CUBeC

Centre for Understanding Behaviour Change

www.cubec.org.uk

Key Stage 4 Accountability: Progress Measure and Intervention Trigger

Simon Burgess and Dave Thomson

December 2013

Short Policy Report No. 13/11

(Funded by Department for Education)

Centre for Understanding Behaviour Change
Centre for Market and Public Organisation
University of Bristol
2 Priory Road
Bristol
BS8 1TX
UK

www.cubec.org.uk

CUBeC delivers evidence and insight into the drivers of behaviour change to inform and improve policy-making. The Centre combines expertise in a wide range of academic disciplines: economics, psychology, neuroscience, sociology, education, and social research.

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

Contents

List of figures

List of tables

1. Introduction

School accountability is a crucial part of any devolved system of education with many autonomous schools. While schools have a great deal of operational freedom in the delivery of education, they are held to account in terms of the results that they help their pupils achieve. The system in England has a number of components but a prominent role is given to a floor target: a minimum level of attainment for a school's pupils below which intervention of some sort is triggered.

This note concerns a revised measure of pupil progress (value-added), including a floor target, to support the proposed accountability regime for secondary schools[1]. We were given a brief to explore different statistical techniques, but also an explicit set of specific requirements to which the measure had to adhere. We set out these requirements, make our proposal, and discuss its implications. We also set out in detail the statistical work we have carried out, and outline other options.

Further detail is given in the Technical Annex published with this report.

[1] https://www.gov.uk/government/consultations/secondary-school-accountability-consultation

2. Requirements

The specification was as follows:

a. Statistical requirements

- An unbiased national progress line that summarises the proposed 'Attainment 8' measure[2] in terms of prior attainment (using key stage 2 marks in English and mathematics);
- Exploration of a number of different technical techniques, including a discussion of their respective merits, to achieve the first requirement;
- Measures of uncertainty (if applicable);
- A consideration of how each of the options might fit into National Pupil Database (NPD) Performance Tables (PT) production cycles

b. Policy requirements

- Provision to schools of 'predictions' of end of year 11 performance for individual pupils when they are in year 9, with all pupils with the same prior attainment receiving the same prediction regardless of any other factors;
- A measure of value-added that compares pupils' end of year 11 attainment to the above;
- A measure that can be disaggregated for different groups of pupils;
- Identification of schools at which pupils are making below average rates of progress;
- Consideration of incentives and perverse incentives that the various statistical options might imply.

Two points were central to the model required:

- It is a progress model, measuring a pupil's attainment given her/his prior test scores. Other factors, known and agreed to affect pupil performance even conditional on prior attainment, were to be excluded from the analysis.

- The model was to be 'fair' in the following sense: given prior attainment, all pupils face the same *ex ante* chance of falling below some threshold; that is, there are no biases in the proposed model which mean that some pupils have an expected progress below zero. Unfortunately, fairness in this sense is not a feature of the current model. Extending this to a school-level: the model should be 'fair' in that conditional on their

[2] This takes the pupils' highest 8 grades at GCSE,

intake by measured prior attainment, all ‘neutral’ schools should have the same chance of triggering intervention regardless of their intake. For example, it should not be the case that schools with low ability intakes start out at a disadvantage (though see the point below).

It was explicitly recognised that these two requirements were in conflict. There are known to be factors other than prior attainment that influence GCSE outcomes. By not taking those into account, and only conditioning on prior attainment, some pupils will be more likely to achieve below their expected level, so defined. By extension, schools with a disproportionate number of those pupils will also be more likely to produce a score below that expected. So the measure is ‘unbiased’ and ‘fair’ given the restriction that it be based solely on measured prior attainment.

3. Proposal

In this section we set out our proposal. The detailed statistical analysis underlying this and justifying our choice are presented below in section 5.

a. Our proposed pupil progress model

We propose a simple model. We show in Section 5 that more complex models add little or nothing to this. We see the simplicity as a virtue. We set out the pros and cons of all the approaches taken in table 5.4 in section 5.

Summary: For each pupil, we average the key stage 2 fine grade[3] in English and the Key stage 2 fine grade in maths. For brevity, call this K. We take each possible value of K (e.g. 3.4, 4.1, 5.5) and group pupils by K. We simply take the mean GCSE performance of each K group as that group's predicted performance. Each pupil's relative value-added score is the difference between this predicted performance and her/his actual performance. Necessarily this means that the mean value-added score is zero for each possible key stage 2 performance, guaranteeing mean fairness in the sense defined above.

As agreed prior to this project, these benchmarks for expected progress could be estimated on a previous year. This would mean, for example, computing progress targets for GCSE performance in 2016 of students about to commence their key stage 4 programme in September 2014 using the data from students who have just taken their GCSEs in 2013. This has two highly beneficial effects: first, schools know the GCSE score expected for each pupil at the start of the GCSE programme; secondly, it will be in principle possible for every pupil and every school to show greater than expected performance.

In Section 5b, we observe that both raw and value-added outcomes at school level are heavily correlated with mean entries in GCSEs and other approved qualifications. We anticipate, but cannot know with any certainty, that the number of pupils entering the maximum of ten qualifications counted in the measure (eight qualifications with double weighting for English and maths) will rise from the current 52 per cent in the next few years as schools respond to the challenges of the new accountability framework. We recommend that benchmarks for expected progress are only produced once stability in entry patterns has been achieved. In the interim, a retrospective (ex post) value-added measure could be used in its place.

[3] Fine grades are measures based on pupils' marks. For example, a pupil who achieves a raw test mark exactly halfway between the lower boundary for level 3 and the lower boundary for level 4 would be assigned a fine grade of 3.5.

There is a legitimate debate about the timing of any switch to an *ex ante* system. And while this is a transition issue rather than a steady-state issue, it is nonetheless important for that. It is clear that ex ante predictions are a good thing. Our point is that because recent incentives will probably drive a rapid increase in entries for approved qualifications, this will create some transitional turbulence that could be avoided by waiting and taking (say) 2014 as base year (by 2014 schools will have had time to respond to the Wolf proposals). The turbulence might discredit this measure despite it being the change in entry patterns that was causing the problem and reduce its credibility in the long run. In the end this is a debate about timing of introduction and not about the statistical or economic merits of the measure. Both *ex ante* and *ex post* models deliver fairness in a statistical sense as defined above. But thinking about likely imminent changes in behaviour, *ex ante* models may not be fair during the transition because we think predictions for lower ability pupils are currently lower than they should be. This will correct itself once schools start entering lower ability pupils for more approved qualifications.

The scaling of grades in GCSEs and other approved qualifications also has an impact on value-added scores. The current points score system, with 16 points for a G and 58 points for A*, rewards entry: four grade G passes yields a higher points score than a single A*. This was not the case under the previous 1-8 scale. Given that the current points score was introduced to incorporate entry level qualifications which will no longer be counted, it is an appropriate time to consider a suitable points score structure from both statistical and policy perspectives.

Note: In our analysis, we have produced models that include all schools and models that include mainstream schools only. For the most part, we have illustrated our analysis with results from example models based on the latter since the floor targets do not apply to special schools. However, it is ultimately a policy decision whether to include them or not. We note that when they are included in a single model structure, the progress of their pupils is significantly lower than that of pupils with equivalent observed prior attainment attending mainstream schools.

b. School and Pupil Group Scores

Summary: A value-added score (and confidence intervals) would be calculated for every school, as now, by averaging the value-added scores (difference between actual and predicted points scores) of its pupils. Unlike under the current multilevel model (MLM scheme, there would be no requirement to shrink (precision-weight) scores although this could be achieved by other means (e.g. empirical-Bayes adjustment) if required. Given the non-constant variance in value-added scores with respect to prior attainment, it may be appropriate to standardise them in order to calculate confidence intervals and associated tests of statistical significance (Schagen, 2006)[4]. The variance around the average GCSE score decreases as we move up the prior attainment distribution: there is much more variation around low k value-addedey stage 2-achievers' GCSE scores than around high key stage 2-achievers' GCSE scores. Schools which have a lot of the former pupils will therefore have much larger 'natural' variation around the average than schools with a lot of high key stage 2 performers. This is very important because it means that a school with low-performing intake will be much more likely to find itself a fixed number of grades below the mean than a school with a high-performing intake. This is the reason for our novel way of approaching floor targets. In terms of school-level VA, we could avoid this by standardising value-added scores before calculating significance tests (although in practice it does not appear to make a massive difference).

Also as now, value-added scores can be disaggregated for different pupil groups. At national level, these show some quite significant biases for some pupil groups, as we discuss in Section 5c, which will obviously influence the overall scores of schools with disproportionately large numbers of pupils from such groups

Based on an ex post value-added calculation, school mean scores (differences between actual points scores and predicted points scores) are heavily correlated with the mean number of qualifications entered by pupils at a school and the outcome measure, the proposed new 'Attainment 8' points score.

Table 1: Bivariate correlations at school level (mainstream schools with at least 50 pupils only)

	Mean key stage 4 Outcome	Mean key stage 4 Entries	Mean key stage 2 APS	% Pupil Premium
Mean KS4 Entries	.883			
Mean KS2 APS	.872	.689		
% Pupil Premium	-.623	-.561	-.700	
Mean VA	.799	.818	.406	-.328

Note: APS – Average Points Score

Put simply, schools which are already entering pupils for the full quota of GCSEs and other approved qualifications are more likely to achieve positive value-added scores. However, we expect other schools to respond to the new accountability regime and begin to enter pupils for more of these qualifications in lieu of vocational alternatives. This is

[4] Schagen, I. (2006). 'The use of standardized residuals to derive value-added measures of school performance', in Educational Studies, 32(2), 119–32.

likely to have a dampening effect on the correlations shown above. Only when entry patterns reach some form of equilibrium will the school value-added scores and associated floor target calculations properly reflect school effectiveness.

Note: As with the results of any value-added modelling, school-level scores are inherently conditional on the prior attainment of their intakes. For example, a school with a highly able intake may achieve the same score as a school with a much less able intake. We cannot assume that the two schools would have achieved the same score had they shared a similar intake: for that reason value-added scores are not directly comparable between schools. This makes them less appropriate for parental choice than for school self-evaluation and inspection.

c. School level floor target calculation

We propose a new way of producing the school-level outcome from the pupil level progress measures. This makes the measure ‘fair’ in the sense defined above and is also more intuitive.

Summary: At each level of K, we capture the standard deviation of GCSE performance as well as the mean. We tag a pupil as “causing concern” if her/his performance falls below 50 per cent of a K-specific standard deviation below the mean. This typically covers around 28 per cent - 30 per cent of pupils for each level of K. The school level outcome is then simply the fraction of pupils identified as “causing concern”. In a ‘neutral’ school, we would expect this to be around 30 per cent simply by the operation of the random process.

Note: the choice of 50 per cent of a standard deviation below the mean is essentially arbitrary and changing it obviously picks out more or fewer pupils. This is discussed in further depth in the flowing section.

d. Intervention trigger

Given this approach, the school intervention trigger is expressed as a fraction of a school’s pupils which are “causing concern”.

