divided by teachers)
Mean_Gross_FTE_Sal	Mean gross salary of FTE teacher in the school

What is **not** in the dataset...?

- ‘Raw’ academic performance
- Teacher and Head Teacher Quality
- Spending per school
- The impact of sitting exams with different Exam Boards
- Data on private schools’ performance: <https://www.isc.co.uk/research/>
- Broader, non-examined curriculum (sports, instruments)
- Other experiences children have at school (e.g. school plays)

...but I kept all outliers since each represented a school and its performance

Some initial visual analysis: Progress *is* related to raw attainment, and schools' value add seems to be broadly consistent across years



There are a number of risks and limitations

Risks

- ‘Off-rolling’ by schools
- Gaming by schools
- Smallest 30% schools (5% of children) excluded from data
- Exam boards
- Some children move between schools

Limitations

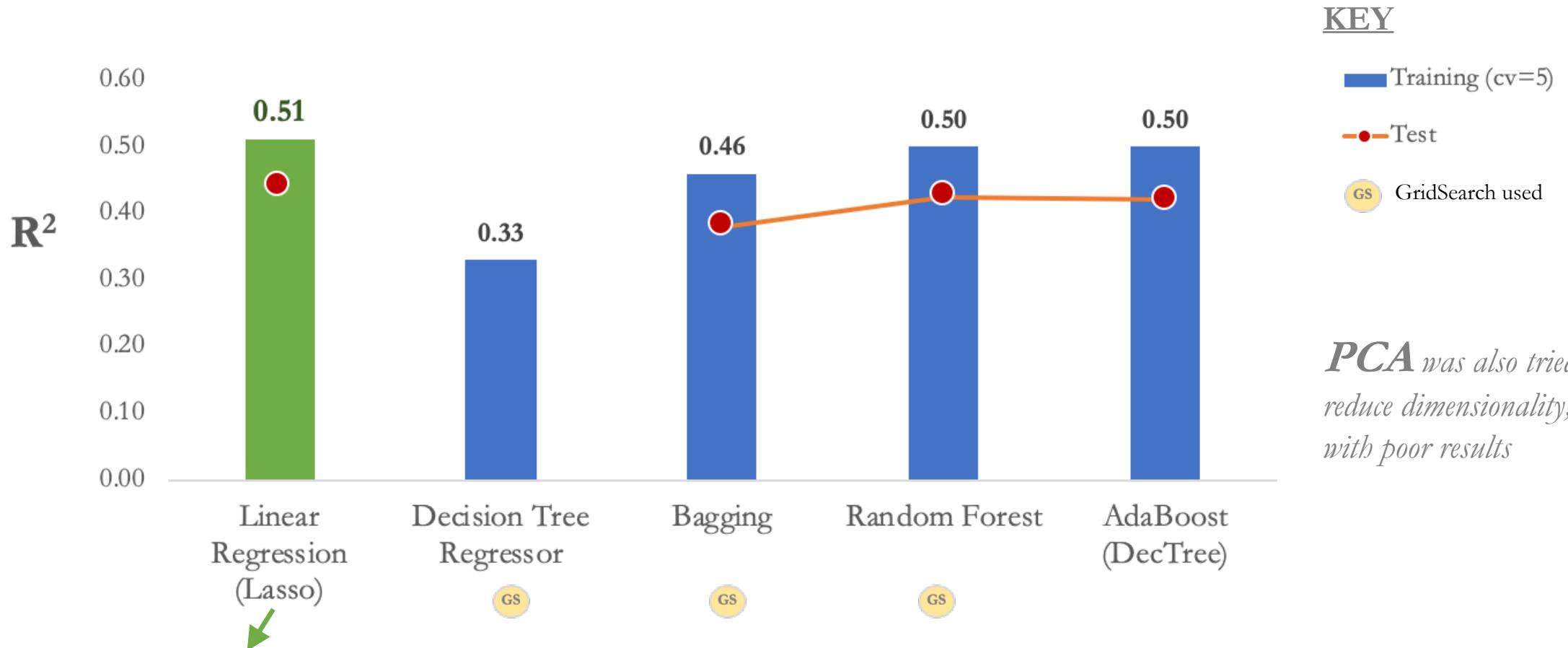
- Some key data isn’t available (teacher quality)
- Difficult to add more data without reducing the number of observations
- Difficult to compare across years (GOV.UK)

Assumptions

- Progress 8 is a good measure, worth pursuing
- Small schools behave similarly to others
- Drivers of progress are constant across school years and pupil cohorts

Modeling

Modeling Results



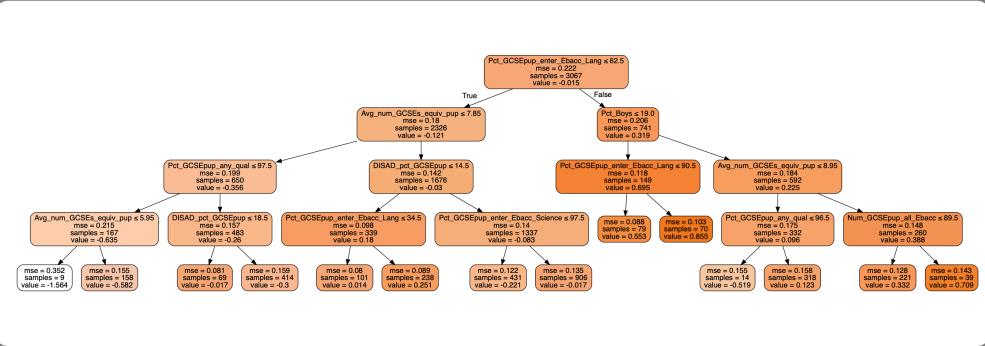
$y = -0.0143 + 0.086(\% \text{ children taking any language at GCSE}) + 0.086(\% \text{ children with any qualifications}) + 0.075(\text{average number of GCSEs and equivalents taken}) - 0.065(\% \text{ disadvantaged children}) - 0.048(\% \text{ boys in the school})$

Note: since the data is standardized, the numbers represent the shift for that factor for every standard deviation of the Progress 8 score

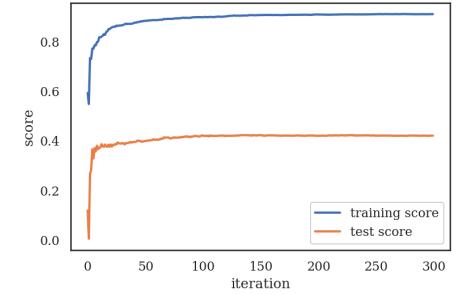
Some highlights from modeling...

**Decision Tree
Regressor had
poor R²**

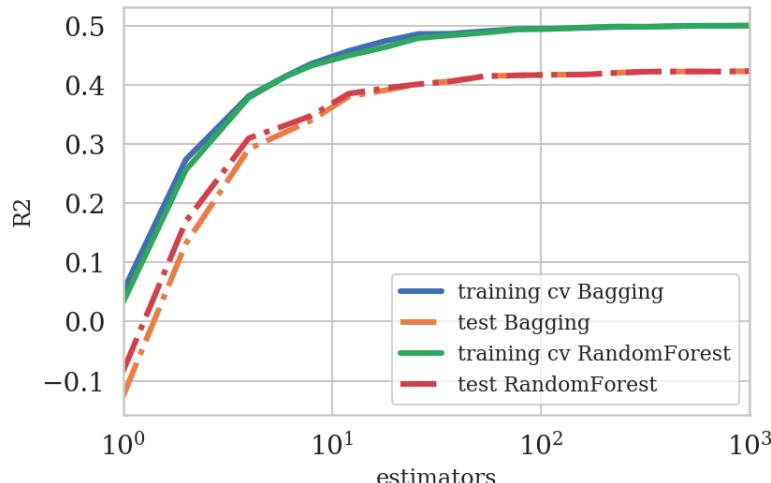
Optimum Depth = 14



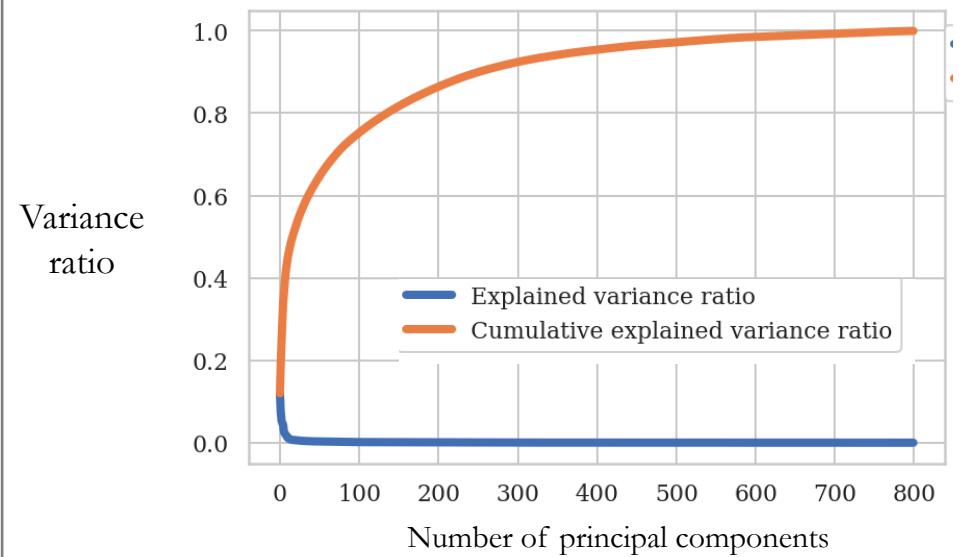
**Ada Boost
Decision
Tree**



**Ensemble of Bagging and Random Forest
reverted to the same R² as Lasso (0.51)
(winner was Random Forest with 1000 Estimators)**



**PCA did not reduce dimensions very
quickly**



...Linear Regression with regularization (Lasso) was the strongest model

Lasso and Ridge performed well...

(with non-regularized performing poorly, indicating continued autocorrelation)

LINEAR REGRESSION MODEL: Results

```
-----  
Training R2:          0.660  
Mean cross-val training R2: -2.75e+28  
Test R2:              -6.19e+24  
Intercept:            -0.0143
```

LASSO MODEL: Results

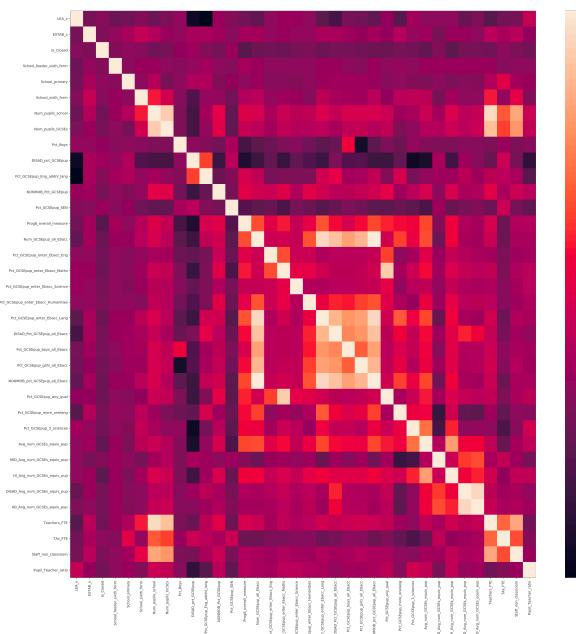
```
-----  
Best alpha:           0.008  
Training R2:          0.610  
Mean cross-val training R2: 0.514  
Test R2:              0.444  
Intercept:            -0.0143
```

RIDGE MODEL: Results

```
-----  
Best alpha:           1291.550  
Training R2:          0.667  
Mean cross-val training R2: 0.501  
Test R2:              0.411  
Intercept:            -0.0143
```

