Cambridge Assessment

Volatility happens: Understanding variation in schools' GCSE results

Research Report

Cara Crawford & Tom Benton

5 April 2017

Author contact details:

Cara Crawford
Assessment Research and Development,
Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
UK

crawford.c@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

How to cite this publication:

Crawford, C. and Benton, T. (2017). *Volatility happens: Understanding variation in schools' GCSE results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Table of contents

Introduction

Schools are often judged by their students' performance. Student performance reflects schools' effectiveness, but it also reflects individual intellectual ability, motivation to succeed, and assessment difficulty. This presents a problem: if students' exam scores are only partially attributable to teaching quality, then to what extent should schools be held accountable for changes in cohort attainment from year to year? The stakes are high—school performance measures are used by government to identify schools requiring intervention, by parents to decide where to educate their children, and by policymakers to evaluate the impact of new teaching initiatives. And there is currently little consensus: schools and their representatives point to unexpected swings in cohort achievement as clear evidence that the exams themselves are inconsistent (HMC, 2012; Mason, 2016), while exam boards have shown that marking quality, and the choice of grade boundaries cannot explain test score volatility (Bramley & Benton, 2015).

At the same time, the UK Department for Education's Office of Qualifications and Examinations Regulation (Ofqual), has acknowledged, but not explained, the problem (Ofqual, 2015; 2016). One recent report notes that "a few hundred centres have variations outside of ± 15 percentage points," despite "the assumption…that an individual centre's results are unlikely to be very different in two consecutive years" (Ofqual, 2016, pp. 4-5). Furthermore, Ofqual writes that the observed volatility is particularly surprising because "in years when specifications and overall cohorts are stable, one might expect the majority of centres with entries in successive years to have similar outcomes (e.g. centre variability within ± 10%), and few centres to have substantial variability (e.g. >± 30%)" (p.4).

This paper argues that volatility is normal, quantifiable, and predictable. We aim to model cohort-level volatility (i.e., changes in results between students taking a subject in different years at the same school) so that teachers and policymakers can have more informed discussions about changes in scores from year to year. We start with a simple premise: that school-level metrics are composites of student-level indicators. It follows, then, that school-level performance is directly influenced by student performance, and furthermore, that uncertainty in school performance is a direct result of uncertainty in student outcomes.

Our methodology stems from these inferences. We first estimate how well each student will perform on an exam, then estimate how well cohorts of students will perform as a group, and finally compare the estimates for different cohorts at a school in consecutive years. This lets us identify two things: the most likely change between any two years' cohorts at a school, and a range around that most likely change, which, if observed, would not be cause for concern. Critically, the methodology allows for different ranges of volatility to be expected of different schools.

We will develop this model in more detail, use it to predict the volatility in school performance that occurred in schools' GCSE results in recent years, and then evaluate the accuracy of the predictions. In doing so, we will show that most of the variation in schools' outcomes occurs for sensible and benign reasons: because of the well-known, predictable-if-not-intuitive influence of chance in an indeterministic system; and because of the measurable changes in students' capabilities in different years.

The implications are substantial—quantifying expected volatility will empower schools to proactively manage stakeholders' expectations, and to interpret changes in performance appropriately. In addition, by conceptualising new performance scores as values selected

from a range of possible outcomes, schools may be less likely to blame exam boards for undesirable—but not unexpected—changes in results.

Sources of uncertainty

Our model considers three types of student-specific uncertainty. First, there is uncertainty in how any individual student will perform on a particular exam, given his or her overall abilities. Second, there is uncertainty about the composition of a cohort relative to the population of students from which the cohort is selected. Third, there is uncertainty about how the population of students feeding into a particular school changes over time. These three sources of uncertainty are additive; changes in a population of students can add to the uncertainty of student outcomes, as can greater variation among students within the student population at a given time. In general, a more stable student population, and a larger cohort (such that the cohort is more likely to represent a similar distribution of ability levels to the population) should be associated with lower levels of variation in school outcomes from year to year. However, a decrease in uncertainty at a higher level cannot reduce uncertainty at a lower level. For example, fewer changes in the student population cannot make up for variation in cohort composition over time due to different groups of students being selected in different years out of a large and diverse population.

The way these different types of uncertainty influence the uncertainty in our school performance measurements depends on how exactly performance is measured. Because the UK education system grades students on a 9-point scale (grades A*-G and U) but evaluates schools on the percentage of their students earning grades at or above a single fixed point on the scale, students of different abilities will exert different amounts of influence on their schools' ratings. For example, two students could have the same level of uncertainty in which letter grade they receive, yet have different levels of uncertainty with respect to their likelihood of achieving a minimum C grade.