Summary: Continuing the case set out in (c), we would expect 30 per cent of pupils to be identified in a ‘neutral’ school. A percentage above that suggests that the school is performing less well than the average. The intervention threshold should highlight schools performing considerably less well than the average. This might be set at twice the expected level, so in this case at 60 per cent.

Note: the two key tuning parameters in this are the pupil level marker (what percentage of a K-specific standard deviation is used to mark concern) and the intervention threshold (relative to the number to be expected by chance, how much higher to put the intervention threshold). Together, they imply the number of schools that are likely to be highlighted for intervention and we set out some examples in section 5 below.

One key design question is how best to set the tuning parameters. There is a pragmatic answer: set them in such a way as to keep the implied burden on DfE and Ofsted at about the current level. We do not really have enough of an evidence base to set them in an optimal manner from an economic or behavioural perspective. Our research[5] has shown that schools given Notice to Improve by Ofsted do raise their performance, so some form of pressure helps. The decision on how many schools to have under intervention is a high-level policy decision. We can illustrate what different values of the parameters are likely to produce, but the decision is up to others.

In this table we illustrate the implications for intervention of different parameter values for the current data. Note that this table is based on ex post facto analysis of the 2012 key stage 4 dataset and it would be possible, in principle, for all schools to exceed floor targets if they were to be given pupil predictions ex ante.

Table 2: Number of schools highlighted for intervention, by different parameter designs

National Average Multiplier	Pupil Trigger (Standard Deviations Below Prediction)			
	0.3	0.5	0.7	1.0
1.5	512	560	594	622
2.0	130	204	268	323
2.5	22	59	99	161

Note the multiplier is more important in affecting the overall number. Selecting the design parameters is a policy decision, but a combination of pupil trigger of 0.5 of a standard deviation below prediction and a multiplier of 2 times the national average seems intuitive and produces (on these ex post estimates) a number not too dissimilar from the current workload on DfE and Ofsted.

The most important issue to highlight is that this measure (like any other) is very sensitive to quantity – to the number of approved qualifications entered. The aim of the progress and intervention measure is to highlight low (quality) performance, but in this case it also picks up quantity variation.

[5] Allen, R. and Burgess, S. (2012) 'How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England. CMPO WP 12/287, CMPO University of Bristol.

e. Commentary: benefits and limitations

We believe that this approach makes a number of substantial improvements over the current model, and is also preferable to the alternatives. We summarise these here and give greater detail in section 5 below.

This approach achieves more-or-less exact ‘fairness’ or lack of bias: if it were true that all that mattered for GCSE attainment was prior attainment and school effectiveness[6], then under our proposal, every child in a ‘neutral’ school would have the same chance of being identified as causing concern whatever their prior attainment, and every ‘neutral’ school would have the same chance of being highlighted as under-performing whatever their attainment intake profile. This is certainly not true (for clear structural reasons) under the current scheme and we see it as a major strength of this approach.

To take a specific instance: this means that low-performing schools with high ability intakes will be more likely to be highlighted by this progress intervention system than under the current scheme, and averagely-performing schools with low ability intakes less likely. As we discuss below, this is strongly related to the number of approved qualifications that pupils are entered for.

It is very simple, which means that it will fit easily into NPD-PT production cycles and is easy and intuitive to explain. It will simplify production cycles as no statistical software is required to calculate it.

It has been historically very stable across years. More complex models risk substantial parameter variation from year to year which is not ideal. This approach yields only minor variations in the estimation. However, this should continue to be monitored in the first few years of the measure as schools respond to the new accountability regime and enter pupils for more approved qualifications.

Note that the information would not be presented to schools in terms of “fractions of K-specific standard deviations”, but in terms of the GCSE score cut-offs that are implied. A non-exhaustive list of examples is shown in the table below.

Table 3: Example thresholds for concern

Prior attainment	Expected GCSE score	Pupil VA score threshold for concern	Pupil GCSE score threshold for concern
3.1	198	-40	158
3.5	233	-44	189
4.0	285	-45	241

[6] Of course, it is not true that prior attainment is all that matters for GCSEs; we return to this in the next section.

4.7	383	-41	342
5.2	459	-35	424
5.6	523	-27	496

We discuss both entries and scaling in further depth in section 5.

f. Discussion

There is an important decision to be made whether to adjust for the non-constant variance in value added scores among pupils in setting a school-level floor target. Such a floor could be set at a variance-adjusted level below the expected level (as we propose) or at a fixed number of grades (measured in points) below the expected level. There are a number of points to make.

The argument in favour of the variance-adjusted measure is that it achieves fairness as defined above. Setting the floor in terms of fixed grades, rather than adjusting for differential variance, will necessarily tip more low-ability intake schools below the floor than high-ability intake schools. This is inevitable given the data, and contrary to the requirement placed on us to achieve absence of bias by intake characteristics. Measuring it the way we propose is a simple and intuitive way of getting round the problem of differential variation and delivers fairness relative to intake characteristics.

Secondly, the measure we propose here uses individual pupils as the unit of measurement. If that pupil is performing way below expectation, then s/he is highlighted by this procedure. And that will be recorded as a cause for concern, regardless of how the other pupils in the school are performing. If there are many such pupils (as defined above) then the school will be highlighted, even if all the other pupils are scoring far above expectation. We believe that this focus on low-achieving pupils, rather than the average, is appropriate in a floor target. By contrast, a school-level average gap of a fixed amount below expected means that high performance by some pupils masks low performance in others.

Both the pupil-based variance-adjusted measure and the school-average gap involve thresholds. In the former proposal, each pupil has a target or expected grade (which s/he will typically know), and a “threshold for concern” grade (which s/he typically won’t). But the same is true in the latter proposal too: the school average gap (actual minus expected) is simply the sum of pupil-specific gaps. A school could sift the list of pupils, predict these pupil-specific gaps and take a view as to which pupils will be more at risk of not achieving the threshold, and act accordingly.

In practice with school behaviour as it is now, the difference between the two approaches is minor. Comparing the 2 methods for defining the floor, simulations show that 24 schools swap category when we account for non-constant variance. 10 schools below the unstandardised floor end up above the floor after standardisation, with 14 schools moving in the opposite direction. Unsurprisingly it is the lowest quintile schools that are the most likely to move from below to above after standardisation.

4. Implications for behaviour

We believe that schools will welcome the new progress measure and the associated intervention strategy. Unlike at present, they will know in advance what is expected for each pupil and all schools will have the opportunity to do better and exceed predictions. It is also unbiased with respect to the distribution of prior ability in each school's intake. In this section we briefly discuss its implications for behaviour by schools.

a. Gaming

It should be acknowledged that high stakes accountability measures are often prone to gaming. Experience suggests that this one too may well eventually generate some clever strategies that at this point cannot be foreseen.

Clearly, knowing the progress expected for each pupil ahead of time is much fairer on schools. Nevertheless it does provide a focal point for their efforts. A school attempting to optimally deploy its resources (its most effective teachers and smaller class sizes for example) might conceivably aim for each pupil to hit their expected level (or indeed the threshold for concern), but no more. This sort of activity is inherent and unavoidable in any system that has threshold-type metrics. Of course this progress intervention is only one element of the accountability framework, and having a different component based on the simple average score will mitigate this. Even so, choosing the language to describe these GCSE targets to imply that they are a minimum to exceed not a level to aim for would be helpful.

The models we outline are unbiased with respect to pupil prior attainment. They are therefore fair to all pupils, regardless of ability, and there will be no advantage to schools to favour more able pupils in admissions rounds. However, the models are biased with respect to other pupil characteristics (see section 5c). Some groups will be less likely to achieve their predicted outcomes. There is therefore a risk that some schools might become reluctant to admit such pupils. For example, pupils attending mainstream schools who had SEN met by School Action Plus at the end of Year 6 achieved, on average, 18 points below expectation at the end of key stage 4. Other pupil groups may also be prone to under-achievement: those with a poor history of primary school attendance or having been subject to exclusion. By contrast, key stage 2 prior attainment may under-estimate the potential of some pupil groups. Pupils flagged as speaking English as an additional language at the end of primary schools who were not assessed at key stage 1 achieved on average above key stage 4 expectation. While it seems clear that schools' ability to work around the Admissions Code is more restricted than it was, this progress measure still offers incentives to schools to try.

Finally, these official predictions for GCSE performance will sit alongside other predictions or expectations in a schools management information system. These may be internal, produced by the school itself, or bought in from a number of providers). There is potential scope for confusion in running these systems and targets alongside each other.

b. Schools in disadvantaged areas may face continuous intervention

Here we set out what is in our view the main drawback of the requirement to exclude all other sources of GCSE performance variation from the determination of an intervention trigger.

As shown below, pupils eligible for the pupil premium perform less well at GCSE conditional on prior attainment, and this effect is not trivial. Therefore, if nothing changes, schools with a high fraction of such pupils will see a lot of them performing below an expectation based on a national average conditional only on prior attainment. There are clearly other factors too: gender, ethnicity, pupils with Special Educational Needs, and so on. It is true that controlling for prior attainment does account for part of the effect of these factors, but not all: they significantly influence progress measures too.

Given this, schools with high fractions of pupil premium-eligible students will on average have a substantial number of pupils "causing concern" and trigger an intervention. We illustrate this in Table 4 (based on using a pupil trigger of 0.5 standard deviations and a multiplier of 2).

Table 4: Number of schools below floor by pupil premium decile

	Below Floor	Total
Least	3	299
2	3	300
3	2	300
4	3	300
5	6	301
6	14	298
7	31	301
8	36	300
9	51	300
Most	55	299
Total	204	2998

It may be that persistent intervention will lead to the exercise of "voice" and parental pressure will lead to dramatic school improvement. It is also possible that persistent intervention will lead to the exercise of "exit" by a lot of teaching staff.

5. Statistical Methods and Tests

In this section we justify our choice of pupil progress model and consider its properties, and we justify the method of arriving at a school level measure. We also discuss a number of outstanding issues in the chosen model. Further detail about our investigations into other statistical approaches can be found in the accompanying technical annex.

a. Justifying our model of pupil progress

Criteria for a fair model

The department presented us with a number of requirements for the Value added measure. The 'Attainment 8' key stage 4 points score proposed by the recent secondary school accountability consultation is the outcome. Secondly, only prior (key stage 2) attainment in English and maths could be included as independent variables. Other factors, which research has shown have a bearing on key stage 4 outcomes over and above prior attainment, will not feature in the revised value-added measure.

The initial step in constructing a measure which is fair to all pupils and schools is to ensure that the model of pupil progress given the bases of input and outcome measures delivers residuals which meet desirable characteristics. They should exhibit:

- Monotonically increasing predictions with respect to prior attainment
- Zero mean (that is, unbiased across the key stage 2 prior attainment range)
- Constant variance ('homoscedasticity')
- A normal distribution (in order to calculate fair tests of statistical significance at school level)

We explored a number of statistical techniques for producing a fair value-added measure. These included:

- Piecewise regression
- Ordinary Least Squares (OLS) regression
- Kernel regression
- Lowess smoothing
- Multilevel modelling (MLM)
- Quantile regression

For each technique, we assess the resulting residuals in terms of their distributional properties. In particular, we pay close attention to the mean score having banded pupils based on their prior attainment. We test that the mean score for each band is simultaneously equal to zero.