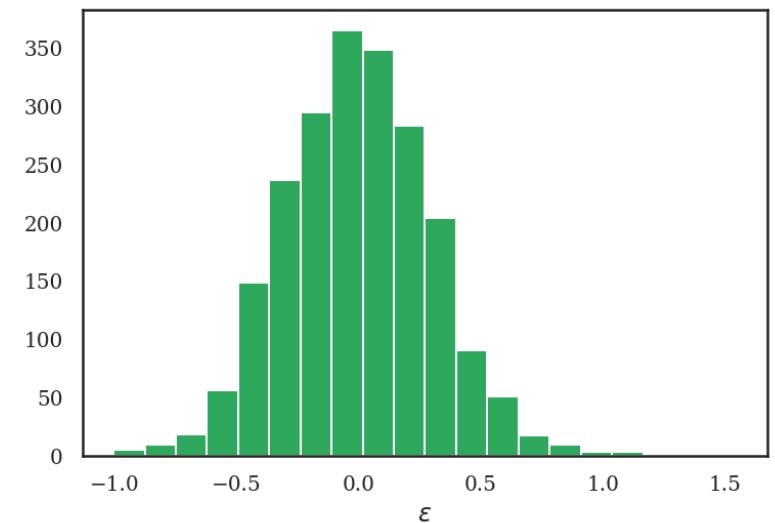
Mean Squared Error: 0.101

Root Mean Squared Error: 0.330



...with Residuals nice and (almost) Normal

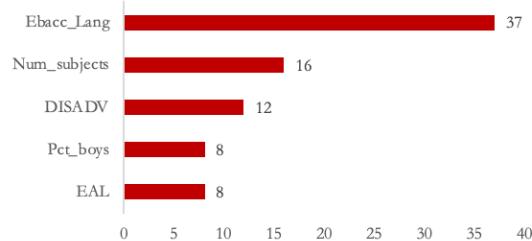
Residuals of Lasso Regression model (histogram)



5 factors came out most strongly in all models

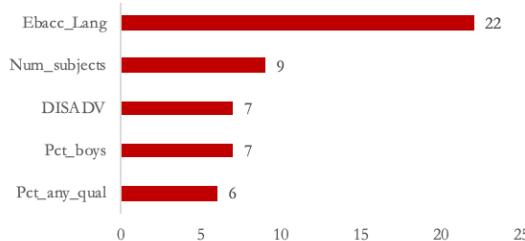
Decision Tree

Top 5 = 81%



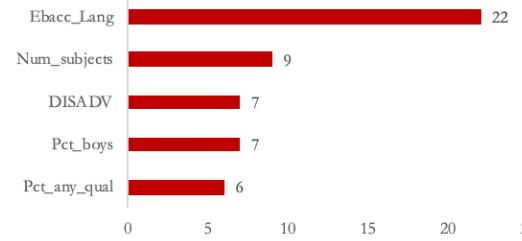
Ada Boost
Decision Tree

Top 5 = 51% *

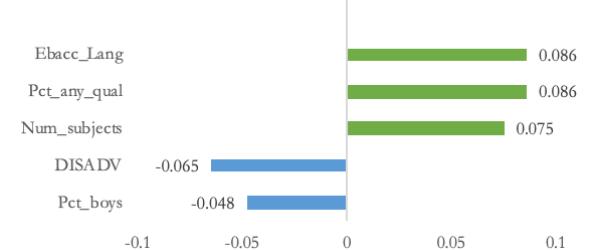


Random Forest

Top 5 = 51%



Linear Regression
(Lasso)



	feature	importance
7	Pct_GCSEpup_enter_Ebacc_Lang	0.368324
8	Pct_GCSEpup_any_qual	0.164744
1	Pct_Boys	0.115616
2	DISAD_pct_GCSEpup	0.084346
3	Num_GCSEpup_all_Ebacc	0.077683
784	Is_Closed_1.0	0.058307
5	Pct_GCSEpup_enter_Ebacc_Science	0.030368
791	School_type_2_UTC	0.020032
12	Teachers_FTE	0.016671
11	Avg_num_GCSEs_equiv_pup	0.016257
0	Num_pupils_GCSEs	0.011317
10	Pct_GCSEpup_3_sciences	0.007528
9	Pct_GCSEpup_more_onelang	0.007153
6	Pct_GCSEpup_enter_Ebacc_Humanities	0.006344
47	Local_Ed_Auth_320.0	0.004406

	features_importance
	Pct_GCSEpup_enter_Ebacc_Lang
	0.222860
	Avg_num_GCSEs_equiv_pup
	0.088295
	DISAD_pct_GCSEpup
	0.071203
	Pct_Boys
	0.068113
	Pct_GCSEpup_any_qual
	0.059236
	Num_GCSEpup_all_Ebacc
	0.038865
	Mean_Gross_FTE_Sal
	0.034184
	TAs_FTE
	0.032741
	Pct_GCSEpup_enter_Ebacc_Science
	0.031657
	Teachers_FTE
	0.031444
	Pct_GCSEpup_enter_Ebacc_Humanities
	0.029752
	Pct_GCSEpup_3_sciences
	0.027962
	Staff_non_classroom
	0.026576
	Num_pupils_GCSEs
	0.026405
	Pupil_Teacher_ratio
	0.024524

	features_importance
	Pct_GCSEpup_enter_Ebacc_Lang
	0.221159
	Avg_num_GCSEs_equiv_pup
	0.087681
	Pct_Boys
	0.069912
	DISAD_pct_GCSEpup
	0.069621
	Pct_GCSEpup_any_qual
	0.061304
	Num_GCSEpup_all_Ebacc
	0.044985
	Mean_Gross_FTE_Sal
	0.033402
	Teachers_FTE
	0.032253
	TAs_FTE
	0.031647
	Pct_GCSEpup_enter_Ebacc_Science
	0.031341
	Pct_GCSEpup_enter_Ebacc_Humanities
	0.030809
	Pct_GCSEpup_3_sciences
	0.027566
	Num_pupils_GCSEs
	0.027100
	Staff_non_classroom
	0.025741
	Pct_GCSEpup_more_onelang
	0.024866

	variable	lasso_coef	abs_lasso_coef
7	Pct_GCSEpup_enter_Ebacc_Lang	0.087034	0.087034
8	Pct_GCSEpup_any_qual	0.085870	0.085870
11	Avg_num_GCSEs_equiv_pup	0.074584	0.074584
2	DISAD_pct_GCSEpup	-0.065384	0.065384
1	Pct_Boys	-0.047618	0.047618
811	School_gender_MIXED	-0.037537	0.037537
802	Rel_Denom_Muslim	0.037165	0.037165
791	School_type_2_UTC	-0.033264	0.033264
9	Pct_GCSEpup_more_onelang	0.028121	0.028121
12	Teachers_FTE	0.025314	0.025314
29	Local_Ed_Auth_302.0	0.024960	0.024960
784	Is_Closed_1.0	-0.024858	0.024858
34	Local_Ed_Auth_307.0	0.023501	0.023501
44	Local_Ed_Auth_317.0	0.023379	0.023379
810	School_gender_GIRLS	0.022730	0.022730
31	Local_Ed_Auth_304.0	0.022343	0.022343

So what *does* explain progress?

Five key drivers explain 50% of the variation in the progress than a child makes

1

Taking any **language** at GCSE is an indicator of greater progression

2

Achieving **at least one GCSE** is important

3

The **more GCSEs** (and equivalent) qualifications that children do, the more they progress

4

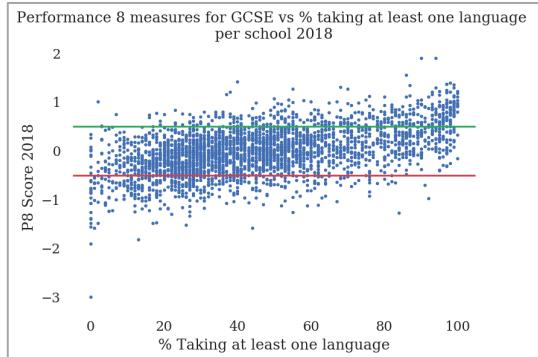
Coming from a **disadvantaged** background hampers progress

5

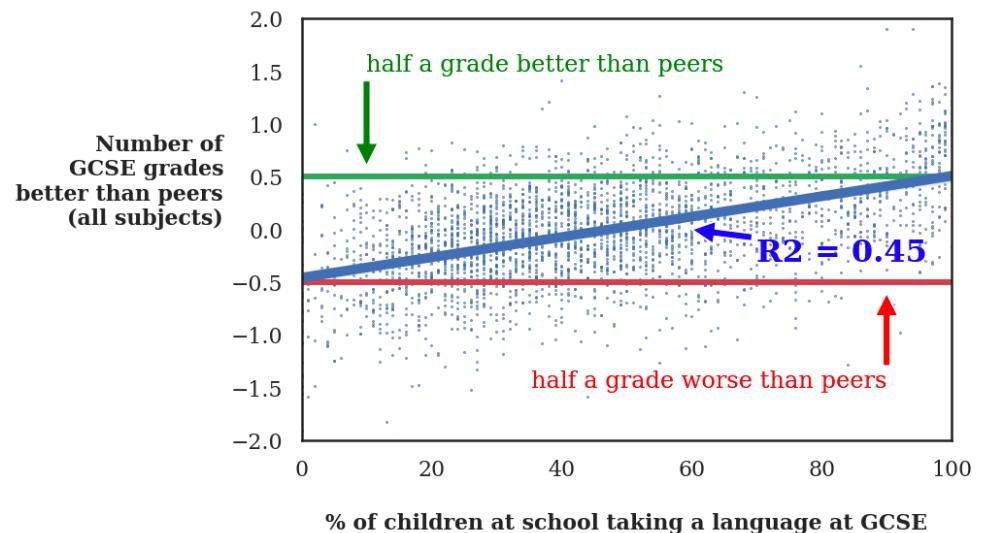
Girls progress faster than boys from 11 to 16 years old

1

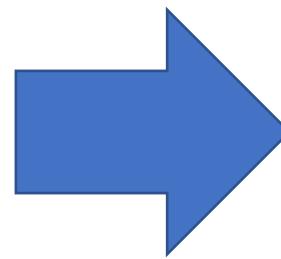
Taking a **language** at GCSE is an indicator of greater progression



Children at schools which take languages progress more in all subjects



47% (234k) of children take at least one language



EBACC all entries	
Science	62%
Humanities	24%
Languages	14%

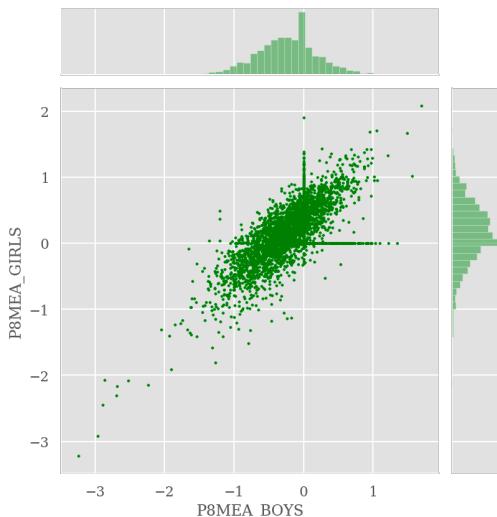
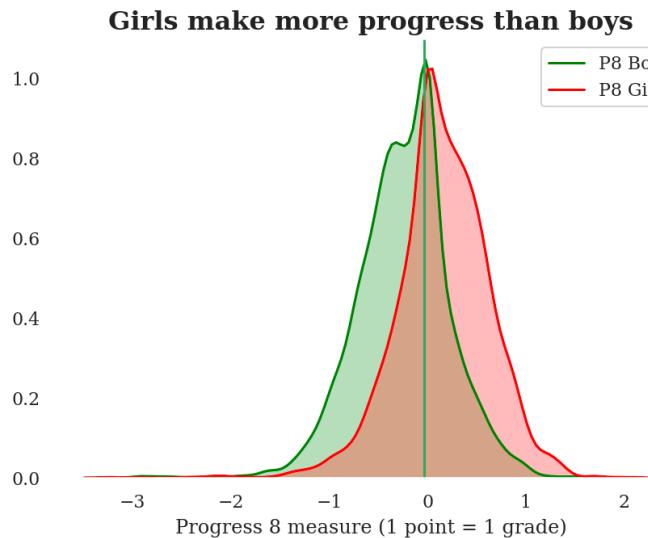
Action: conduct further research to try to understand why this is so