School performance indicators

The UK Department for Education assesses schools by examining student performance on national examinations taken at the end of Key Stage 2 (KS2) and at the end of Key Stage 4 (KS4). This paper considers student performance on the most common type of KS4 assessment, the GCSE. GCSEs are offered in many subjects; each subject is considered a standalone qualification that can be taken independently from all other qualifications. Qualifications may contain one or more components, or sub-tests. Each component is marked by itself, and then marks from all components in a qualification are combined into an overall score. Finally, these scores are translated into letter grades that are awarded to students. Currently, most GCSE qualifications are graded on an 9-point scale on which A* is the highest possible score, followed by the letters A, B, C, D, E, F, G, and U. Letter-grades closer to the beginning of the alphabet indicate higher achievement. Typically, students take 8-10 qualifications that cover subjects studied over two years.

In the UK, school performance is often assessed by calculating the percentage of students in a cohort scoring at least a specified grade on one or more qualifications. The most frequently used cut-off grades are A and C. The percentage of students in a cohort who score at least Grade C in a qualification is the most common metric used by schools to evaluate performance in particular subjects. This statistic is also used by exam boards to ensure comparability between different versions of an assessment. The UK government sets floor standards for schools in terms of the minimum percentage of students that must score at least a C on five different GCSEs, two of which must be English and Maths. This paper addresses changes in the proportion attaining a C or above on a subject-by-subject basis; future work can apply this methodology to more complex outcome measures.

An additional measurement used by some schools, especially those with many high-achieving pupils, is the percentage of students in a cohort achieving an A-or-A* on a subject test. Which of these two boundaries is more informative for a given school (A-or-above or C-or-above) will depend on the typical attainment of its students: to track school performance changes via test scores, one must select measurements that are sensitive to changes in performance. In schools that selectively admit candidates with the strongest academic performance, the percentage of students achieving grade C is likely to be close to 100% in all years. Therefore, these schools often monitor cohort performance relative to grade A. If selective schools draw many of the country's brightest students, then non-selective schools may tend to enrol mostly students of average ability, and these schools may be unlikely to see many students achieving an A-or-A*'s. In this case, the C-grade would be a more informative boundary.

Despite the upcoming changes to the grading scale (which is due to change from one ordinal scale, based on letters A*-U, to another ordinal scale, based on integers 9-1 and U) and changes to the methods used to rank schools (instead of %C's these will soon have a new measure called "Progress 8"), the methods used to evaluate the soon-departing grade scale can be easily adapted to the new scale. In addition, the percentage of students scoring at least grade C and A (or their integer equivalents) will continue to be a meaningful indicator of performance as, for example, exam boards continue to rely on these percentages when setting grade boundaries for individual qualifications.

To summarise, this paper will operationalize school performance using two variables: (1) the percentage of students in a cohort at a single school achieving a grade between A* and C on a single GCSE qualification (% C-or-above); and (2) the percentage of students in a cohort at a single school achieving a grade of A or A* on a single GCSE qualification (% A-or-above). Volatility will refer to changes in these percentages from year-to-year at a single school.

Defining Volatility

Volatility can mean many things: some apply the term to any frequently-changing value; others reserve it for unpredictable changes (e.g., financial markets treat volatility as risk). Any type of numerical value that is measured repeatedly over time can be volatile. In addition, volatility is not restricted to the change in a value itself—the rate at which a value changes (first derivative) could be volatile, as could the rate of change in the rate of change (second derivative), and so on.

Educators often claim that test scores are volatile when the performance of a cohort (in terms of the percentage of students scoring above a C or above an A) changes by what they consider to be a large amount. The problem is that there is no justification for the claims about what qualifies as a "large amount."

Volatility is also a statistical term that refers to the standard deviation of a distribution. That is, volatility describes the typical variation of a value around its average. For example, imagine that the last Thursday and Friday had the same average temperature of 17 degrees Celsius. On Thursday, the temperature varied throughout the day from a minimum of 16 degrees to a maximum of 18 degrees. On Friday, the temperature varied from 10 to 24 degrees. Despite the same average, Friday's temperature varied more around the average value, and therefore we could say that the temperature on Friday was more volatile than the temperature on Thursday.