The proposed ‘attainment 8’ points score

We have used a version of the 2012 key stage 4 final dataset provided by the Department for Education. It contains a prototype version of the ‘Attainment 8’ points score measure proposed by the recent departmental consultation on secondary school accountability. The design of the measure is still currently in development.

For the most part, the analysis is conducted on the subset of pupils included in the calculations for the existing 2012 value-added measure. A number of analyses are performed based on the subset of pupils attending mainstream schools only.

The proposed measure has more tractable and convenient statistical properties compared to the current ‘Best 8’ measure which make it a more desirable indicator to model. Table 5 shows that the measure is much less ‘peaked’ (kurtosis is minuscule) and substantially less skewed, but we note that its variance is substantially larger. It is also, from Table 6 below, more strongly correlated with key stage 2 prior attainment (the average of English and maths fine grades) than the current measure.

Table 5: Key stage 4 Outcome Measures 2012

		N	Min	Max	Mean	Std. Deviation	Skew-ness	Kurt-osis
Mainstream	Proposed	525532	.00	580.0	370.2	121.9	-.553	-.110
	Current	525532	.00	580.0	429.1	91.9	-1.453	3.715
All schools	Proposed	533062	.00	580.0	365.6	127.3	-.638	.045
	Current	533062	.00	580.0	424.6	99.5	-1.589	3.799

Table 6: Bivariate correlations (pupils attending mainstream schools only)

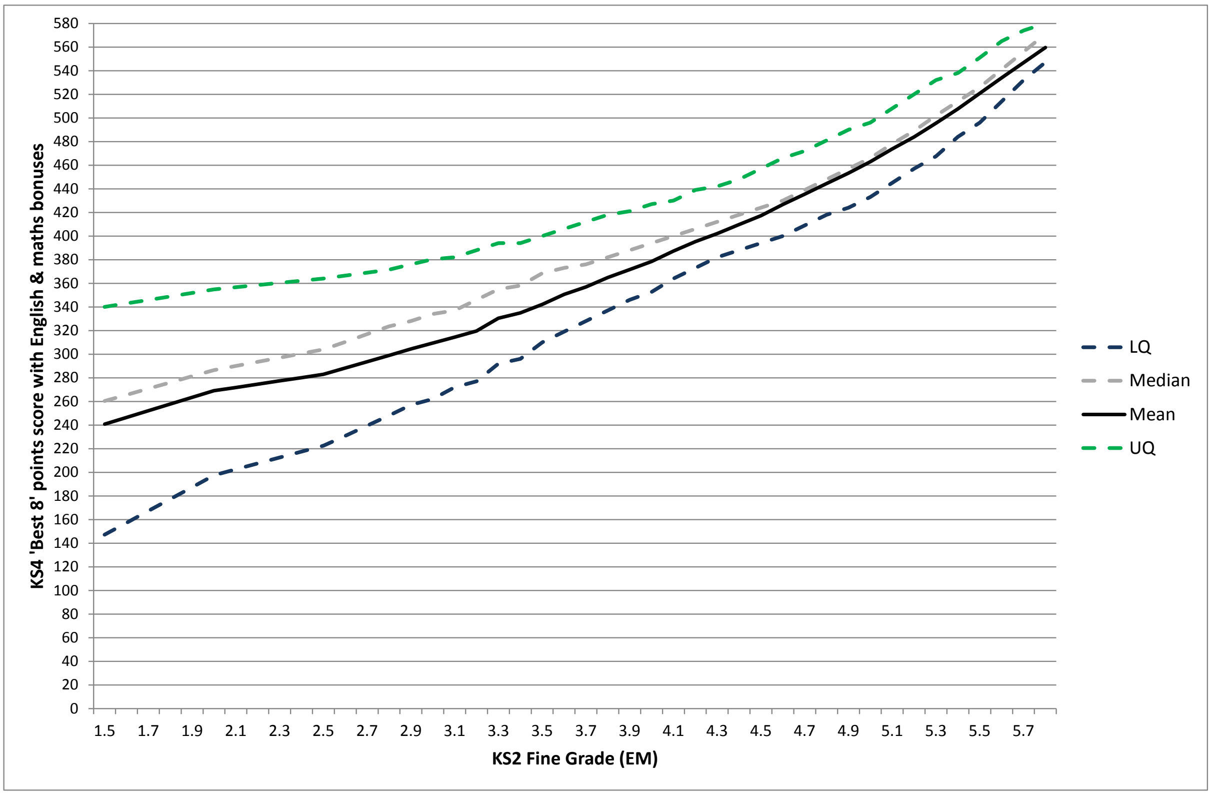
	Current KS4	Proposed KS4
KS2 (EM)	0.647	0.749
Current KS4		0.871

In Figures1 and 2 we summarise respectively the current and proposed points scores by key stage 2 prior attainment. Having calculated the mean key stage 2 fine grade in English and mathematics, we have rounded to one decimal place (hence creating a discretised variable). To avoid sparseness of pupil numbers at the lower end of the prior attainment range[7][8]:

- pupils with a mean fine grade below 1.5 are assigned to 1.5;
- pupils with a mean fine grade between 1.6 and 2.0 are assigned to 2.0;
- pupils with a mean fine grade between 2.1 and 2.5 are assigned to 2.5;
- pupils with a mean fine grade between 2.6 and 2.8 are assigned to 2.8; and at the top
- pupils with a mean fine grade of 5.8 or above are assigned to 5.8.

Linear interpolation has been used in the charts.

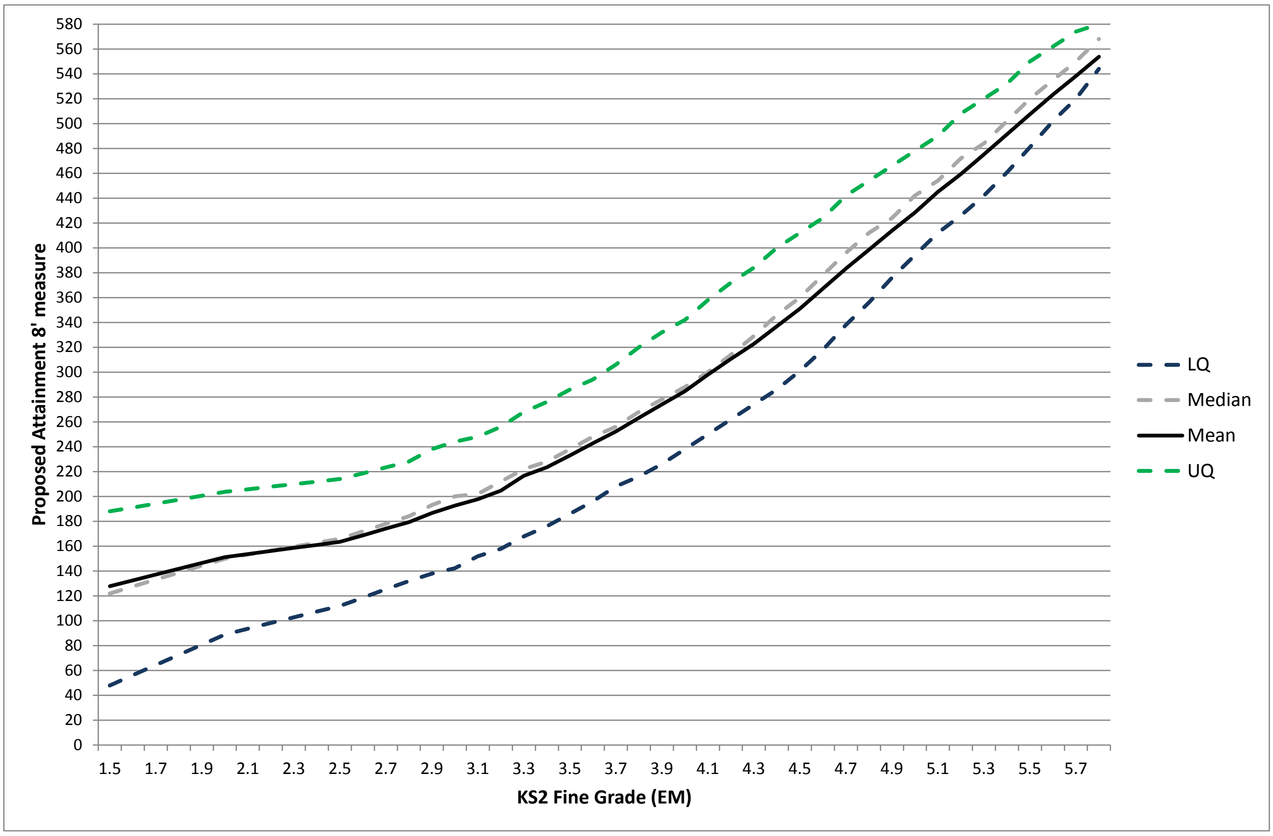
Figure 1: Current key stage 4 'Best 8' points (with English and maths bonuses) by prior attainment 2012 (mainstream schools only)



[7] There is a certain amount of arbitrariness here. There would be no difference in practice to coding 2.6 to 2.8 to 2.7 rather than 2.8. We do grouping simply to ensure monotonically increasing predictions. We would end up with spikes if we included 2.6, 2.7 and 2.8 individually.

[8] This banding at the very bottom end of the distribution may make the measure vulnerable to year-on-year changes. However, we have looked back over the last 3 years and the bands are stable. This will need to be checked as pupils are entered for more approved qualifications but they should be sufficiently large (in terms of pupil numbers) to remain stable.