Hypotheses:

1. Only the most able, in schools with strong language teachers, are steered towards taking a languages GCSE
2. The type of person who studies a language is someone who will achieve particularly well
3. The act of learning a language enables the development of more advanced cognitive functioning than learning other subjects
4. The quality of teaching has the biggest impact, and high quality teachers work with other high quality teachers
5. There is another confounding factor(s) that is yet to be discovered

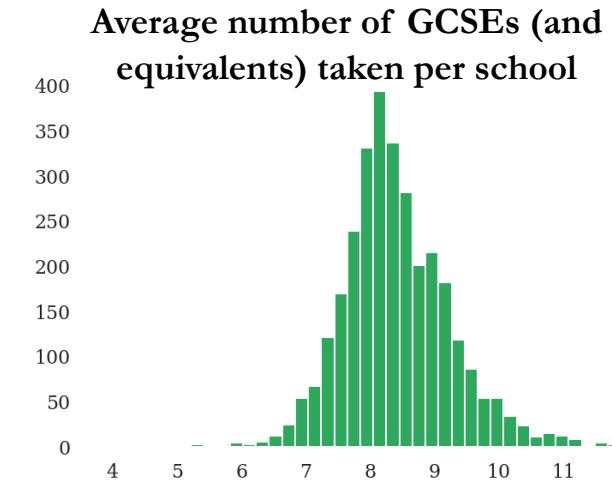
5

Girls progress faster than boys from 11 to 16 years old

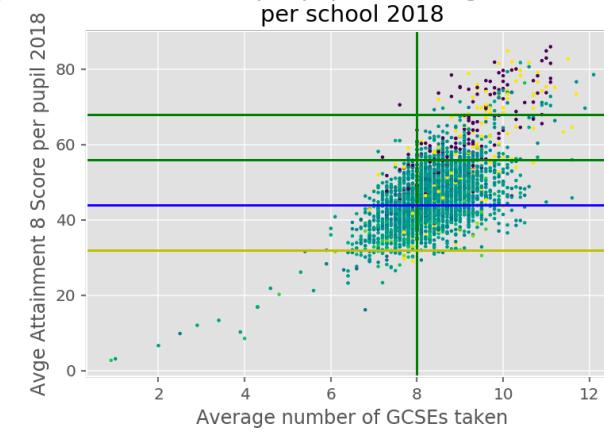


3

The **more GCSEs** (and equivalent) qualifications that children do, the more they progress

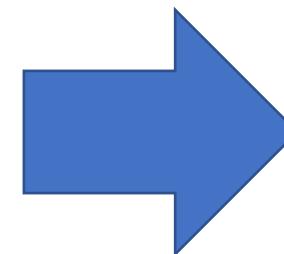
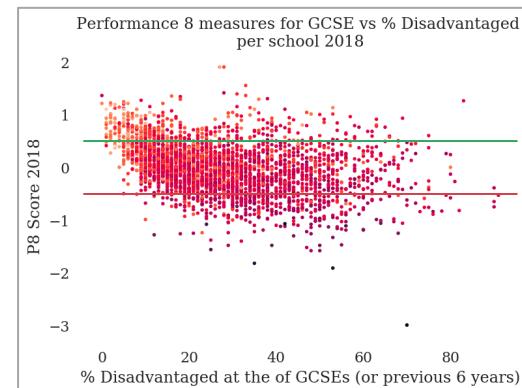
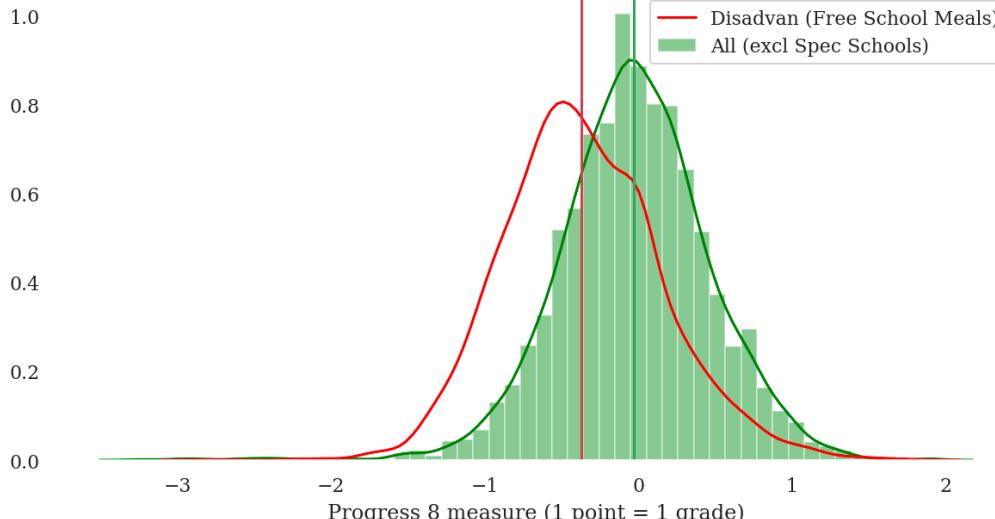


Average Attainment 8 Score per pupil vs Average number of GCSEs taken per school 2018

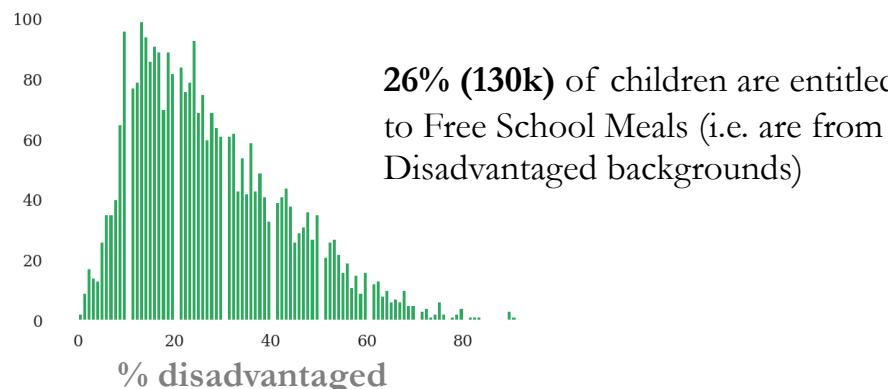


Coming from a **disadvantaged** background hampers progress

Disadvantaged children progress less (on average one third of a grade)



Action: review the impact of 'Pupil Premium' since 2011 and either change policy or boost spending



Model: MODEL_10_Just_Cat_REDUCED-LOGs (end)

<https://www.suttontrust.com/>

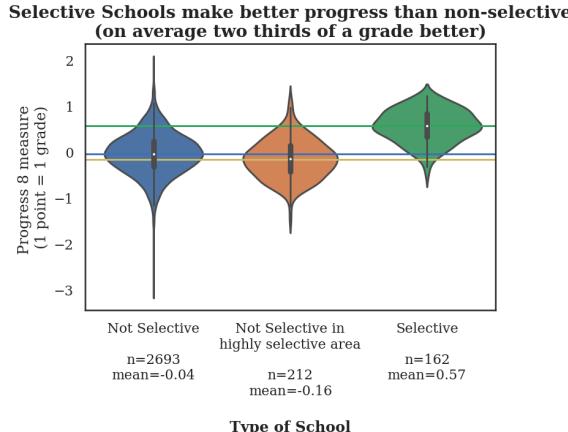
<https://www.gov.uk/guidance/pupil-premium-information-for-schools-and-alternative-provision-settings>



There were a few surprises too that showed up in the analysis...

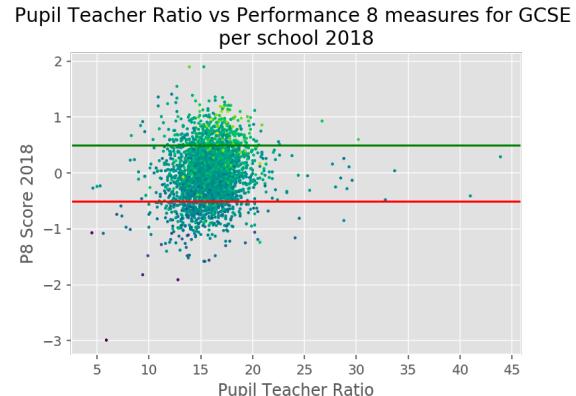
Selective Schools

Children make 2/3 grade more progress than in non-Selective Schools



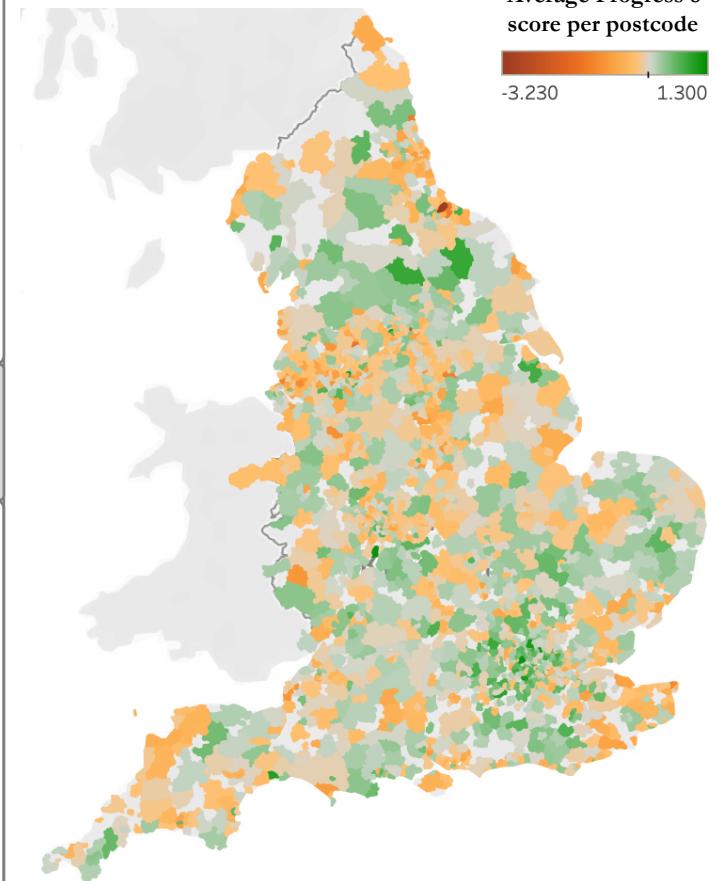
Pupil Teacher Ratio

Not a significant impact on progress made



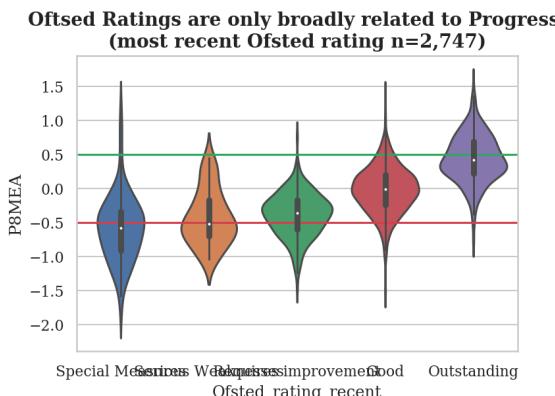
Geographic location did not itself explain progress

Average Progress 8 score per postcode
-3.230 1.300



Government Inspection (Ofsted)

Most recent rating was not a big predictor of progress, with very varied results across categories



Establishments (groups of Academies)

May be too early to see their impact?