This paper uses the statistical definition of volatility: volatility is the standard deviation of a distribution of values, such that a higher standard deviation indicates greater volatility. By distribution of values, we mean that each individual value, such as the average exam score in a particular year, is a member of a larger set that contains values of the same measurement taken at different time points. For example, we could talk about the set of percentages of C-or-Above grades on the Maths GCSE for the past 10 years in a particular school, and this set would contain 10 values, each calculated based on student performance from a different year.

There are several advantages to using the statistical definition of volatility to investigate variability in cohort performance across multiple years. First of all, we can get a more precise estimate of the typical level of performance of cohorts at each school. If we consider pairs of years, then one year of extremely high or low scores will skew the reference value for the following year. However, by averaging over a few years, new scores can be compared against a realistic estimate of typical performance. In addition, value judgments about how much volatility is "too much" can be replaced by quantitatively-informed expectations about how much scores typically vary from a school's long-term average. Finally, by comparing schools on their statistical distributions of scores, we can use statistical theory to gain meaningful insights into how factors such as the number of students in a cohort and the average grade obtained by these students will affect the amount of variation in school-level scores that will occur even across stable cohorts.

Measuring Volatility: from student grades to school grades

First, we shall examine how school performance is calculated from student performance. Students receive grades from A*-U. These grades are then partitioned into two groups, separated by a grade boundary. We use the following student-level variables:

Student performance variables

C or above.
This variable equals 1 if an individual student scored a grade between A* and C on a single qualification and equals 0 otherwise.

A or above.
This variable equals 1 if an individual student scored an A or A* on a single qualification and equals 0 otherwise.

The table lists each possible student grade on a qualification and the corresponding values that would be recorded on the two student-level binary performance variables.

Table 1. Student performance indicators

Letter Grade	C or Above	A or Above
A*	1	1
A	1	1
B	1	0
C	1	0
D	0	0
E	0	0
F	0	0
G	0	0
U	0	0

These variables are summed over each cohort in each school and then divided by the number of pupils in each cohort, resulting in the following cohort-level variables:

School performance variables

% C or above.
The percentage of students in a cohort who score a grade between A* and C on a single GCSE qualification. This is a value between 0 and 100 percent.

% A or above.
The percentage of students in a cohort at a single school who score an A or A* on a single GCSE qualification. This is a value between 0 and 100 percent.

Our model uses student characteristics to predict the probability of a student achieving a grade at or above the A or C boundary (i.e., the probability that the A-or-Above variable is equal to 1, etc.), and then groups students together into cohorts to estimate the percentage of students at or above each grade in a given cohort (the school-level performance variables).

Building the Model: probability and random sampling

Sampling students
First, let's clarify our assumptions. We consider schools to be made up of cohorts and cohorts to be made up of random samples of students from a local area. This makes sense if one considers that proximity is one (if not the only) main determinant of which school a student attends: each school draws from a "population" of students that live in the surrounding neighbourhoods; these students are assigned to cohorts based on their dates of birth and nothing else. Because the ages of children in one family should be independent of the ages of children in other local families, we can treat the ages of different children as independent and identically distributed random variables. Furthermore, we can say that cohort selection occurs by independent random selection of students from this population.

Probability theory has developed simple ways to discern the relationships between random variables and to estimate how changes in individual units will yield changes to a group of these units (i.e., the joint distribution of many random variables). The fact that we can reasonably treat cohorts as random groups of students who happened to be born in the same year means that we can use these theories about random variables to understand and predict volatility in cohorts of students. Specifically, we can estimate the expected year-to-year variation in scores that will typically occur, just because of chance, within each school.

The methodology will be explained with an example from an imaginary school; this will motivate the approach used to analyse real performance data.

Sample student
Imagine that a single student is preparing to take a GCSE exam in Maths. This student is generally very good at maths and generally a good test-taker. While there is no way to predict ahead of time exactly how that student would perform on the particular version of the GCSE maths exam taken on a particular day, the student's maths teacher could make a reasonably accurate guess as to the student's most likely grade on the exam. Imagine that this student's maths teacher predicted that he would get an A on the GCSE Maths test.

Despite the student being likely to get an A, there is some uncertainty. Perhaps on exam day, he is particularly well-rested, energised by a filling breakfast and confident in his abilities. Furthermore, perhaps the questions on exam happen to exactly match the subset of course materials that the student used for revision. In this situation, he might score an A*.

Alternatively, perhaps the student happens to perform much worse than usual, and does not accurately demonstrate his mathematical prowess. In this case, he might receive a B on the assessment. Even less likely, but still possible, is that he performs so poorly that he receives a D or an E on the exam. This might be very surprising, but it would not be impossible.