Figure 2: Proposed key stage 4 'Attainment 8' points (with English and maths bonuses) by key stage 2 fine grade 2012 (mainstream schools only)



Compared to the current measure, the proposed measure:

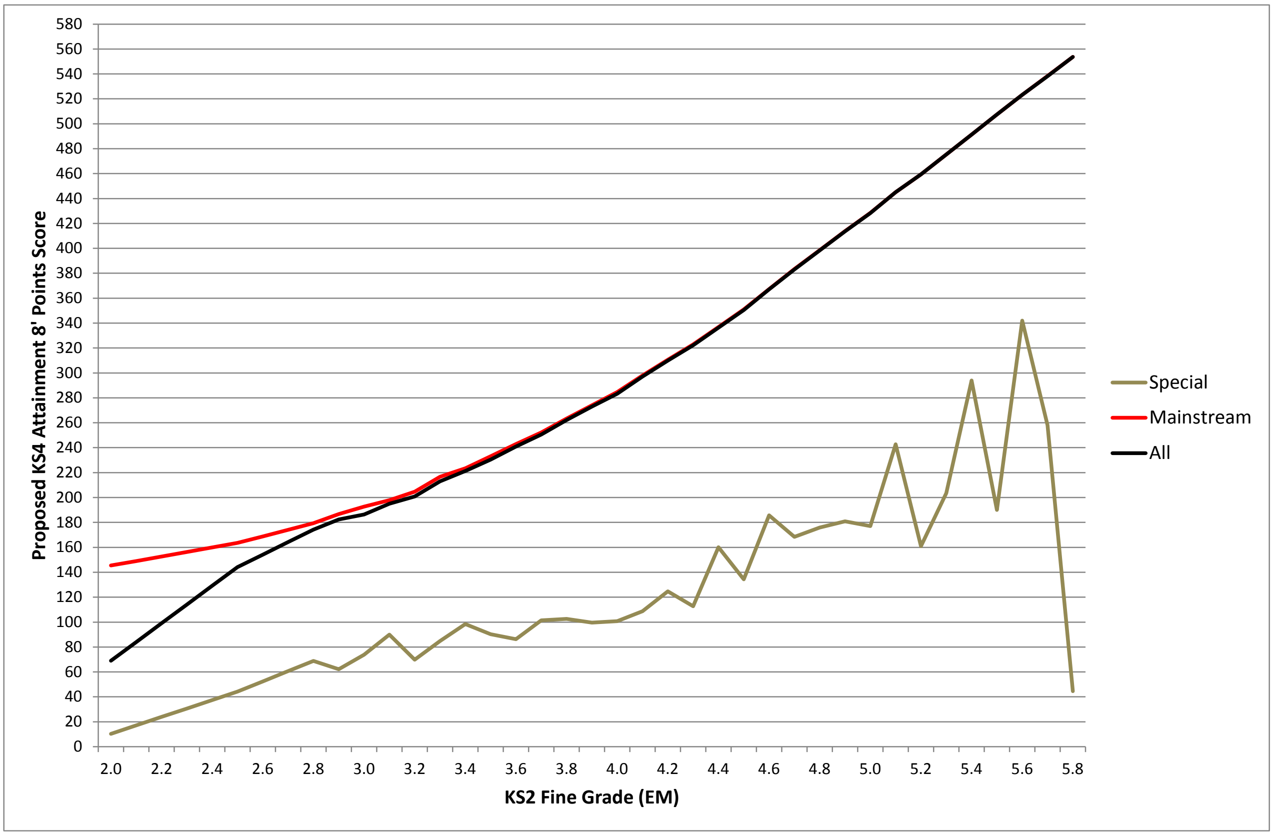
- appears to exhibit more equal variation across most of the key stage 2 prior attainment spectrum;
- increases more sharply as prior attainment rises; and
- has a mean that is generally closer to the median (indicating a more Normal distribution].

Visually, the mean line of Figure 2 has a more pronounced curvature *within the key stage 2 fine grade range observed* but with a discernible linearity at the upper levels of the key stage 2 and key stage 4 outcomes.

The difference between the Lower Quartile (LQ) and mean line is around 48 points for the key stage 2 fine grade range from 3.0 to 4.7 inclusive. This range covers over half of pupils nationally in mainstream schools. The gaps are wider at the lower end, and narrower at the top end. These will be important considerations if setting a threshold for concern' for pupils relative to the national line and also for the calculation of school-level scores, particularly for those schools with disproportionately large cohorts of low-ability or high-ability pupils.

The current value-added model is based on pupils in mainstream schools only, primarily because the fitting of school lines for special schools in the MLM structure disrupted the fixed part of the model. Pupils in special schools could be included in the calculation of a national pupil progress line. However, as Figure 3 below shows, this will have a significant bearing on the position of the lower end of the national line and, therefore, the value-added scores of schools with disproportionately large numbers of pupils with low prior attainment, most of which will have scores which are significantly below average. We adopt a strategy that creates both 'mainstream only' and 'all pupils' versions of models.

Figure 3: Mean proposed key stage 4 'Attainment 8' points (with English and maths bonuses) by key stage 2 fine grade 2012 (all pupils)



The simple model

Figure 2 above shows a simple method of calculating value-added. For each value of key stage 2 fine grade (rounded to one decimal place), the mean outcome score can be used as a prediction (or 'benchmark' or 'expectation') in a value-added measure. It is axiomatically unbiased with respect to prior attainment. That is to say, for each value of key stage 2 fine grade (e.g. 3.5, 4.6), the mean value-added score nationally is zero.

We refer to this as the **simple** model and it is the model we recommend as the basis of a value-added measure. Firstly, its very simplicity is attractive and therefore it avoids the scepticism that accompanies the all too often misunderstood statistical models. Secondly, schools tend to find transition tables and charts easy to use and charts such as those shown in Figure 2 are easily produced and can be overlaid with scatterpoints (and quartiles) representing each pupil at the school.

In our view, any further complexity is unnecessary, firstly because it would render the methodology less comprehensible and secondly, as we show below, it adds very little predictive power to the model.

The simple model can be conceived as a piecewise regression model in which each of the 33 values of mean key stage 2 fine grade shown in Figure 2 is a ‘piece’. In this way, additional variables could be added to the model.

Other statistical models

All of the models we propose are founded upon pupils’ mean finely graded points scores in English and mathematics at key stage 2. They cover a range from 3 (working below level 1) to 36 (level 6). The mean fine grades used in the previous section can be obtained by dividing by six.

In addition to the mean, we also included the English subject differential (the difference between the English points score and the average points score[9]) in some models. Our choice of explanatory variables was limited by design to available measures of prior attainment, and restricted in the case of lowess and kernel regression by the capability of the software (SPSS, STATA) to produce a smoothed relationship from a single explanatory variable.

In the OLS (and MLM and quantile) models, we included one or more ’pieces’ to locally tune the relationship between prior attainment and outcome to yield unbiased residuals. We grouped pupils into twenty equal-sized bands based on increasing level of mean prior attainment and each group formed a piece in the models. Table 7 summarises the explanatory variables used in the various models.

Table 7: Explanatory variables used in models

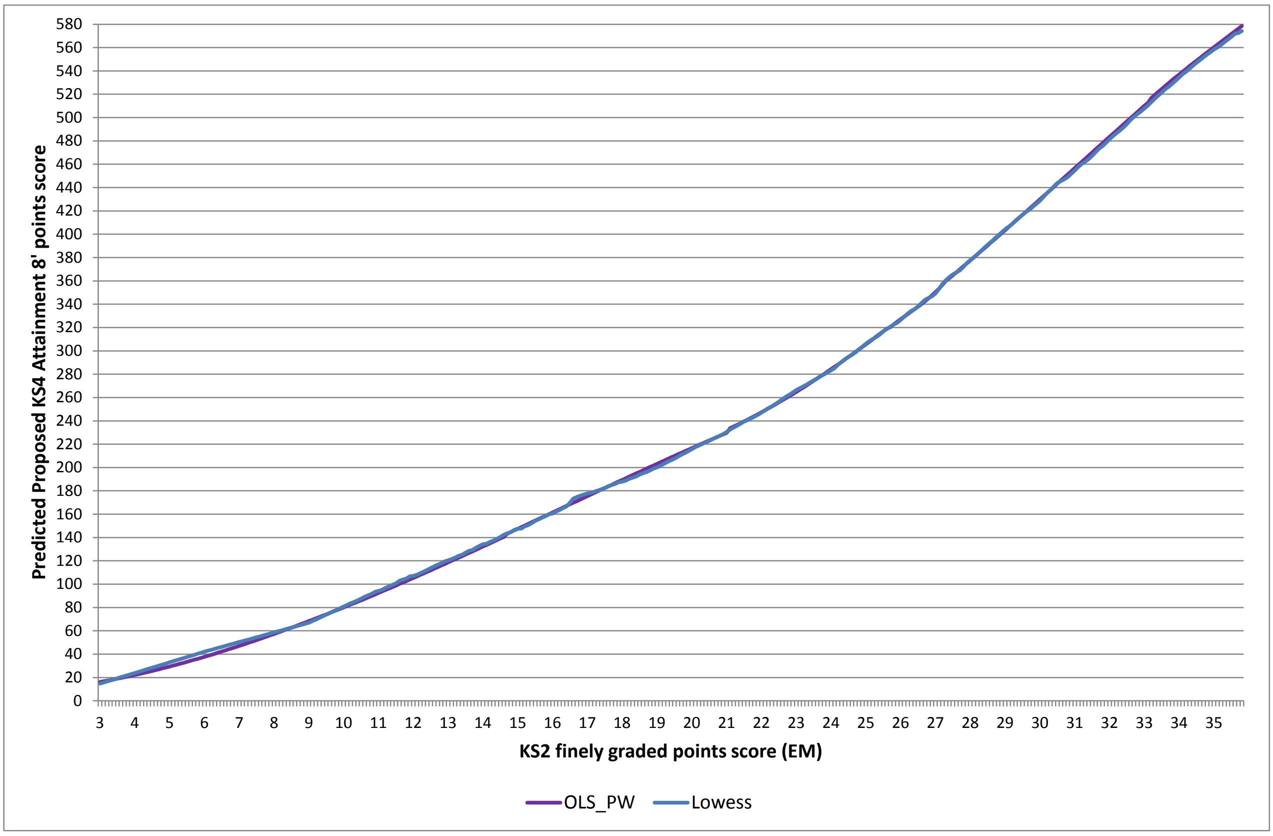
Simple	Mean finely graded points score divided by six, rounded to 1 decimal place (30 discrete values)
Simple Extended	As simple, plus the English subject differential
OLS (cubic piecewise)	Mean finely graded points score with quadratic and linear terms English subject differential A piece to define the top 10% of pupils in terms of prior attainment A piece to define the bottom 10% of pupils in terms of prior attainment Various interactions between the above

[9] In effect, half of the difference between points scores in English and maths

OLS (percentile)	Mean finely graded points score (linear term only) English subject differential A piece for each prior attainment band Various interactions between the above
Kernel	Mean finely graded points score
Lowess	Mean finely graded points score
MLM	As OLS
Quantile	As OLS

All the approaches we tried yield broadly similar pupil progress lines. For instance, Figure 4 compares predictions from the lowess and cubic piecewise models. The cubic piecewise model produces slightly higher predictions for pupils with high prior attainment while the inverse is true for pupils with low prior attainment. However, most of this difference is accounted for by holding the English subject differential constant for the cubic piecewise line. Pupils with high prior attainment tend to have negative English subject differentials, whilst they tend to be positive for pupils with low prior attainment.

Figure 4: Comparison of predicted values from the lowess and cubic piecewise models



The various models vary in terms of complexity and the degree to which they can be said to be unbiased. Their advantages and disadvantages are summarised in Table 8.