Recommendations

• • • assuming that Progress 8 remains a priority

- 1) Commission research into determining whether learning languages strengthens a child's overall cognitive development, and if so, return to making languages compulsory at GCSE
- 2) Review the effectiveness of the 'Pupil Premium' funding that was introduced in 2011, and either boost that funding or change the policy for improving the progress that children make
- 3) Examine why children in Selective Schools make more progress than their peers in non-Selective schools
- 4) Encourage schools to offer as broad an education as possible since children who take more GCSEs progress more

Given that the models explained approximately 50% of the variation in progress that children made, there is **one further hypothesis that I believe needs to be explored further: that teacher and Head Teacher quality in schools is the primary driver of the amount of progress that children make.**

However, before these recommendations are implemented, the data for other school years needs to be analysed, to ascertain whether the explanatory factors for those years are similar to 2017-18.

Main learnings (so far)

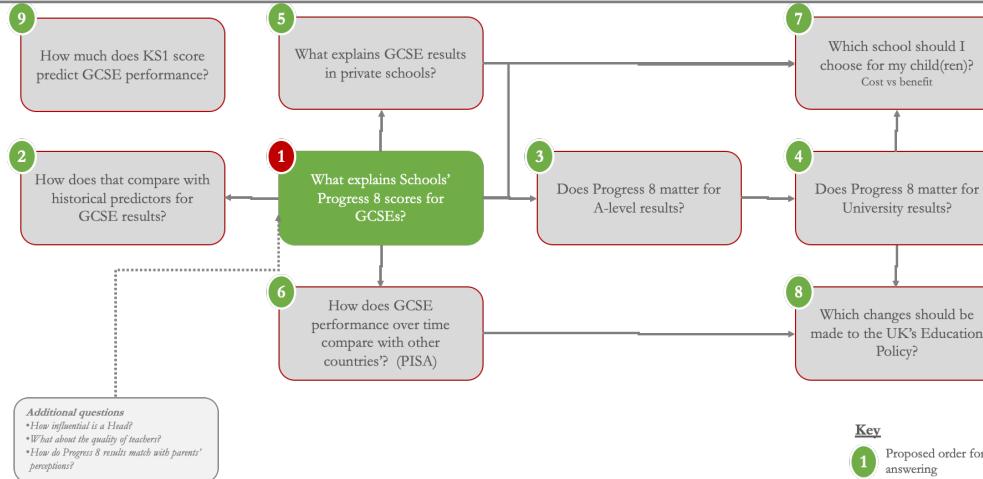
- 80% of time on EDA is about right...but it's also super important
- Considerable amount of subjectivity: domain knowledge is key
- Data science is an **iterative** process:
 - Need to prioritise
 - Deliver in increments, and
 - Establish when it's time to stop
- Sometimes the simplest model is the best
(0.5 isn't bad considering data excluded)
- In addition there have been many technical learnings

Next steps

Current Project and dataset (in priority order)

1. Investigate findings further (especially languages)
2. Add new features to the dataset (without reducing number of schools in it)
3. Try additional models to see whether they provide different results (k-NN, SVM, Bayesian)
4. Try classification approach (logistic regression) to see how accurately schools' P8 can be predicted
5. PCA: try sub-groupings of variables to see whether they can be reduced to fewer dimensions

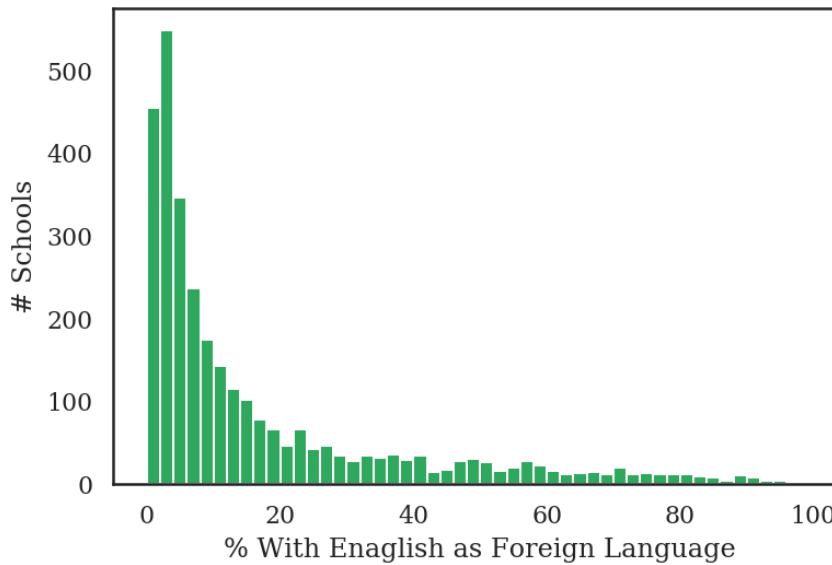
Expand the scope



Thank you

Appendices

Pupils with **English as an additional language** (EAL) progress more



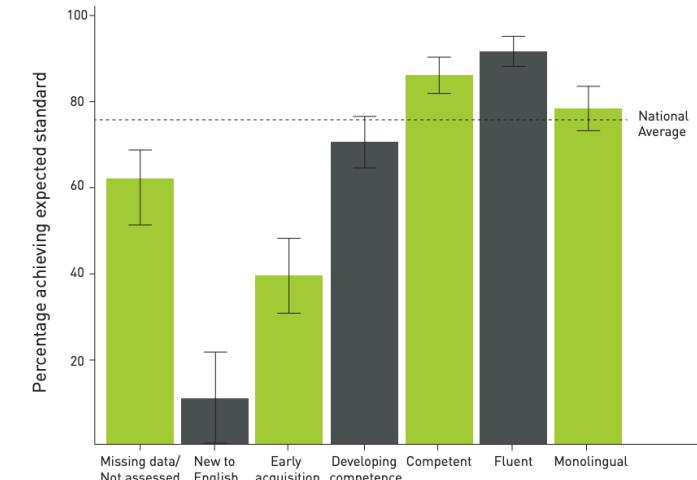
16% (81k) of children are entitled to Free School Meals (i.e. are from Disadvantaged backgrounds)

“Bilingualism can have positive associations with achievement”

<https://www.bell-foundation.org.uk/wp-content/uploads/2018/10/EAL-PIE-and-Educational-Achievement-Report-2018-FV.pdf>

“English proficiency can explain up to 22% of the variation in EAL pupils’ achievement”

Age 7 reading achievement by proficiency in English

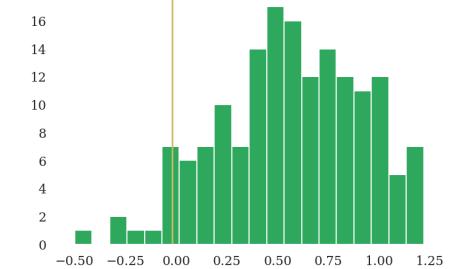
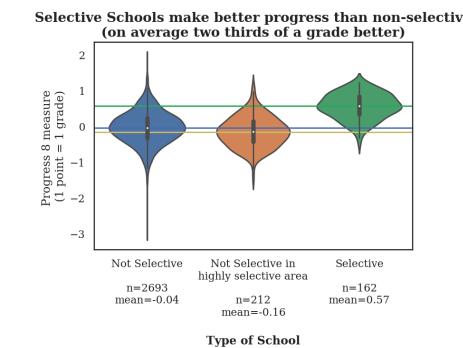
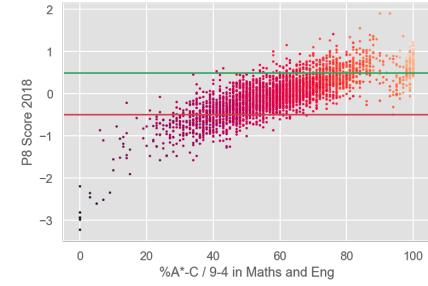
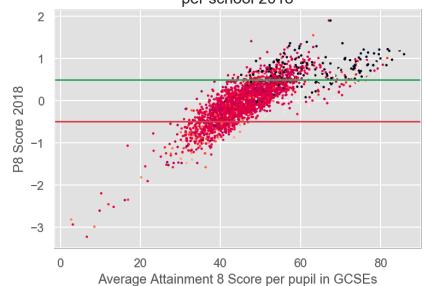
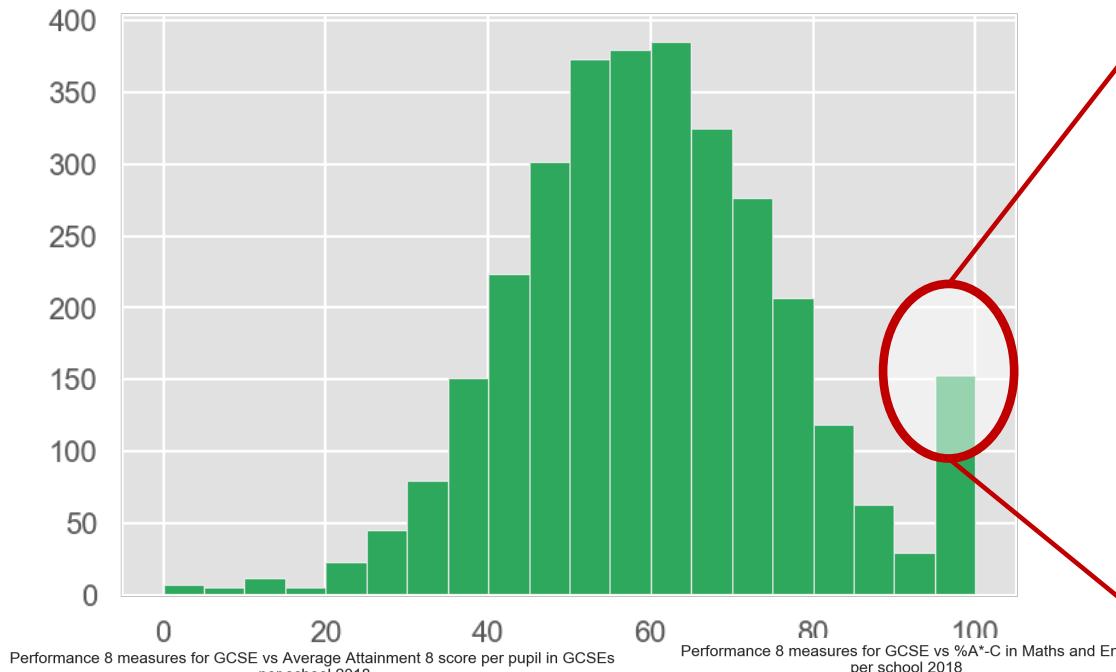