Next, imagine that instead of simply thinking about the most likely outcome of the exam, we went so far as to assign numerical probabilities to each possible score that our imaginary student might receive on the test. The probabilities would represent the relative likelihood of the student receiving each possible mark, and each value would lie between 0 (impossible) and 1 (certain). If we were to add up the probabilities of each possible grade, the sum would be 1, because if our probability distribution includes all possible grades, then it is certain that the student will get a grade somewhere in that range. We could list the possible grades and the associated probability that the student achieves each of the grades in a table.

Table 2. Likelihood of possible outcomes for a sample student

Letter Grade	Probability grade achieved on maths GCSE
A*	0.31
A	0.38
B	0.24
C	0.06
D	0.09
E	0.005
F	0.003
G	0.0015
U	0.0005

In mathematical terms, this list of all possible outcomes and their associated probabilities is called a *probability mass function (pmf)*, and can be represented graphically as shown in Figure 1.

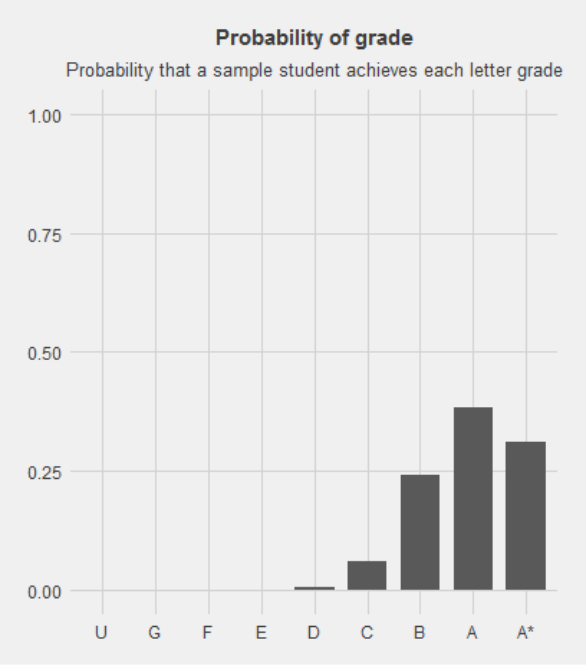

Figure 1. Probability of sample student achieving each possible grade

Next, let's consider two more students who are also preparing for the Maths GCSE. While our first sample student was most likely to get an A, the second is most likely to get a C, and the third, an E. Let's assume that despite differences in their propensities for high scores in math, each student has the same amount of uncertainty in which actual grade he will achieve given his abilities. Figure 2 shows all three students' probabilities of achieving each grade.

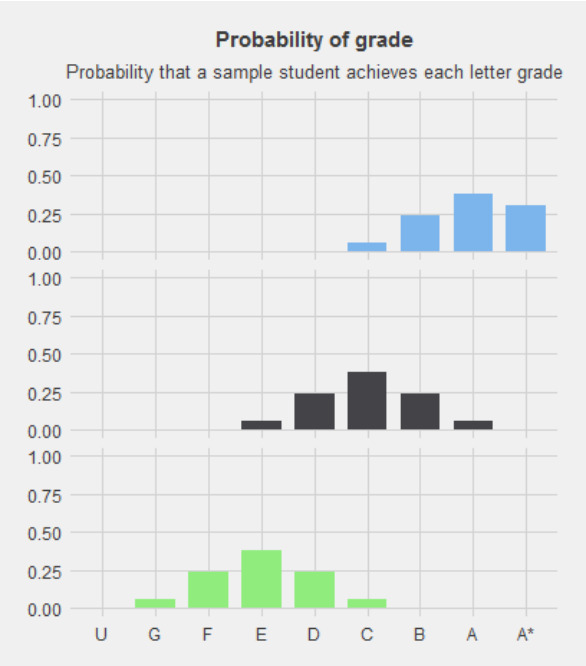

Figure 2. Probability of sample students of different abilities achieving each grade

However, schools are not measured by students' exact grades, but rather by the percentage of students achieving a minimum grade. Thus, to get from student outcomes to school outcomes, we need to translate the probability mass function for a student into a *cumulative distribution function (cdf)*, accounting for all possible ways that a grade at or above the minimum grade could occur. In other words, if a student received an A* on the exam, he would have achieved a grade at or above C, but if the student received an A, a B, or a C, he would also fall into the C-or-above category. To determine the overall likelihood that this student will get a minimum grade of C, we add up the probabilities for all grades within that category. We will start off with a consideration the C-grade boundary; that is, the likelihood that a student will achieve a grade of A*-C on an exam. In Figure 3, a dashed vertical line shows the C-grade boundary, and the darker shading for grades A*-C shows the probabilities of all grades that would fall within the C-or-above category. To the right of each probability mass graph is a cumulative density graph that shows, in dark shading, the sum of the shaded columns in the probability mass graph. For each sample student, the cumulative density graph represents the probability that he or she will achieve a grade of C or higher.