Table 8: Summary of the advantages and disadvantages of value-added modelling options

Models	Advantages	Disadvantages
Simple	No statistical modelling required. Easy to understand. All information contained in a single table or chart. Unbiased residuals with respect to mean fine grade.	Explains slightly less variation in KS4 points scores than other methods.
Simple Extended	Explains slightly more variation than simple model. Unbiased residuals with respect to mean fine grade.	Adds slightly more complexity to explanation for users
OLS (cubic piecewise)	Provides unbiased residuals from a generalised relationship between prior attainment and outcomes	Adds extra complexity to explanation compared to simple model. Marginal increase in overall explanatory power
OLS (percentile)	By default provides totally unbiased residuals (with respect to prior attainment quantile)	Adds extra complexity to explanation compared to simple model
Kernel	Produces unbiased residuals from a single explanatory variable	Requires some manual intervention to tune the smoothing features. An advanced technique requiring significant effort to attempt to make the mathematics comprehensible to non-statistical audiences. Restricted to a single independent variable (though a user-written extension is available)
Lowess	Produces unbiased residuals from a single explanatory variable	Adds additional complexity to the weighting and smoothing features of kernel regression, and thus less intelligible to non-statistical audiences. Restricted to a single independent variable.
MLM	Individual school pupil progress lines created within a standardised national relationship between KS4 points scores and explanatory variables. More statistically efficient	Models yield biased residuals. More sophisticated model structures will be considerably less comprehensible to non-statistical audiences

Models	Advantages	Disadvantages
	estimates of VA scores	
Quantile	Produces a series of average relationship lines corresponding to different percentiles of pupil KS4 points score	Cannot be used to define an average progress line with unbiased residuals.

In Table 9, we compare two key statistics- the mean squared residual and the proportion of variance explained- from the various models. All provide a broadly similar degree of explanatory power.

Table 9: Model Diagnostics (mainstream schools only)

	% variance explained	Root mean square residual	Unbiased residuals by prior attainment
Simple	57.9	79.0	Yes
Simple Extended	58.2	78.6	Yes
OLS (cubic piecewise)	58.3	78.6	Yes
OLS (percentile)	58.3	78.6	Yes
Kernel	58.0	78.9	Yes
Lowess	58.0	78.9	Yes
MLM	58.0	78.8	No
Quantile	57.8	79.3	No

With the exception of MLM and quantile (median) regression, all the models we developed could be used to create an unbiased national pupil progress line. However, without rescaling points scores associated with grades in GCSEs and other approved qualifications, they all yield residuals with significant heteroskedasticity (variance that varies with respect to prior attainment), as we outline in the following section.

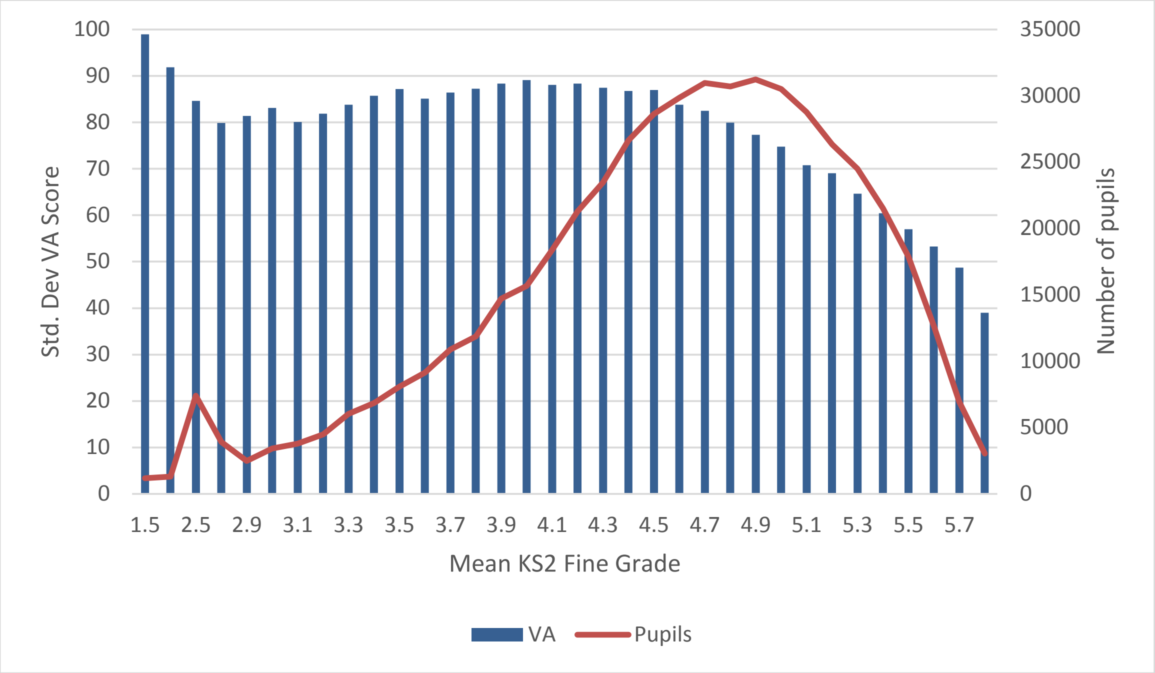
b. Justifying our approach to computing the school level floor measure

In the previous section, we outlined various approaches to producing a national pupil progress line that is unbiased with respect to prior attainment. However, and regardless of the method used, there remains non-constant variance in value-added scores with respect to prior attainment. Pupils with low prior attainment show a much greater dispersion of GCSE outcomes. As we describe in section 5e, this can be alleviated to some extent by rescaling the points scores associated with GCSE grades.

Defining a pupil-level trigger

As noted section 3b, a school level value-added score would continue to be produced by calculating the mean value-added score of its pupils. These school level scores could be used to define a school level floor measure. However, Figure 5 shows how the variance in value-added scores is associated with prior attainment. Schools with a disproportionately large number of high attaining pupils, whose progress is less variable, would be less likely to fall beneath a floor target defined by schools' mean value-added scores.

Figure 5: Standard deviation in value-added scores and number of pupils by mean key stage 2 fine grade (mainstream schools)



An alternative approach would be to define a pupil-level trigger, e.g. 0.5 of a standard deviation below prediction, and calculate the proportion of pupils below the trigger at each school. These proportions would be used to identify schools below the floor target. This approach effectively standardises value-added scores with zero mean and unit variance at pupil level by mean key stage 2 fine grade and these standardised value-added scores could also be used to calculate school level tests of statistical significance[10].

[10] In practice, only 62 schools (2%) changed significance state as a result of standardising Value added scores. 40 schools ended up with a lower significance state and 22 ended up with a higher significance state following standardisation. Lower ability schools tended to end up with lower significance states and higher ability schools with higher significance states.

For the purposes of exemplification, we continue to use the ‘simple’ model for mainstream schools only outlined in the previous section. We compare the two methods of defining the pupil-level trigger:

1. Use the standard deviation in value-added scores for all pupils (78.9)
2. Use the standard deviation in value-added scores for each group of pupils based on mean key stage 2 fine grade (e.g. for pupils with a mean fine grade of 4.0, the standard deviation shown in Figure 5 above is 90.6)

Pupils at least half a standard deviation below prediction are considered to be below the trigger. Table 10 shows the proportion of pupils below the trigger by mean key stage 2 fine grade. Based on a normal distribution, we would expect 30 per cent of pupils to be 0.5 standard deviations below their prediction.

Table 10: Percentage of pupils below the pupil-level trigger (mainstream schools only)

Mean key stage 2 fine grade	Method 1	Method 2	Pupils
1.5	39%	35%	1,192
2.0	33%	31%	1,316
2.5	30%	30%	7,414
2.8	28%	28%	3,900
2.9	29%	28%	2,475
3.0	30%	29%	3,420
3.1	27%	27%	3,797
3.2	27%	27%	4,497
3.3	28%	27%	6,023
3.4	29%	27%	6,859
3.5	27%	26%	8,089
3.6	27%	26%	9,138
3.7	27%	26%	10,875
3.8	28%	26%	11,857
3.9	28%	26%	14,703
4.0	29%	27%	15,688
4.1	29%	26%	18,379
4.2	29%	27%	21,293
4.3	29%	27%	23,475
4.4	30%	28%	26,640
4.5	29%	27%	28,633
4.6	28%	27%	29,847
4.7	27%	27%	30,956
4.8	26%	26%	30,701
4.9	24%	24%	31,239
5.0	24%	25%	30,507
5.1	22%	24%	28,768
5.2	22%	25%	26,333
5.3	21%	25%	24,504
5.4	20%	25%	21,513
5.5	19%	25%	17,902
5.6	16%	24%	12,642

5.7	15%	20%	6,949
5.8	12%	20%	3,035
Total	**26%**	**26%**	**524,559**

Method 2 results in a much more uniform set of proportions (apart from at the extremes) although the overall proportion of pupils is the same. 26 per cent of pupils are found to be 0.5 standard deviations below the mean, lower than the expected of 30 per cent. This indicates a degree of non-normality (caused by extreme values) in value-added scores.

There were 195 secondary schools below the 2012 floor target[11], plus a further 20 that were below the floor target prior to becoming a sponsored academy. If we were to rank the 2998 mainstream schools in the dataset by the proportion of pupils who were 0.5 standard deviations below prediction conditional on prior attainment, then the threshold for 215 schools would be 51.4 percent. 52 per cent or more pupils (twice the national average) are below the trigger at 204 schools.

The choice of 0.5 standard deviations in defining a pupil-level trigger is arbitrary, as indeed is the school-level floor target based on this measure. In Table 11**,** we show the number of maintained mainstream schools (out of a total of 2998) at which more than d*p pupils are below the pupil level trigger, where p is the appropriate national average from Table 8 and d is a multiplier. For example, for a trigger of 0.3, 1.5*33 per cent = 48.5 per cent.

Table 11: Number of schools below the floor

National Average Multiplier	**Pupil Trigger (Standard Deviations Below Prediction)**			
	0.3	**0.5**	**0.7**	**1.0**
1.5	512	560	594	622
2.0	130	204	268	323
2.5	22	59	99	161
% pupils below trigger	33%	26%	20%	14%

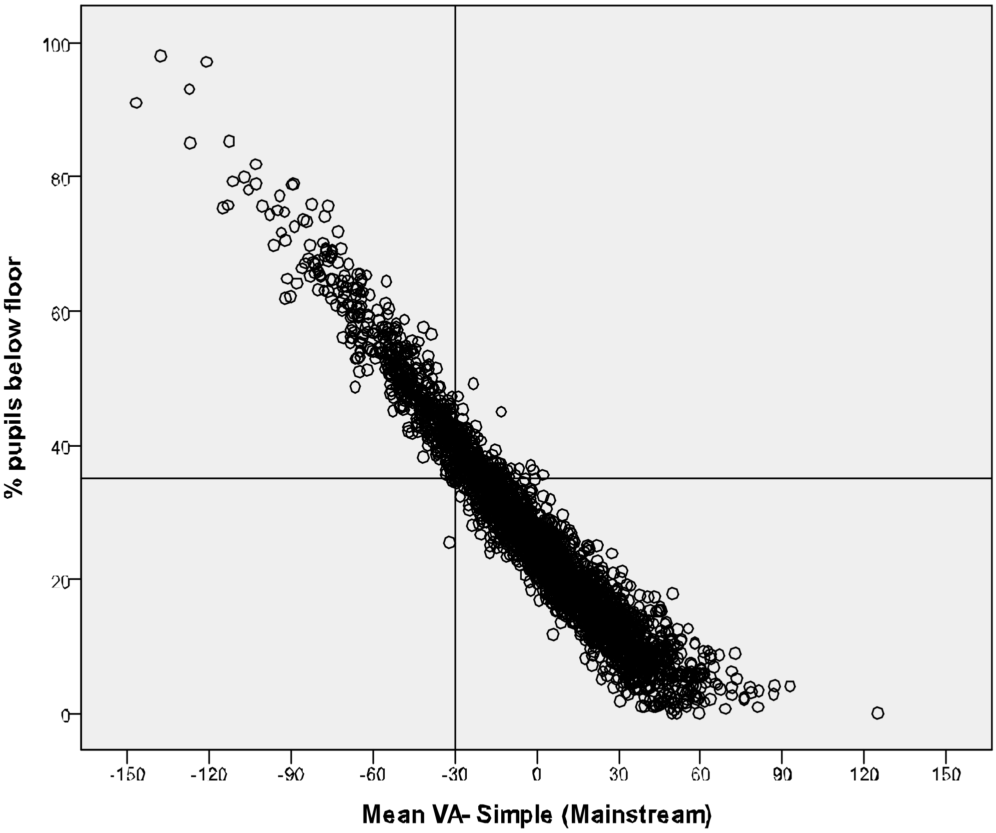
The choice of pupil trigger matters less than the choice of multiplier. A pupil trigger of 0.5 and a multiplier of 2 (2x26%=52%) produces a broadly similar number of schools below the floor as at present. Of the 204, just 64 are below the current floor target. Moreover, four schools below the 2012 floor target achieved value-added scores based on our proposed methodology that were significantly above average.