Action: review the October 2018 Bell Foundation report

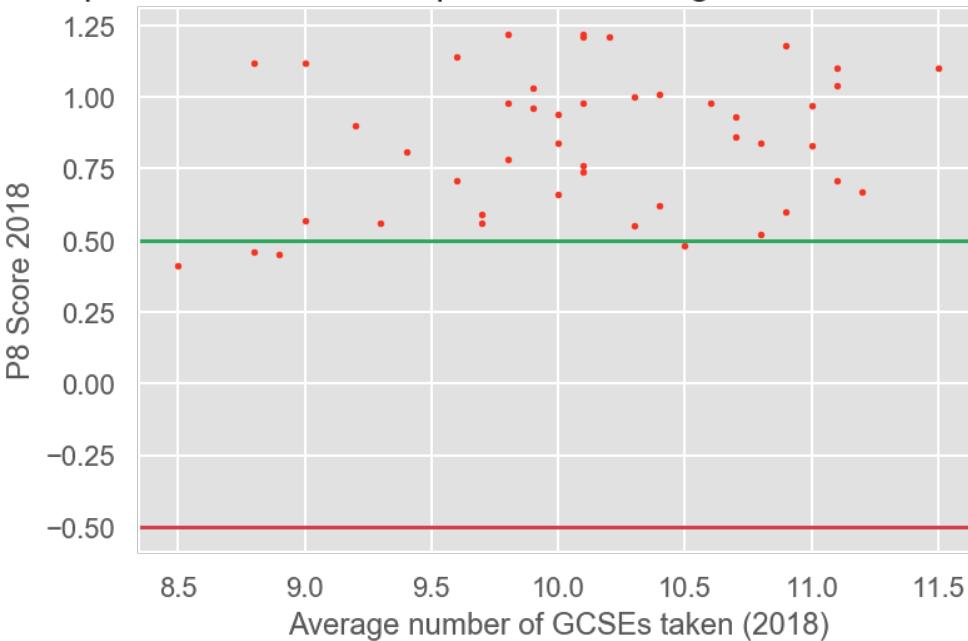
- Report into EAL in education found that proficiency proficiency in the language of instruction is key.
- Once children are competent or fluent in English, they outperform other children.
- One of the research's findings was that the proficiency in English of children increased considerably between 11 and 16 years of age
- English language support is most needed in the early years and KS1

Selective Schools

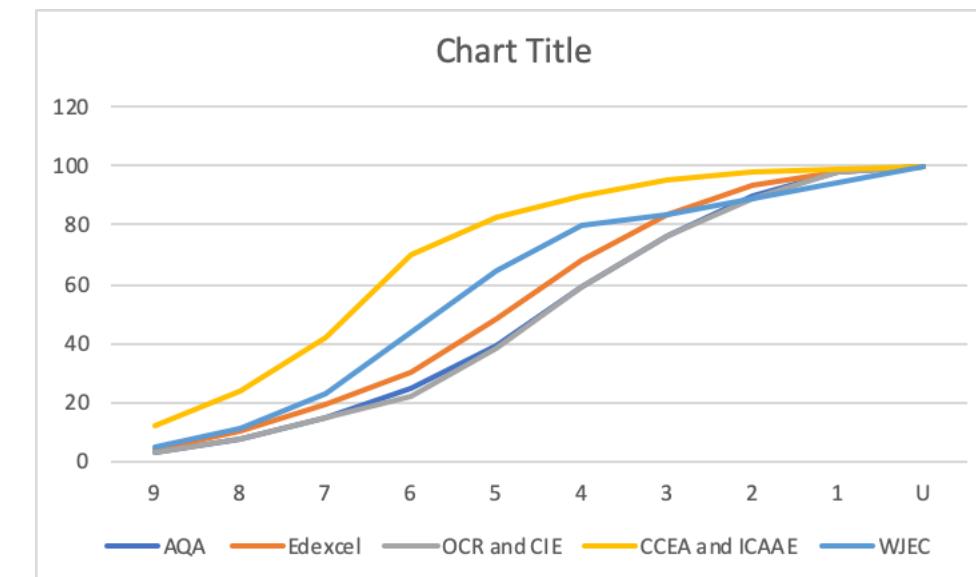
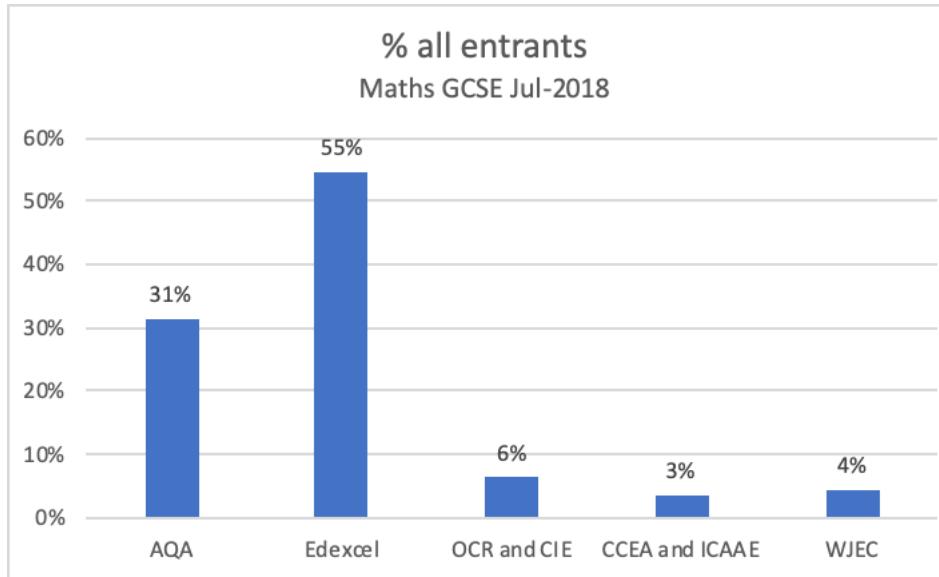
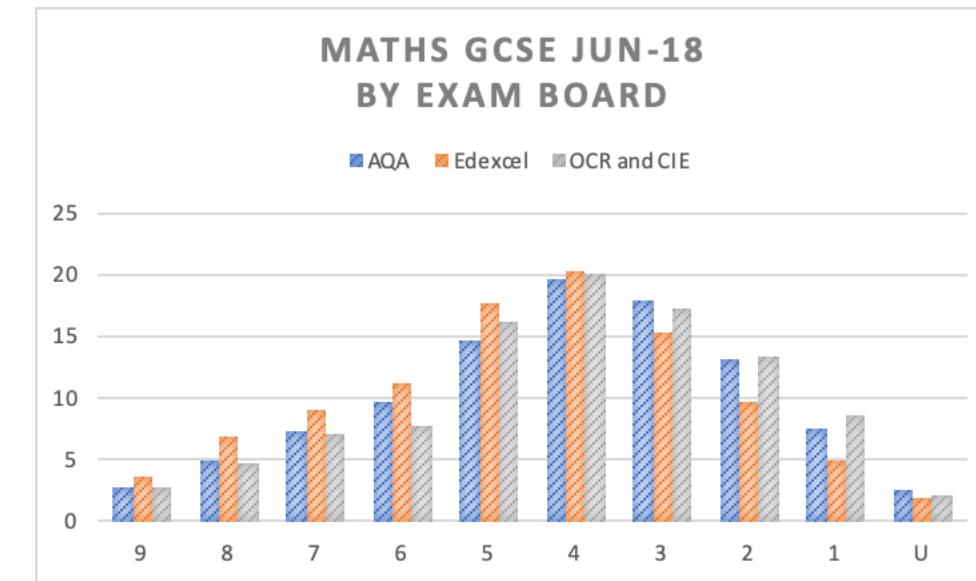
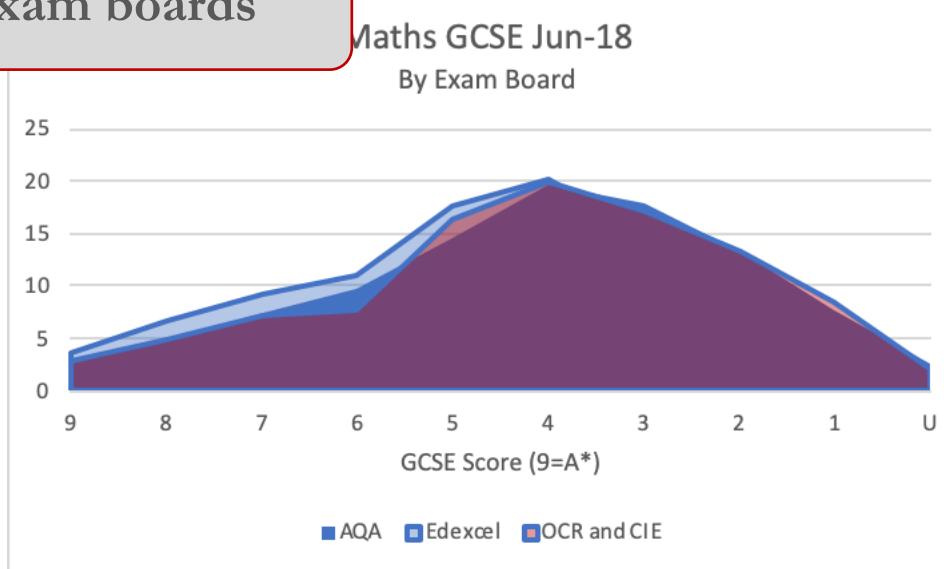
% getting A*-C in English and Maths GCSEs



Performance 8 measure for GCSE vs Average number of GCSEs taken per school with 100% pass rate for English and Maths 2018