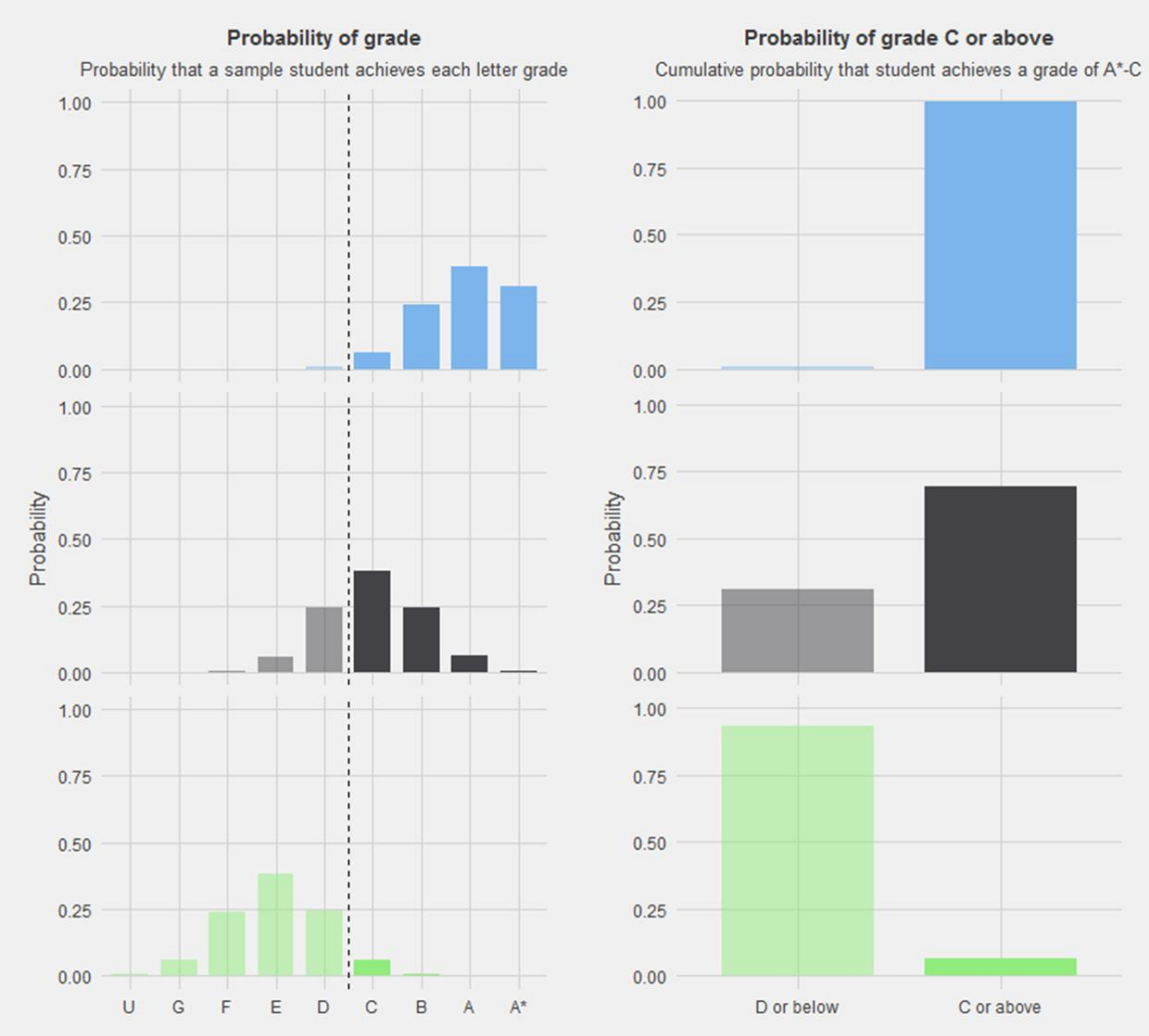

Figure 3. Probability mass function and cumulative density function for sample students' grades (and grades relative to a C).