Even within schools below the floor, there will be a proportion of pupils who have achieved above their prediction. In Figure 4, we show the proportion of pupils at each

[11] Less than 40 per cent of pupils achieving 5 A*-C including English & maths and below average proportions making expected progress in English & maths.

school below the pupil trigger set at 0.5 standard deviations. The vertical lines represent national average (29 per cent) multiplied by each of the values shown in the first column of Tables 9 to 14. There are a number of differentially effective schools at which between 39 per cent and 52 per cent of pupils achieved 0.5 standard deviations *below* prediction and at which at least 20 per cent of pupils achieved 0.5 standard deviations *above* prediction.

Figure 6: Percentage of pupils 0.5 standard deviations above/ below prediction



Due to the current relationships between prior attainment, entries and outcomes described in the following section, schools with lower attaining intakes and schools which tend to enter pupils for fewer approved qualifications would be more likely than other schools to fall below the floor target. Of the 204 below the floor based on Table 11 above, 93 are in the lowest quintile (out of 299) compared to just 5 (out of 299) from the highest quintile. Similarly, no schools that entered the overwhelming majority of their pupils in the maximum of eight (ten with double weighting for English and maths) approved qualifications counted in the Attainment 8 measure would be below the floor as illustrated in Table 12.

Table 12: Number of schools below floor by quintile of mean entries in approved qualifications counted in Attainment 8 measure

Quintile	Below floor	Total
Highest	0	599

Second	0	600
Middle	1	600
Fourth	10	600
Lowest	193	599
Total	204	2998

Policy decisions have to be reached about how many schools should be below the floor and when the floor target should be introduced. Schools should have time to adjust to the new accountability regime. Our analysis in this section is necessarily retrospective and, in principle, it could be that no schools would be below the floor target if predictions were calculated ex ante.

c. The influence of the number of subject entries on the progress measure

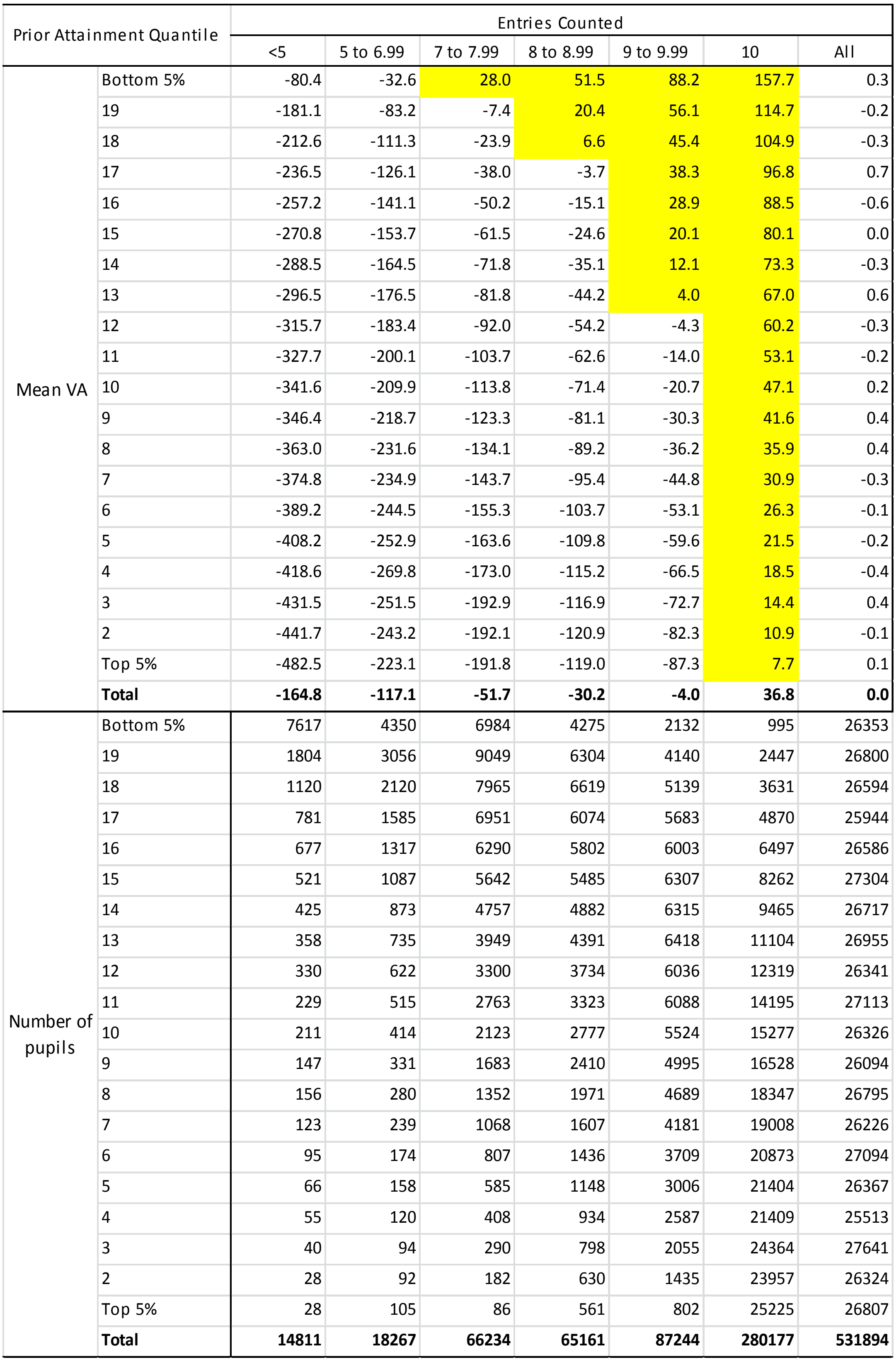
The introduction of the proposed ‘Attainment 8’ key stage 4 outcome accountability measure for Performance Tables (PT) will lead schools, particularly those which have tended to enter disproportionate numbers of pupils for vocational qualifications, to adapt their entrance policies and introduce additional GCSE options. We can certainly envisage that it will take some time for changes to curriculum patterns to ‘bed in’, especially at schools where the vocational qualifications that will no longer be approved for PTs have been more widely used. For these schools, we might expect relatively lower value-added scores over the next few years.

Pupil key stage 4 outcomes are a function of qualifications entered and the quality of the grades achieved. Schools will need to ensure that all pupils follow a curriculum that suits them but which also meets the precepts of the key stage 4 outcome measure. Some pupils will not take eight approved qualifications and so a balance will have to be struck by schools between entries and grades achieved. The choice of scaling options, as described in section 5e, will play a part here.

The relationships at school level between key stage 2 prior attainment, key stage 4 entries, key stage 4 outcomes and value-added were summarised in Table 1 above. value-added scores are heavily correlated with mean key stage 4 entries and indeed with the outcome measure. In other words, for many schools, raw and value-added measures tell essentially the same story.

At pupil level, those pupils who enter the full ten qualifications (eight plus double weighting for English and mathematics) achieve above average value-added scores regardless of prior attainment (Table 13). Indeed, when number of entries is included in the simple model, the proportion of variance explained increases from 58 per cent to 84 per cent.

Table 13: Mean value-added scores by prior attainment and entries (all schools)