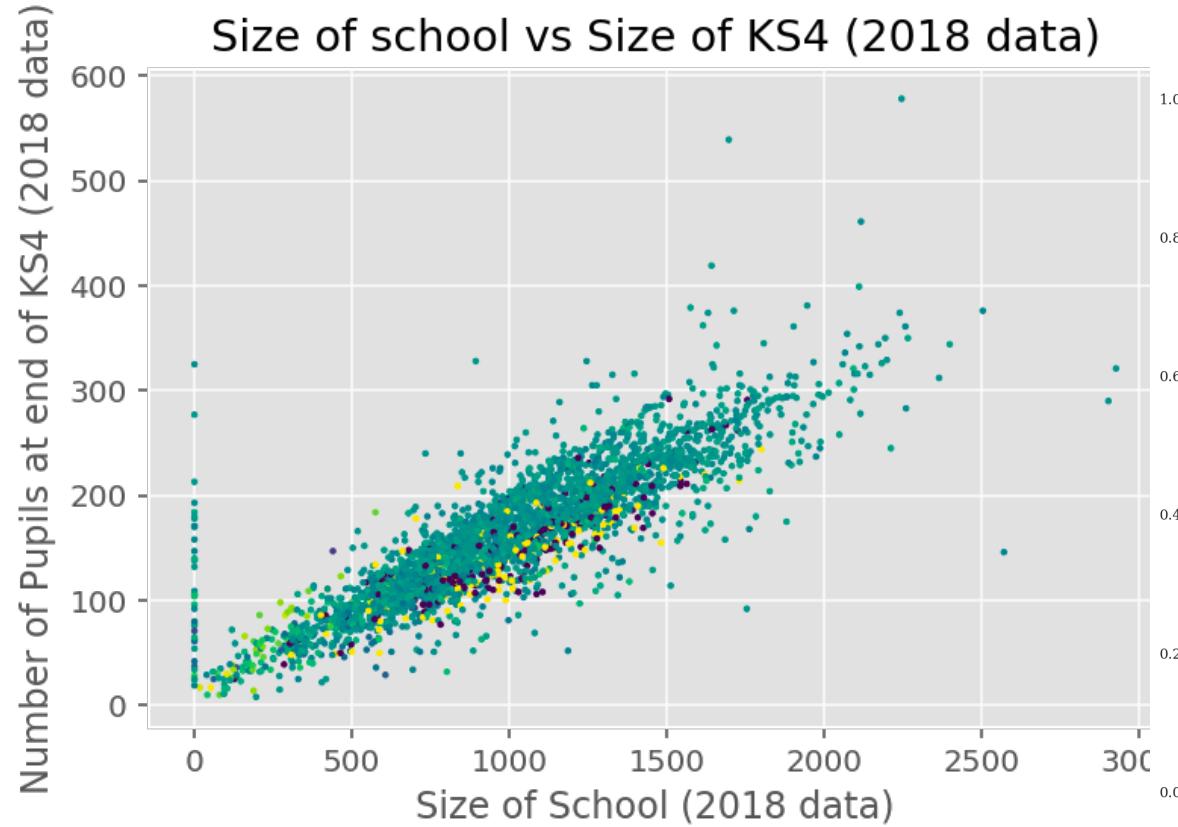
Exam boards



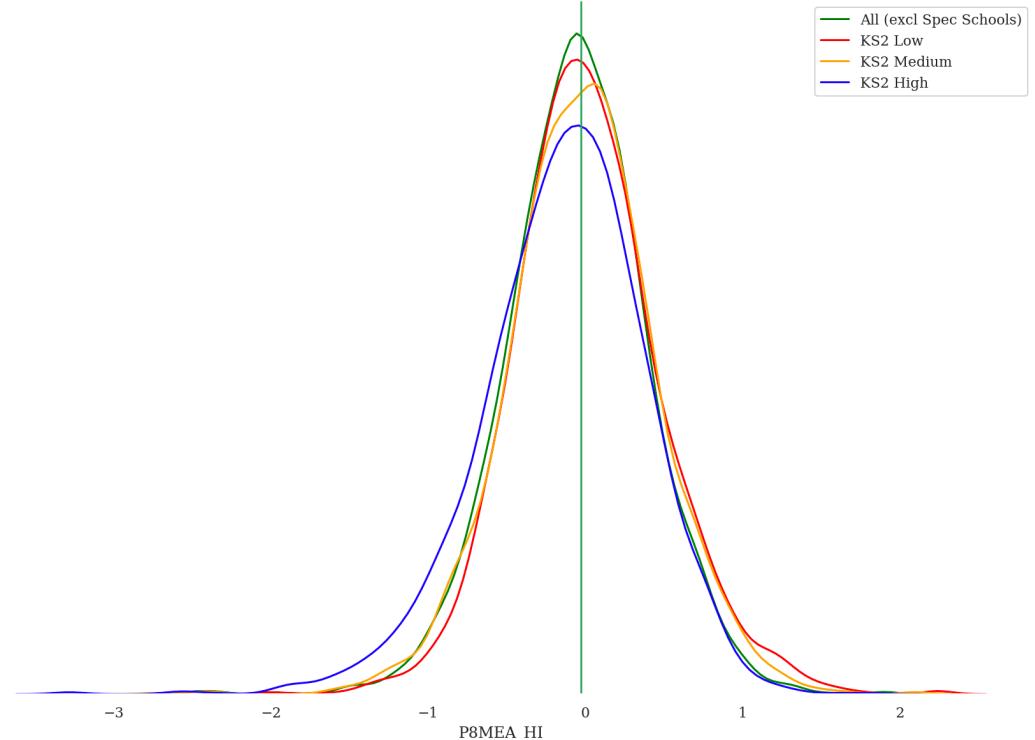
Note: CCES and WJEC are on old GCSE grading scale (A* - G and U)

Off-rolling...?

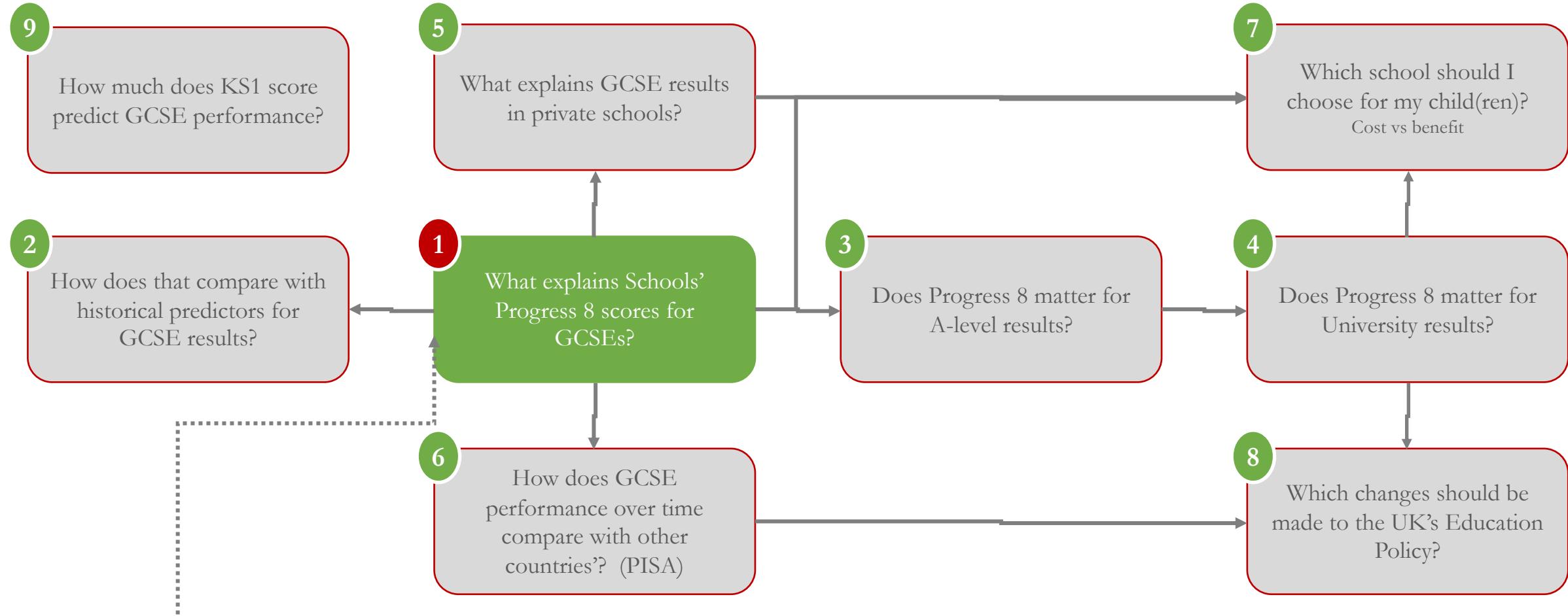
(bottom right of graph)



High achievers progress a bit less than their peers between 11 and 16 years old
(n = 2,843 because of lack of low attainment data...)



Future focus to broaden this project



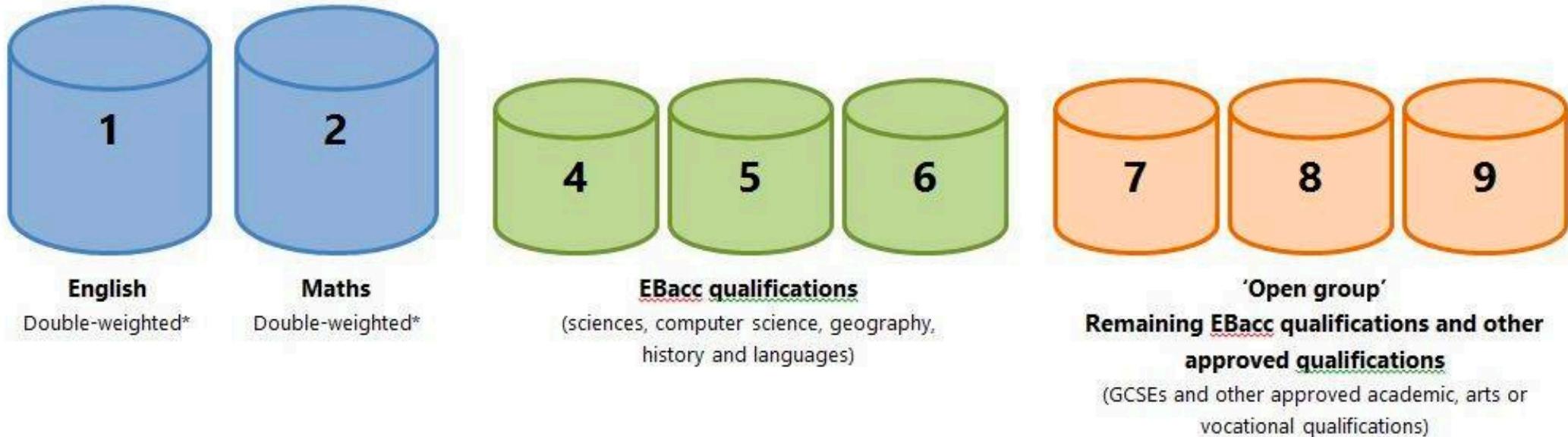
Additional questions

- How influential is a Head?
- What about the quality of teachers?
- How do Progress 8 results match with parents' perceptions?

Key

- 1 Proposed order for answering

Progress 8 measure



*Higher score of English Language OR English Literature
double-weighted if a student has taken both qualifications

Source:

<https://edcentral.uk/edbloggerreader/progess-8-and-attainment-8-two-new-accountability-measures>

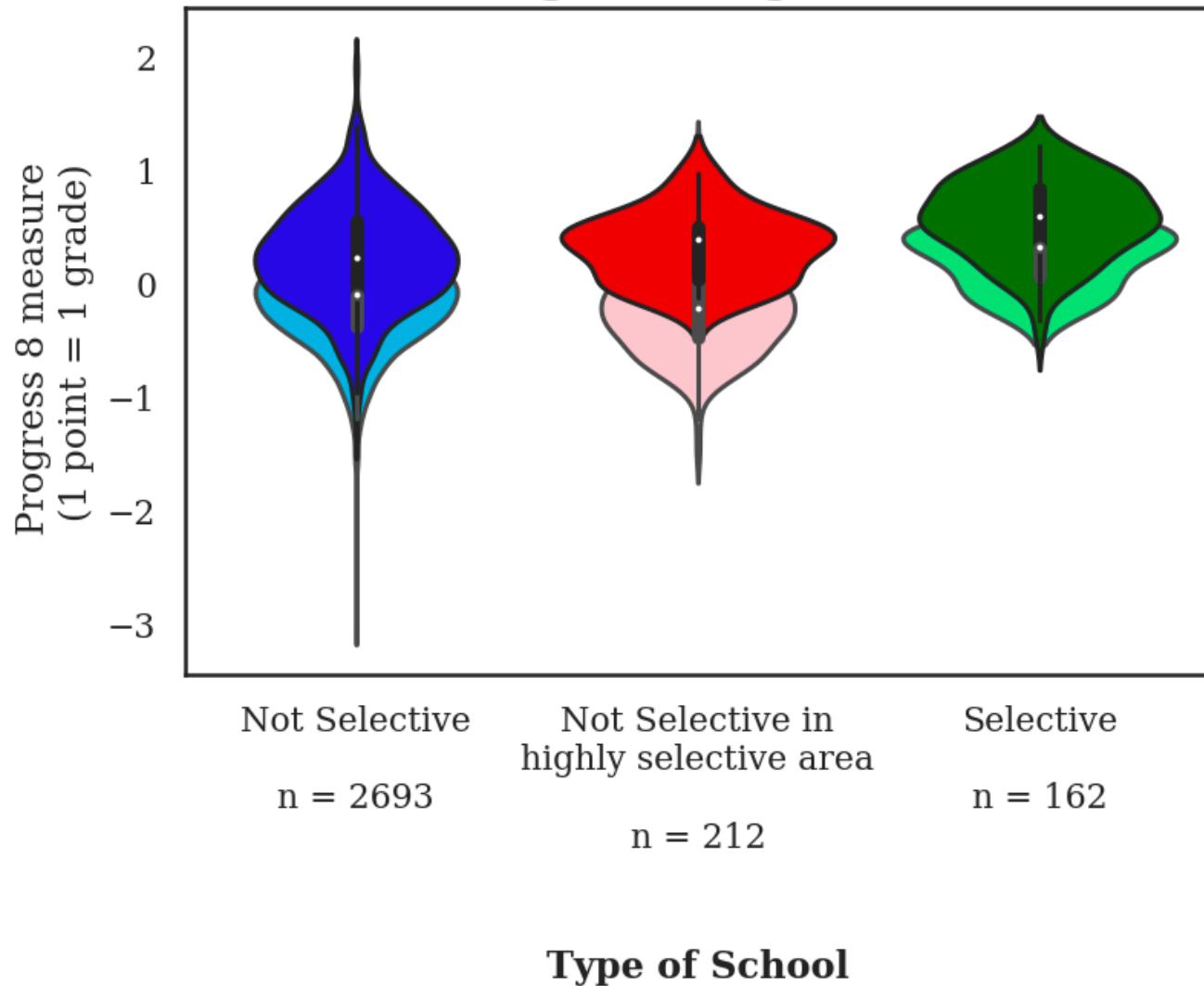
How we calculated the bandings for 2017 and 2016¹