If we were to consider the level of uncertainty in the actual grade each student received, they would be similar, since in fact all of their predicted grades were drawn from a normal distribution (capped at the endpoints of the grading scale). However, this is not the case when we consider the uncertainty in each student's likelihood of getting at least a C: the student whose performance in general is closest to that required for a C will have much more uncertainty in the C-or-above measurement that the other two, who are both far from C-level candidates, albeit in different directions. This shows an important component of volatility in grades: because we are looking at a fixed boundary drawn somewhere along a larger scale of possible outcomes, students that have a relatively equal likelihood of ending up on both sides of the boundary are most likely to be affected by small changes in any test-related variable. For example, a small change in the test conditions (e.g., temperature of the exam room) might cause a candidate who would have achieved a C to fall just below the C boundary and get a D instead. Whereas the A-level candidate could also fall a whole letter grade for the same reason, it is only the effect on the C-level candidate that would be visible when looking at the distribution of grades at or above grade C (because that candidate would switch sides from at grade C to below grade C, whereas the higher-achieving student would fall by the same amount and yet would still end up comfortably above the C boundary). We therefore are able to distinguish smaller differences in performance among students, and by extension, among cohorts, with many students around a C-level of ability, than among cohorts either above or below a C in ability. This could mean that we are able to pick up earlier on improvements in teaching methods, but it also means that we could see large fluctuations in performance due to other reasons that are not related to the school.

Because we are treating students as independent random variables with a particular probability of a C-or-above on a binary axis, we can easily estimate the likelihood of different possible outcomes for a cohort of many similar students. In the next set of graphs, we imagine that each of the three students is in a cohort of 100 students, the other 99 of whom have the exact same likelihood of success as he or she does. Because low-probability outcomes for one student could be balanced out by higher probability outcomes for other students, there is less uncertainty in the overall percentage of a cohort of 100 that will achieve at least a C than in a cohort of 1 (i.e., our single student with a percentage C-or-above of either 100% or 0%). However, the effect of the increase in students on the uncertainty in the group's outcomes will differ based on the ability levels of the students---uncertainty in a C-level student does not disappear when we have a lot of them. This is an important point: it means that even in a big school with very large and stable cohorts, depending on the ability level of typical students, there may still be substantial volatility in school performance simply because of uncertainty in students' outcomes. This relationship is illustrated in Figure 4, which shows the probability of different numbers of students achieving a grade C or above for a cohort of 100 students of equal ability to each of the three individual students described above.