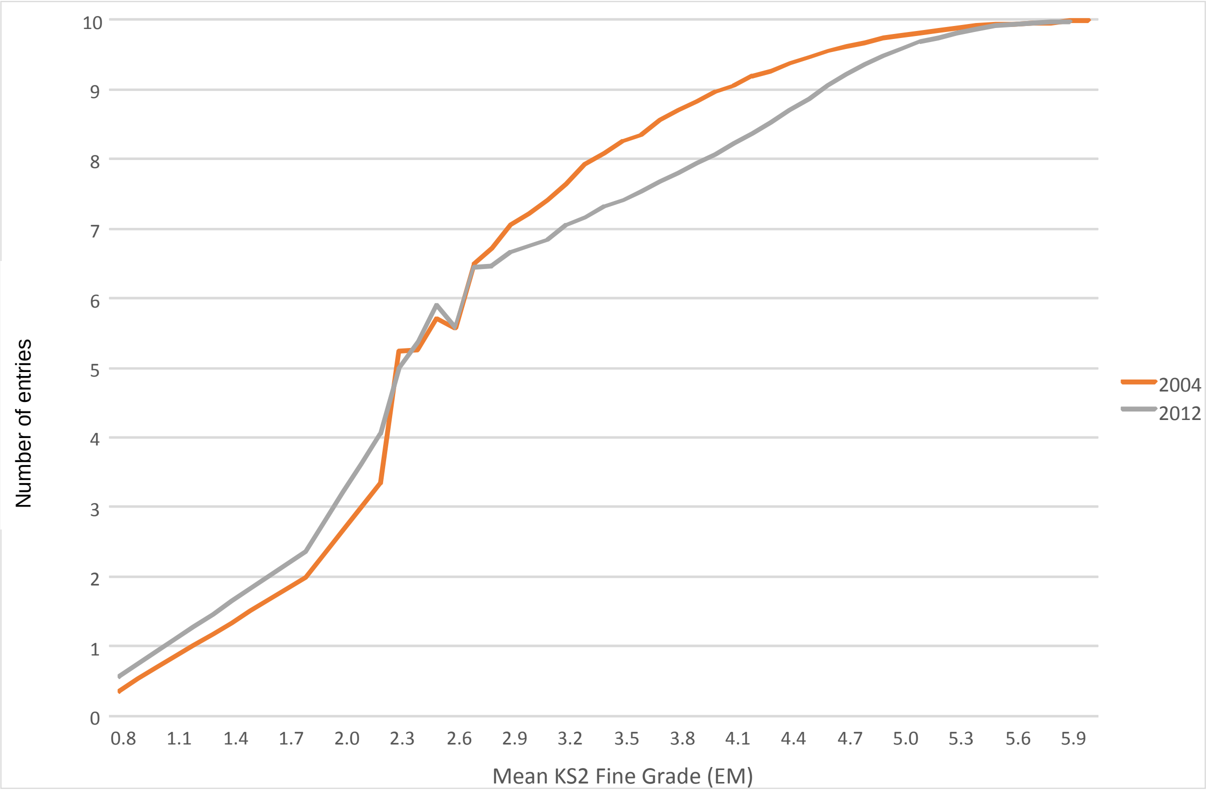


Prior Attainment Quantile		Entries Counted						
		<5	5 to 6.99	7 to 7.99	8 to 8.99	9 to 9.99	10	All
Mean VA	Bottom 5%	-80.4	-32.6	28.0	51.5	88.2	157.7	0.3
	19	-181.1	-83.2	-7.4	20.4	56.1	114.7	-0.2
	18	-212.6	-111.3	-23.9	6.6	45.4	104.9	-0.3
	17	-236.5	-126.1	-38.0	-3.7	38.3	96.8	0.7
	16	-257.2	-141.1	-50.2	-15.1	28.9	88.5	-0.6
	15	-270.8	-153.7	-61.5	-24.6	20.1	80.1	0.0
	14	-288.5	-164.5	-71.8	-35.1	12.1	73.3	-0.3
	13	-296.5	-176.5	-81.8	-44.2	4.0	67.0	0.6
	12	-315.7	-183.4	-92.0	-54.2	-4.3	60.2	-0.3
	11	-327.7	-200.1	-103.7	-62.6	-14.0	53.1	-0.2
	10	-341.6	-209.9	-113.8	-71.4	-20.7	47.1	0.2
	9	-346.4	-218.7	-123.3	-81.1	-30.3	41.6	0.4
	8	-363.0	-231.6	-134.1	-89.2	-36.2	35.9	0.4
	7	-374.8	-234.9	-143.7	-95.4	-44.8	30.9	-0.3
	6	-389.2	-244.5	-155.3	-103.7	-53.1	26.3	-0.1
	5	-408.2	-252.9	-163.6	-109.8	-59.6	21.5	-0.2
	4	-418.6	-269.8	-173.0	-115.2	-66.5	18.5	-0.4
	3	-431.5	-251.5	-192.9	-116.9	-72.7	14.4	0.4
	2	-441.7	-243.2	-192.1	-120.9	-82.3	10.9	-0.1
	Top 5%	-482.5	-223.1	-191.8	-119.0	-87.3	7.7	0.1
	Total	**-164.8**	**-117.1**	**-51.7**	**-30.2**	**-4.0**	**36.8**	**0.0**
Number of pupils	Bottom 5%	7617	4350	6984	4275	2132	995	26353
	19	1804	3056	9049	6304	4140	2447	26800
	18	1120	2120	7965	6619	5139	3631	26594
	17	781	1585	6951	6074	5683	4870	25944
	16	677	1317	6290	5802	6003	6497	26586
	15	521	1087	5642	5485	6307	8262	27304
	14	425	873	4757	4882	6315	9465	26717
	13	358	735	3949	4391	6418	11104	26955
	12	330	622	3300	3734	6036	12319	26341
	11	229	515	2763	3323	6088	14195	27113
	10	211	414	2123	2777	5524	15277	26326
	9	147	331	1683	2410	4995	16528	26094
	8	156	280	1352	1971	4689	18347	26795
	7	123	239	1068	1607	4181	19008	26226
	6	95	174	807	1436	3709	20873	27094
	5	66	158	585	1148	3006	21404	26367
	4	55	120	408	934	2587	21409	25513
	3	40	94	290	798	2055	24364	27641
	2	28	92	182	630	1435	23957	26324
	Top 5%	28	105	86	561	802	25225	26807
	Total	**14811**	**18267**	**66234**	**65161**	**87244**	**280177**	**531894**

We are looking here, of course, at historical entry patterns. It is likely that many schools will adapt their curriculum offer to meet the challenges of the new accountability framework. We might therefore expect higher proportions of pupils to be entered for the full ten approved qualifications in future years. In 2012, 52 per cent of pupils entered the full ten. 93 per cent entered ten when other qualifications (those that will no longer be counted in Performance Tables) are included.

In the past, pupils tended to be entered for more qualifications approved under the new proposed accountability framework (Figure 7). In 2004, the year Section 96 qualifications were first included in Performance Tables, 70 per cent of pupils were entered for 10 or more qualifications. This was especially the case in the key stage 2 mean fine grade range from 3.0 to 5.0. On average, pupils with a mean fine grade of 4.0 were entered for 0.9 fewer GCSEs in 2012 than 2004.

Figure 7: Mean 'Attainment 8' entries by prior attainment (English and maths double weighted), 2004 and 2012

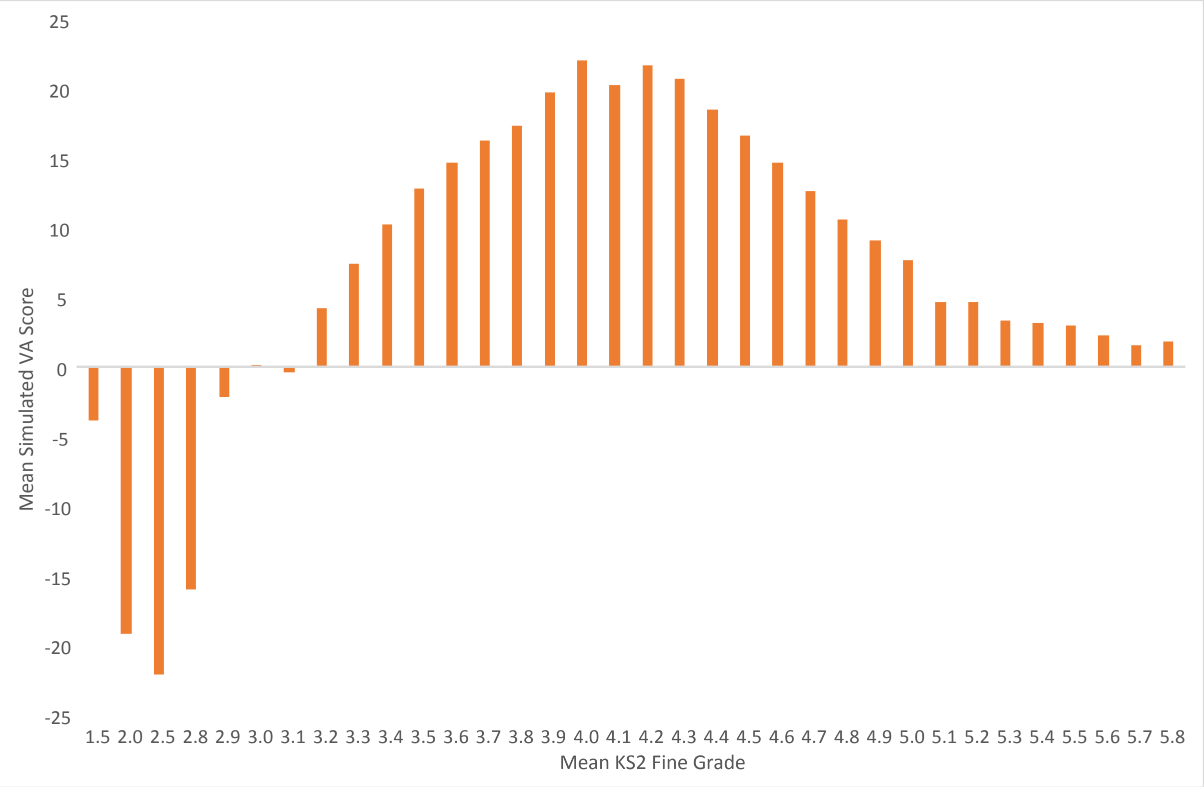


Pupils who reached the end of key stage 4 in 2012 had higher levels of key stage 2 prior attainment than their peers in 2004. Using information about 2004 entry patterns, we undertook a small simulation exercise in order to examine the possible impact of schools changing their entrance policies. We estimate that 76 per cent (rather than 52 per cent) of all 2012 pupils would have entered the full 10 approved qualifications counted in the measure based on 2004 entry patterns.

For each pupil in the 2012 dataset attending a mainstream school, we generated a random percentile from a uniform distribution. We then looked up the number of entries from the 2004 dataset for pupils of equivalent prior attainment (same key stage 2 finely graded points score in English & maths rounded to one decimal place) based on the random percentile. For example, the 50th percentile (median) for pupils with a key stage 2 Average point score of 25.0 in 2004 was ten.

We then simulated the effects of increased entries on pupils' 'Attainment 8' points scores. To do this, we calculated the average points per entry in approved qualifications from 2012 and multiplied by the *simulated* number of entries. Retaining the predictions from our simple model, Figure 8 summarises the consequential impact on value-added scores by mean key stage 2 fine grade.

Figure 8: Mean simulated value-added by mean key stage 2 fine grade (all schools)



Aside from the small number of pupils with low prior attainment, who tended to enter more qualifications in 2012 than in 2004, pupils in the middle of the prior attainment distribution would be likely to achieve above average value-added scores if predictions from the 2012 model were used as a basis for comparing actual attainment in future. This simulation exercise hints that any predictions would under-estimate the potential of lower-middle to middle ability pupils until entry patterns in approved qualifications achieved equilibrium.

Changing entry patterns will therefore have an impact on the stability of predictions over the next few years. This will be particularly important if predictions are to be given to schools ex ante based on the progress of a previous cohort of pupils. It would be advisable to defer a decision on doing so until such time that a degree of equilibrium has been achieved. This could be monitored each year, for example, by producing an internal diagnostic value-added model that includes entries as an independent variable and calculating its effect size. When this falls below a certain threshold (e.g. 0.2), a suitable level of equilibrium could be said to have been achieved.

d. Bias with respect to pupil groups and types of school

Although the models outlined above are unbiased with respect to prior attainment, they are nonetheless biased with respect to other pupil characteristics as shown for a selection of pupil groups in Table 14.