Progress description	GCSE approved qualifications, a school/college was given the description if...	
	2017 ²	2016
Well above average	About 12% of schools/colleges in England Score is 0.50 or higher, and lower confidence interval limit is higher than 0.	About 5% of schools/colleges in England Score is 0.50 or higher, and lower confidence interval limit is higher than 0.
Above average	About 18% of schools/colleges in England Score is higher than 0 but lower than 0.5, and lower confidence interval limit is higher than 0.	About 25% of schools/colleges in England Score is higher than 0 but lower than 0.5, and lower confidence interval limit is higher than 0.
Average	About 40% of schools/colleges in England Lower confidence interval limit is 0 or lower, and the upper confidence interval limit is 0 or higher.	About 40% of schools/colleges in England Lower confidence interval limit is 0 or lower, and the upper confidence interval limit is 0 or higher.
Below average	About 18% of schools/colleges in England Score is -0.5 or higher but less than 0, and upper confidence interval limit is lower than 0	About 20% of schools/colleges in England Score is -0.5 or higher but less than 0, and upper confidence interval limit is lower than 0
Well below average	About 12% of schools/colleges in England Score is lower than -0.5 and upper confidence interval limit is lower than 0.	About 10% of schools/colleges in England Score is lower than -0.5 and upper confidence interval limit is lower than 0.

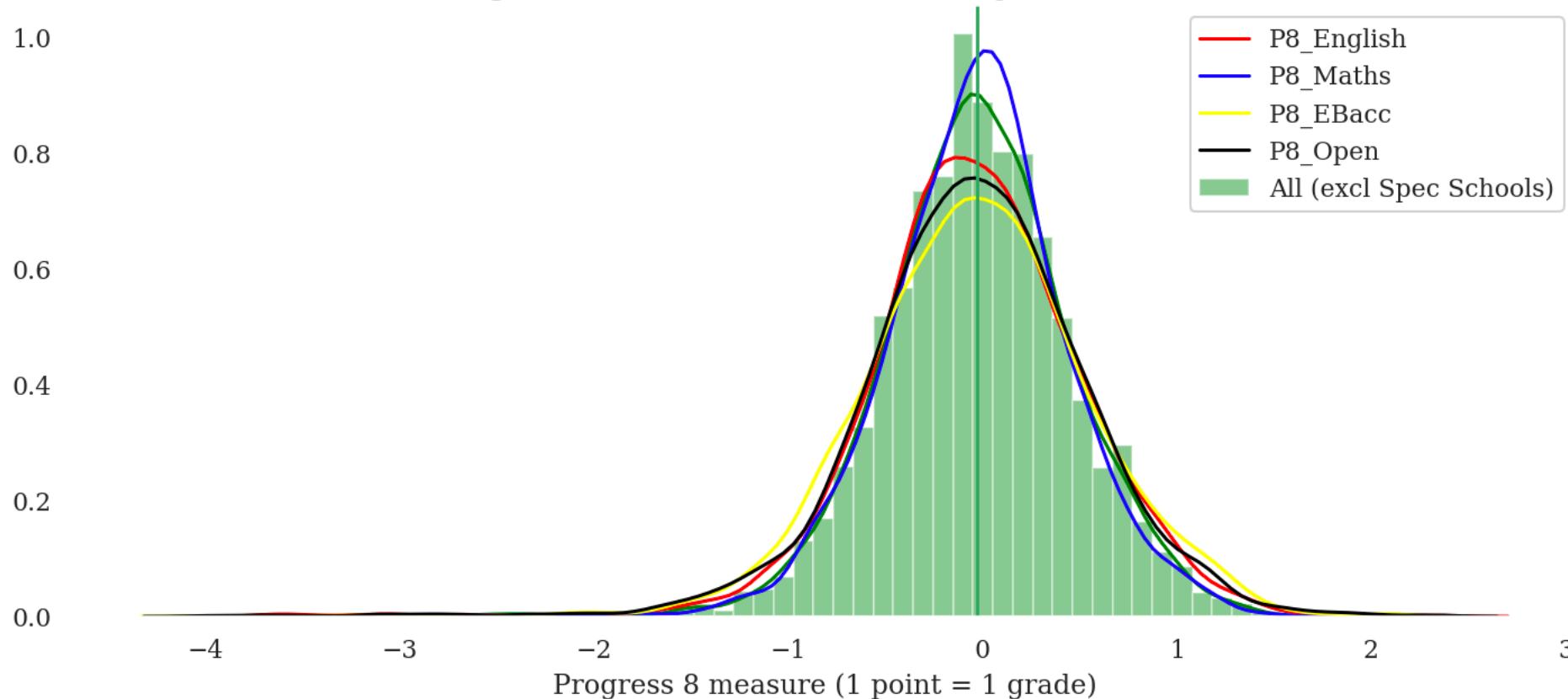
Source:

chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/676173/Key_stage_4_progress_bandings.pdf

**Children at schools with more than two thirds
of children taking a language
make better progress
(on average half a grade better)**

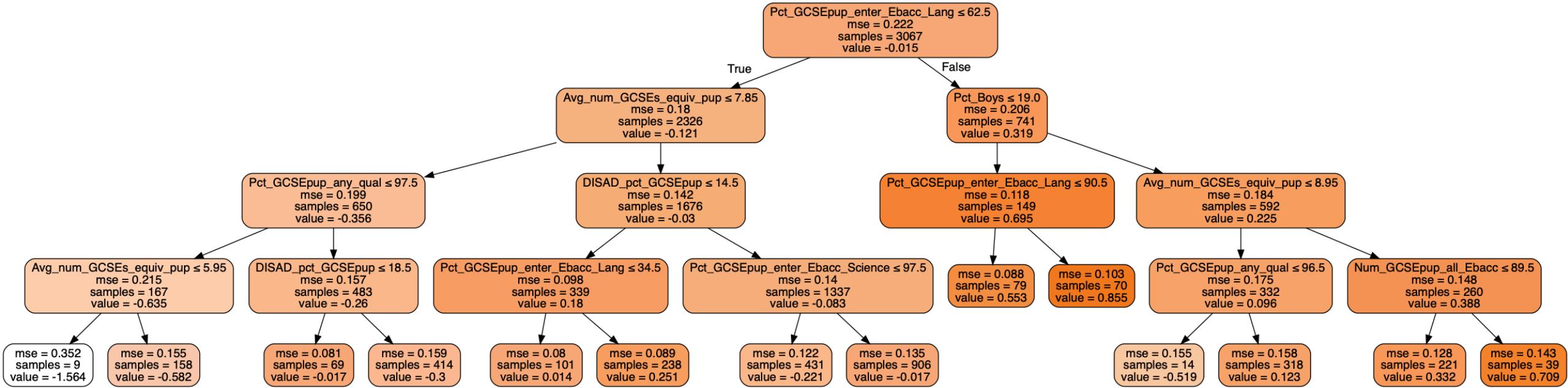


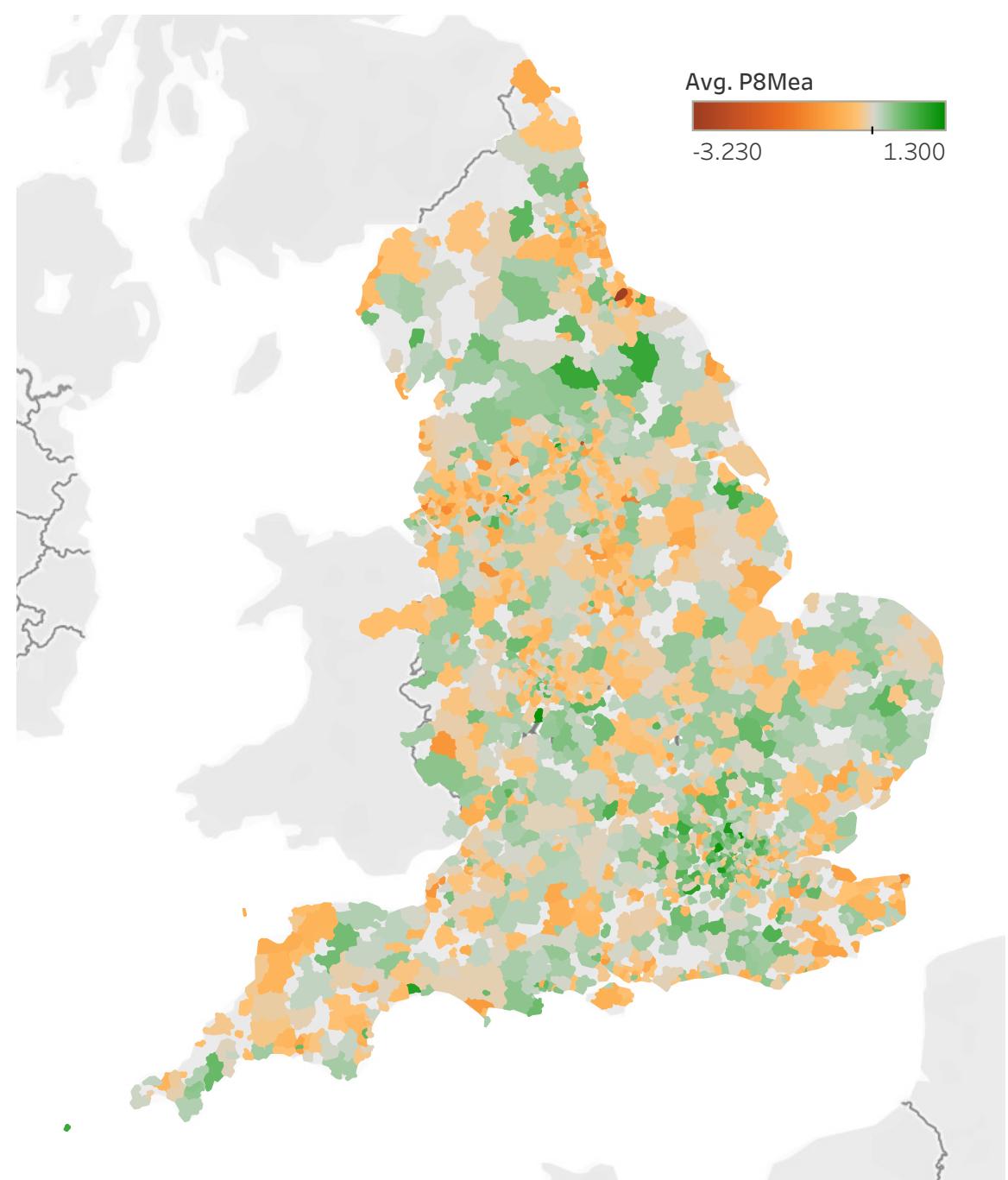
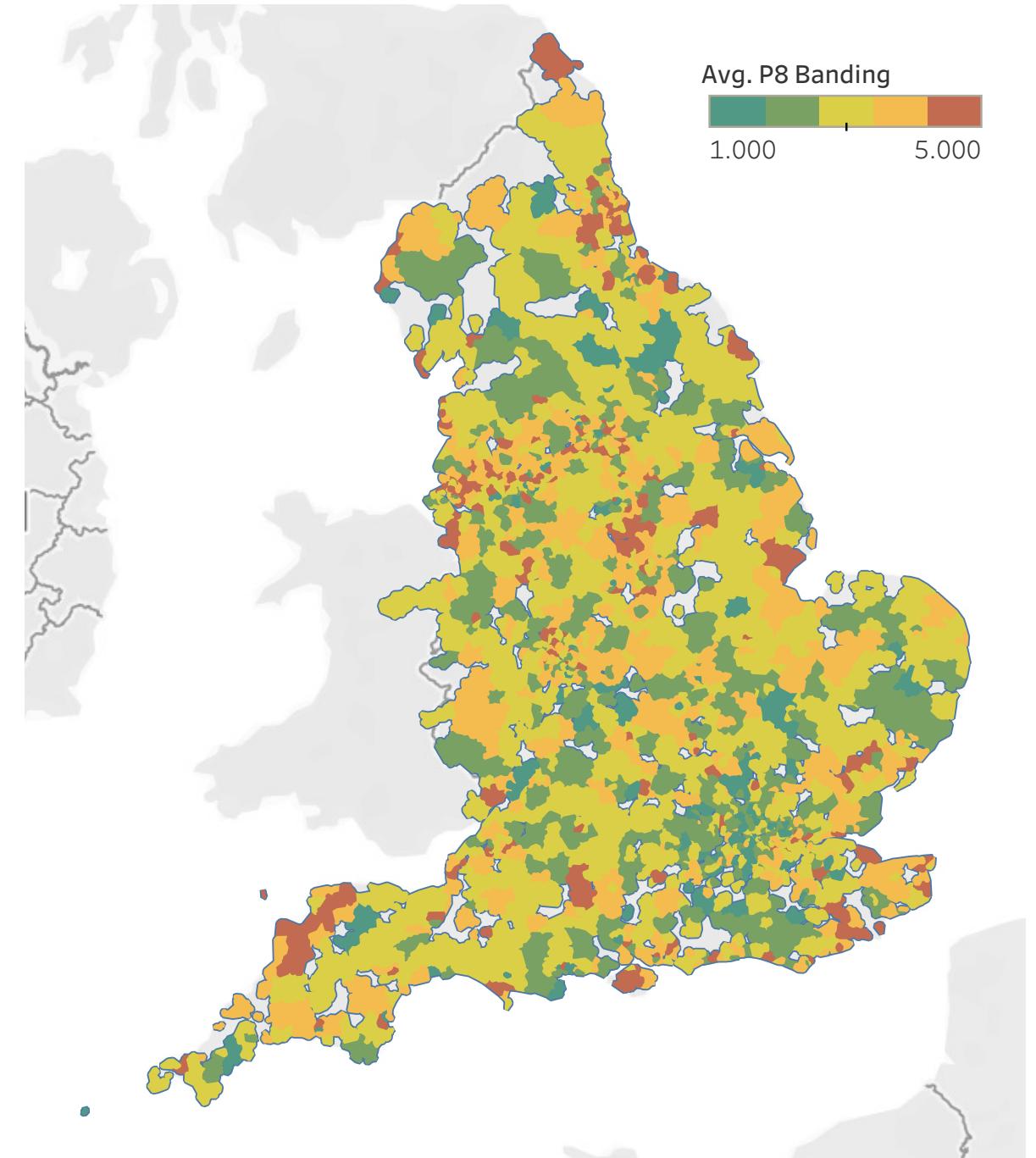
P8 scores across subject buckets are relatively similar across all schools...