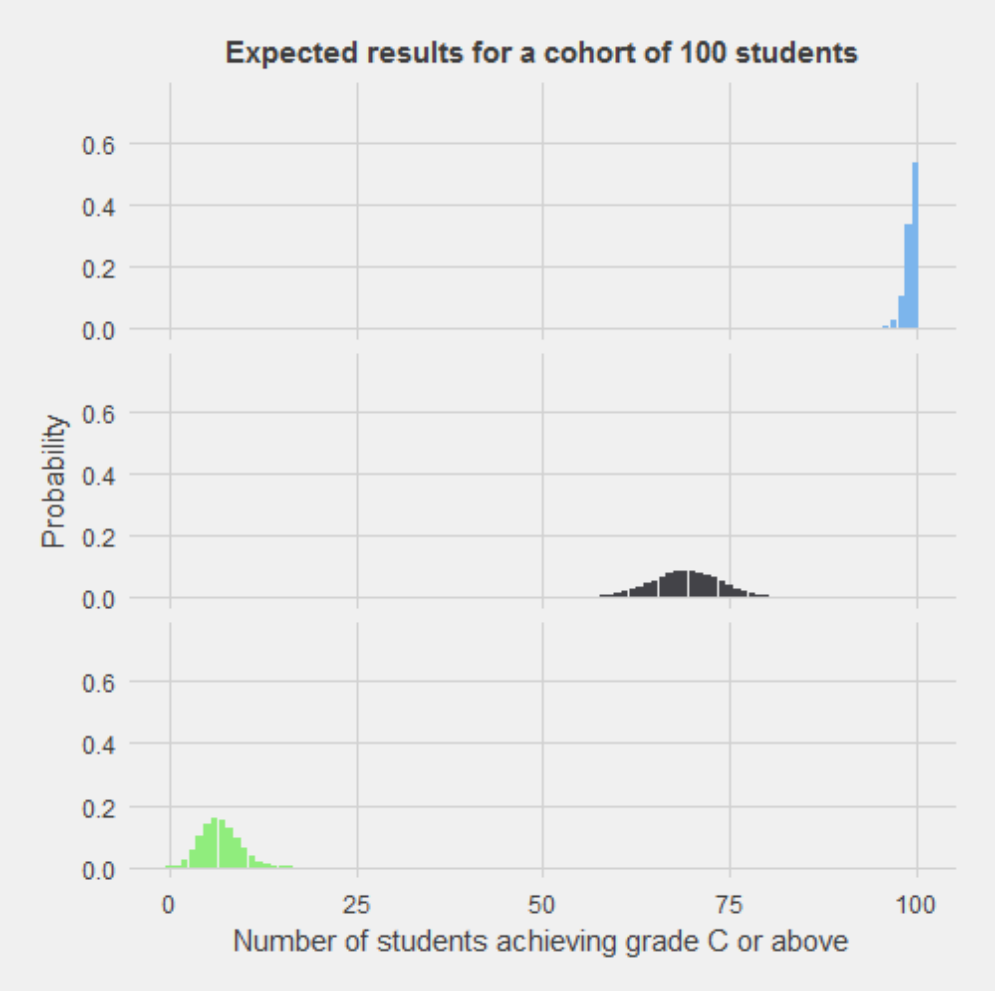

Figure 4. Probability mass functions for the number of students out of a 100-student cohort achieving a grade of C or above.

Another way to think about these differences in uncertainty stemming from uncertainty in students is to consider the likelihood of a change of a certain size in the percentage of a cohort in one year relative to a different year. For example, how likely is it that if our three populations of identically-talented students remain constant, one cohort of 100 would result in a percentage of candidates scoring at least a C that is more than 5 percentage points above or below the percentage for a different cohort of 100 similar candidates? To calculate this, we need to look at the joint distribution of possible outcomes for both cohorts and sum the probabilities of all possible ways that the two groups could differ by more than 5 percentage points. The likely change in the number of students getting a C or above for two identical cohorts of 100 students is shown in Figure 5. The shading represents a change of 5 or more students.