Table 14: Key stage 2 to key stage 4 value-added scores by pupil group

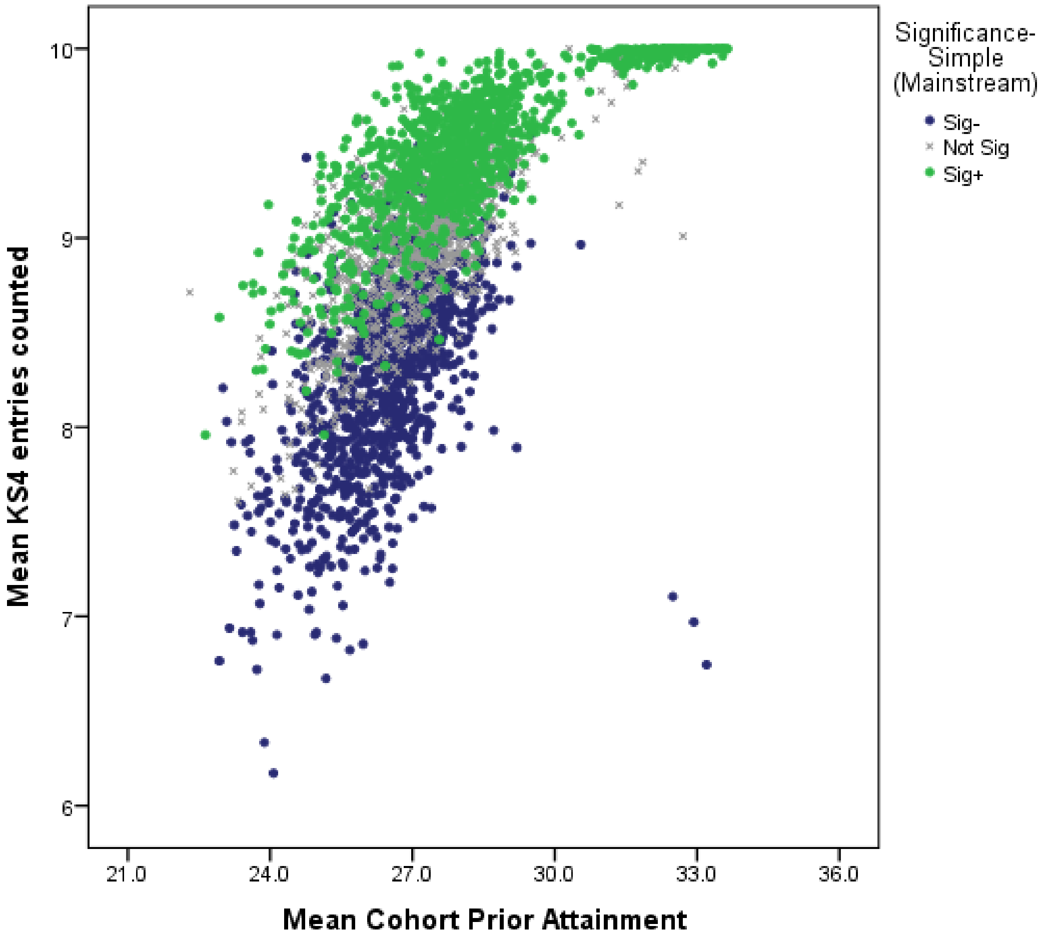
	Group	Mean VA Score	% below trigger	% above upper bound	% all pupils
Gender	Girls	10.0	22%	39%	49%
	Boys	-9.8	30%	27%	51%
Ethnicity	White British	-6.5	28%	29%	80%
	White Irish	4.1	23%	35%	<1%
	White Irish Traveller	-95.5	61%	8%	<1%
	White- Gypsy/ Roma	-66.4	55%	11%	<1%
	White (other)	35.9	15%	52%	2%
	Mixed White/ Black Caribbean	-17.5	33%	25%	1%
	Mixed White/ Black African	13.5	20%	40%	<1%
	Mixed White/ Asian	12.3	20%	41%	1%
	Mixed (other)	10.6	22%	39%	1%
	Indian	42.6	10%	56%	2%
	Pakistani	21.1	19%	42%	3%
	Bangladeshi	31.6	16%	48%	1%
	Asian (other)	49.6	10%	60%	1%
	Black African	40.9	23%	34%	2%
	Black Caribbean	5.6	12%	53%	1%
	Black (other)	14.4	21%	39%	<1%
	Chinese	52.5	8%	67%	<1%
	Any other group	49.0	12%	58%	1%
	Not obtained	-4.2	30%	36%	<1%
	Refused	5.7	24%	36%	1%
Pupil Premium	Not eligible	8.6	21%	36%	76%
	Eligible	-26.6	39%	23%	24%

Special Educational Needs	Not SEN	7.1	23%	35%	79%
	School Action	-14.8	34%	25%	12%
	School Action Plus	-53.2	49%	17%	6%
	Statement	-21.8	37%	24%	2%
	All pupils	0.0	26%	33%	524,677

There is a 35 point difference between pupils eligible for the pupil premium and their peers, for example. Some of these differences are partially explained by variations between groups in entry patterns. Pupils eligible for the pupil premium were entered for 8.2 qualifications (out of a maximum of 10) on average, compared to 9.2 among their peers. The gap would effectively be closed by the pupil premium group taking another GCSE and achieving it at grade D.

Moreover, value-added scores based on the new ‘Attainment 8’ measure currently favour those schools with either a high ability intake or an above average number of entries at key stage 4 (Figure 9), because schools with lower attaining intakes will probably have equivalencies as part of their curriculum offer, whereas less able pupils in schools with generally more able intakes are likely to enter the same non-equivalency based curriculum as their more able peers.

Figure 9: Mean cohort prior attainment and mean key stage 4 entries counted in the value-added added measure by statistical significance of value-added score, mainstream schools



416 of the 599 schools in the top quintile based on cohort prior attainment achieved a value-added score significantly above average at the 95 per cent confidence level. This compares to 119 of the 599 schools in the lowest quintile.

e. Rescaling

Provided the point score intervals between grades are grades are perceived to be reasonable and fair – that is, the achievement of each higher grade represents the same degree of teaching difficulty - then the incentive for schools to concentrate their efforts on specific grades will diminish. However, other performance indicators may still encourage a focus on grade C.

At present, points scores associated with GCSE grades range from 16 (G) to 58 (A*) with equal intervals (6 points) between grades. The sizeable gap between no achievement (0 points) and a grade G means that quantity of entries has a sizeable, in theory if not in practice, impact on points scores. For instance, four grade G passes currently amass more points than a single A* pass.

However, this was not always the case. Prior to the inclusion of Section 96 qualifications in Performance Tables calculations, points scores ranged from 1 (G) to 8 (A*). Under this scheme, four G passes amassed half the points of a single A* pass.

Given that entry level qualifications, which amass fewer points than grade G, are no longer to be counted, now is an appropriate juncture to consider a points score that strikes a suitable balance between quality of grades achieved and quantity of qualifications entered.

We show in Table 15 five versions of a rescaled points score. All apart from V5 (the previous 1-8 scale) maintain the range of 1 point for Grade G to 10 points for Grade A*. V1-V3 are arbitrary attempts to impose a 1-10 scale on the current set of GCSE grades. V4 was determined by examining the intervals from an inverse normal distribution function of GCSE English and mathematics grades achieved in 2012. Consequently, this attempts to represent the level of difficulty between grades.

Table 15: Rescaled points scores

	V1	V2	V3	V4	V5
A*	10	10	10	10	8
A	9	8	8	8	7
B	7	7	6	6.5	6
C	6	5	5	5	5
D	4	4	4	4	4
E	3	3	3	3	3
F	2	2	2	2	2
G	1	1	1	1	1

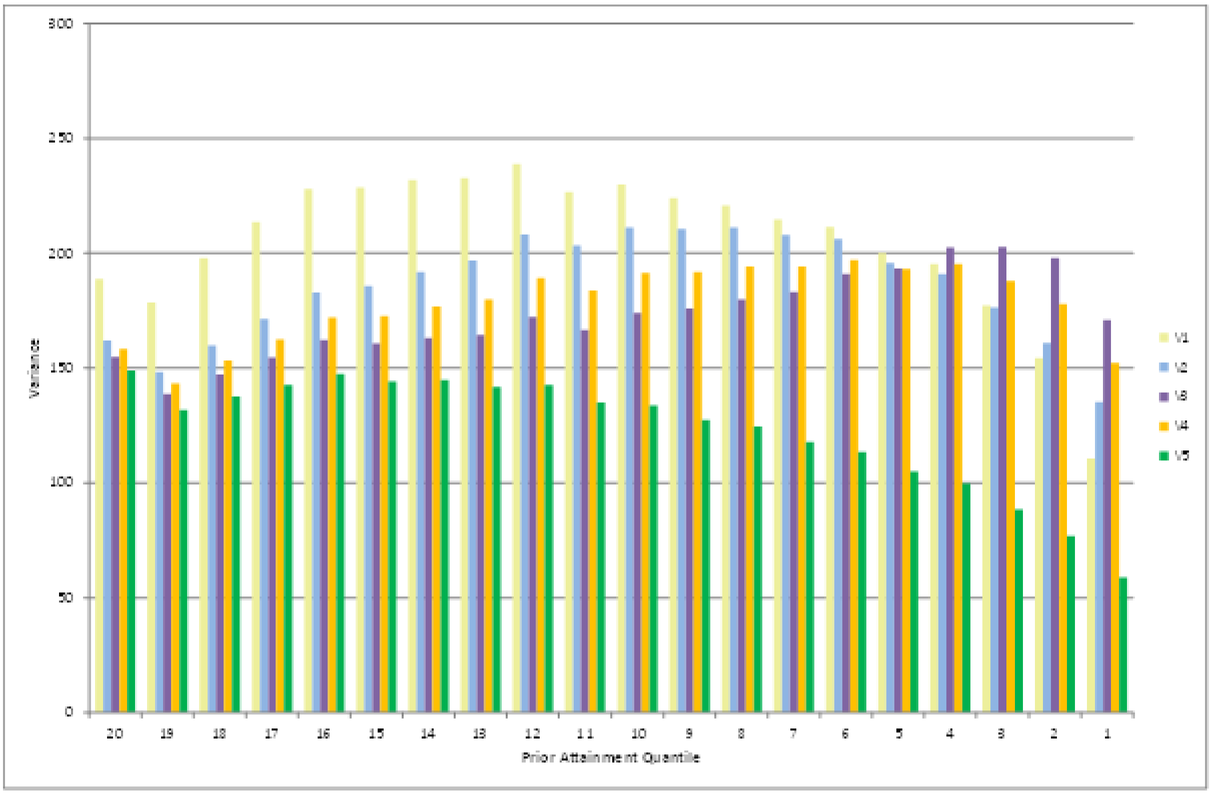
We note that different scales have different quantity: quality ratios. For example, under the V1 scale, 5 Grade Cs (30 points) achieves more points than 7 Grade Ds (24 points), whereas under the V4 scale, the points achieved are 25 and 28 respectively.

Figure.10 (below) compares and contrasts the variance in rescaled points scores. V3 and, especially, V4, have much flatter functions which we argue is more desirable. A uniform distribution of unexplained variance in pupil value-added scores will ensure that the standard error and ensuring confidence interval of the school value-added score is fair to all schools whatever the key stage 2 distribution of their pupils.

Rescaling therefore has the advantage of reducing heteroskedasticity. Consequently, a simpler method of defining the floor target (e.g. based on a school’s mean value-added score) could be used. However, scaling could also influence behaviour in both intended and unintended ways. Under V4, equal intervals between grades are not assumed. This might encourage a focus on pupils at grade C and above, and on the A*/A border in particular, at the expense of lower ability pupils.

Note that all the analysis in this section is necessarily *ex post*.

Figure 10: Variance by prior attainment quantile, rescaled points (All Schools)


Variance
Prior Attainment Quantile
V1
V2
V3
V4
V5