Decision Tree

First 5 layers (optimum depth is 14 layers; only 5 shown here for visibility)

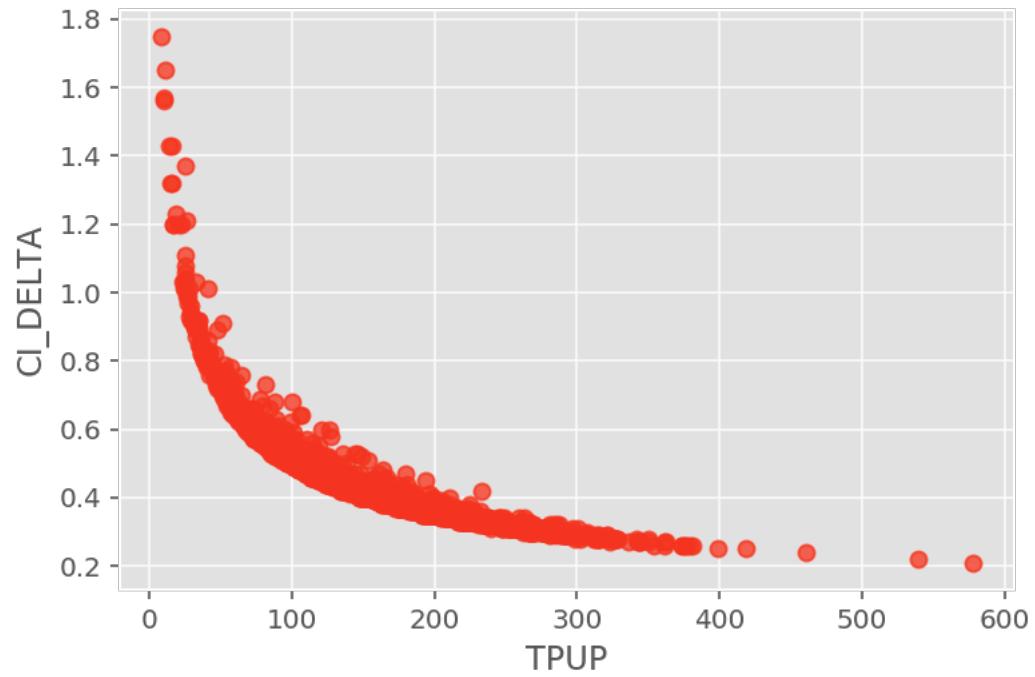
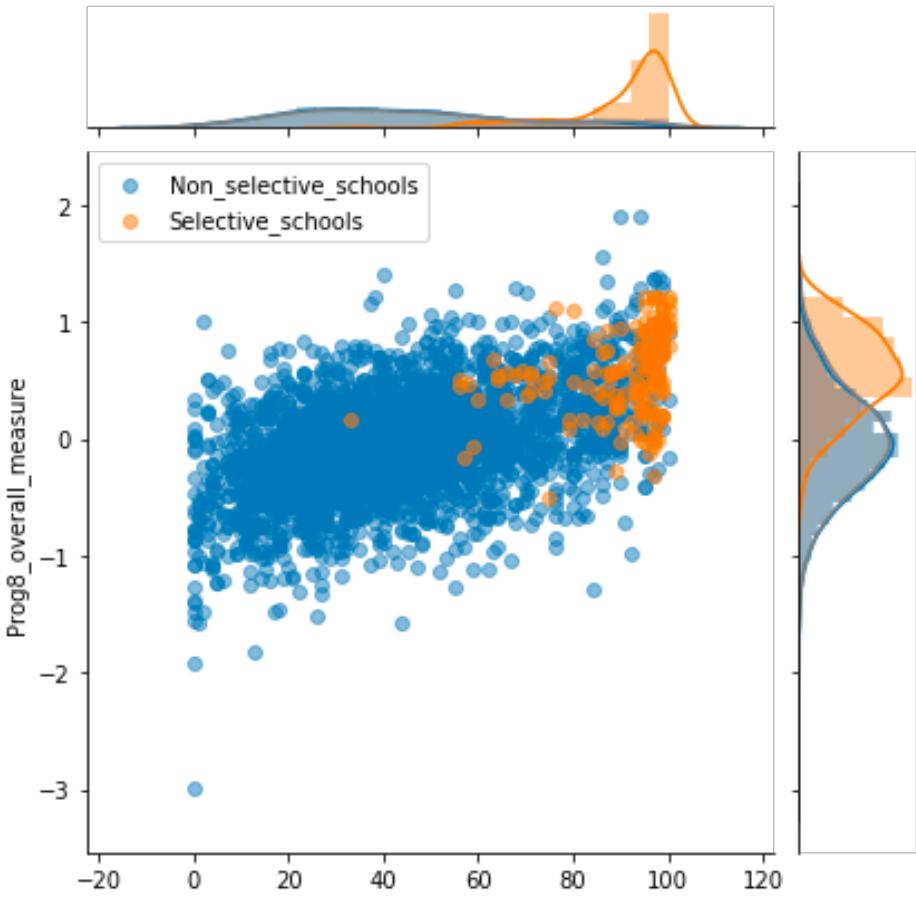




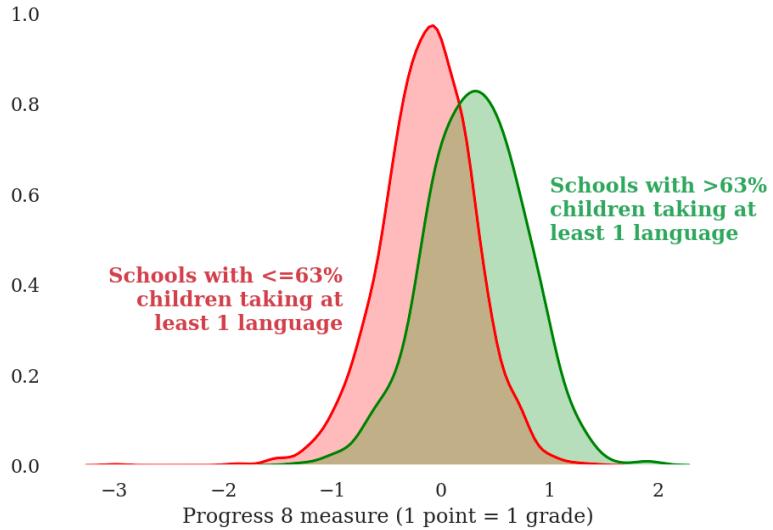
London: Progress 8 score per postcode

Size of bubble represents number of schools

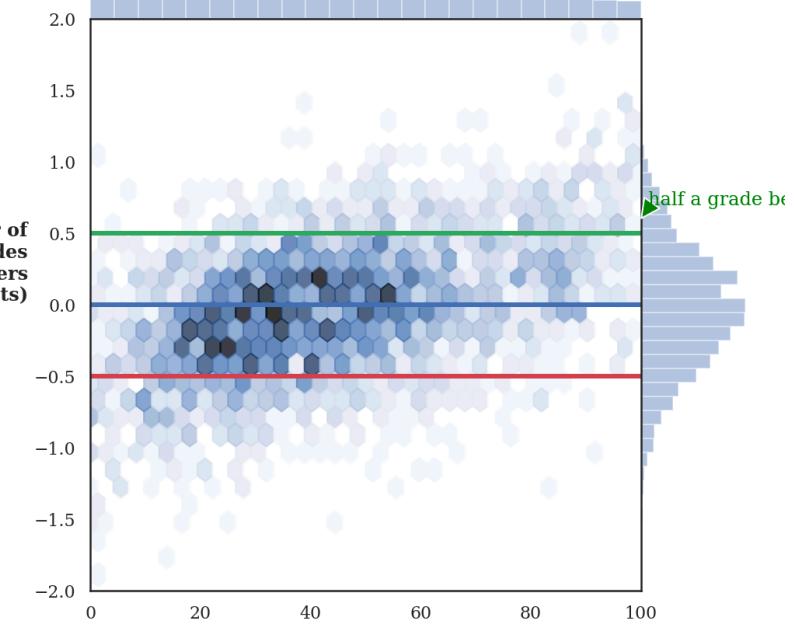




Children at schools with at least 2/3 doing a language progress more



Children progress more at schools where more people take foreign languages



Children at schools which take languages progress more in all subjects