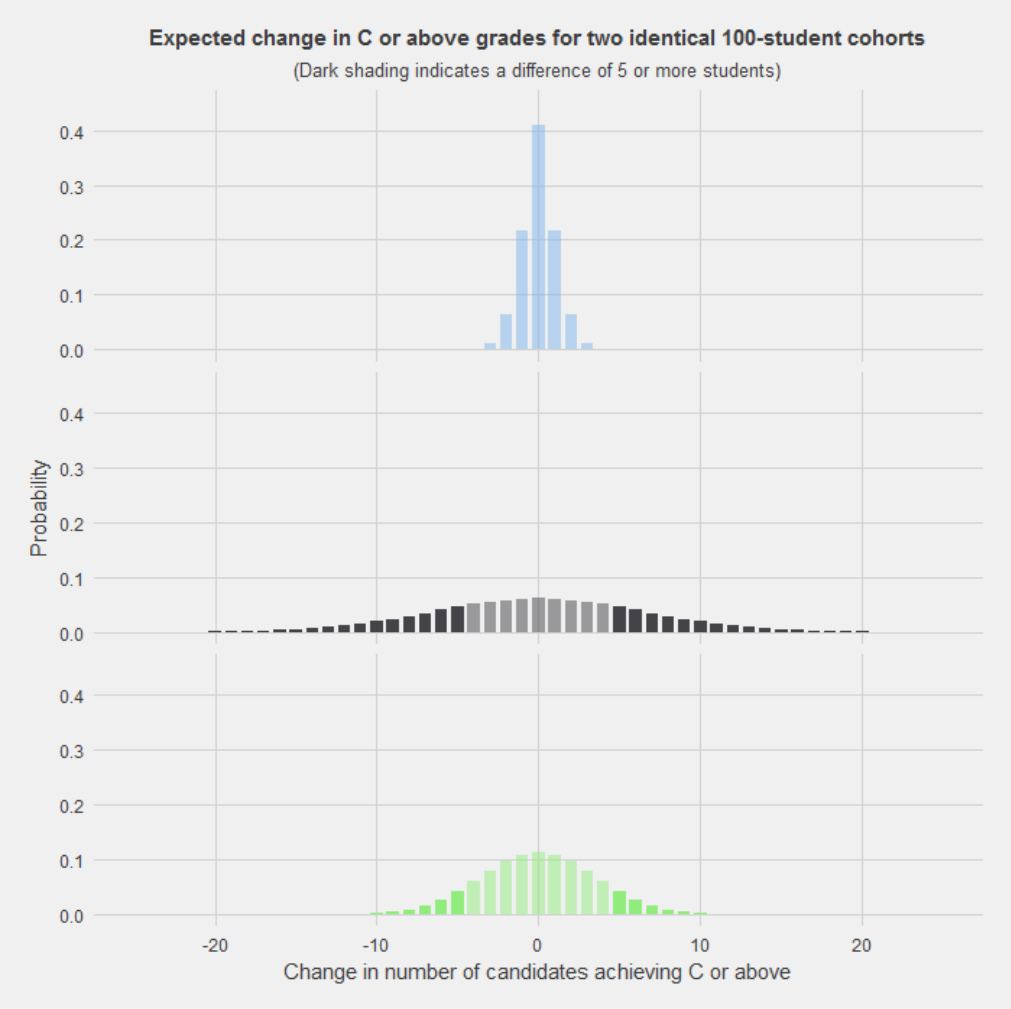

Figure 5. Probability of a change of more than 5 students in the number of students in a cohort of 100 achieving a C or above (the top graph represents a cohort with students of high ability; the middle of students with mid-range ability; the bottom of students of low ability).

In reality, the probability mass functions differ for each student, with some students of very high ability in a cohort with other students of a much lower ability. When the likelihood of getting a C and an A is not constant across students, the calculations get much more complex. However, the intuition is the same: the greater the uncertainty in student scores, the higher the variance will be for the cohort's score.

In addition, over time changes in school performance are not only due to the laws of chance but also due to visible changes in the characteristics of students. This could be due to a shift in the population of students living in the local area (changing neighbourhoods, opening or closing of other nearby school(s), etc.). Alternatively, it could just be a reflection of the underlying heterogeneity in the population, because it is unlikely that the distribution of abilities in any subset of students will be perfectly representative of the distribution of abilities in the population.

We have established how we will look at variance in school outcomes and at the relationship between student grades and volatility from changes in student grades, but we still have not determined how much of the change in cohort scores in recent years can be explained by these changes in students. In order to explore this we leave theoretical notions of schools and turn to empirical work.

Volatility in Schools' 2011-2015 GCSE Outcomes

Returning to the question about the sources of variability in school performance, we need to look not just at the total variation between cohorts but also at how much of this variation is explained by changes in the students, changes in the test, or changes in the school's effectiveness. Furthermore, it is crucial to understand the extent to which year-on-year changes in results are commensurate with the extent of unpredictability in individual students' results. Without these steps, we are only able to note differences that have occurred, but we cannot answer questions about the appropriateness of the observed volatility or about how much volatility we can expect in the future. The best way to identify the sources of all variability in scores is to use the previously outlined relationship between school outcomes and student outcomes, and to construct a model of student-level attainment as a function of each student's characteristics, the cohort he or she belongs to, and the school he or she attends. Because uncertainty in student outcomes is propagated into uncertainty in school outcomes, a model of student achievement can discern the variability in school rankings that is due to something else about cohorts besides the students that they contain.

In the following sections, the Maths GCSE, and specifically, the volatility in the percentage of students in different years achieving grade C or above, will be used as a case study to explain our methodology in detail. We selected Mathematics as the subject to review in depth because it is the largest GCSE by total entries. We use the same methodology to disentangle the sources of variance in cohort-level performance in four additional GCSE subjects: History, Additional Science, English Literature, and Biology. These subjects (and Mathematics) were selected by the number of entries and by content. All five are in the top 10 subjects with the most entries over the five-year period analysed, and they reflect a variety of assessment types. Because of the comparative nature of our analysis, the major results will be presented for all subjects together.

Data

Data were obtained from the National Pupil Database (NPD), supplied by the UK Department for Education (DfE), and contained results for all Year 11 students who took a GCSE in one of five subjects in a June exam session between 2011 and 2015. Data were obtained separately for each subject, and thus, the specific students and schools used in each analysis may differ. To be included in the analysis, we restricted the data by cohort size and by student entry patterns. Specifically, student eligibility required grades for at least three additional GCSEs taken in the same exam session as the subject of interest, and school eligibility required at least 20 eligible pupils entered in each of at least two of the five exam years. Tables 3-7 show, by subject, the number of students included from each year and the percentage of these students achieving a minimum grade of C and A. The bottom row of each table shows the total number of schools, students, and overall percentage of students across the five years achieving grades at or above a C and an A. Note that although the "Students" entry in the bottom row of each table is the sum of the students from each year, the "Schools" total is not the sum of the number of schools in each year because all schools were required to have data from more than one year.

Table 3. Description of Mathematics GCSE data

			Student achievement (percent of students nationally)	
Year	Schools	Students	Grade C or above	Grade A or above
2011	3,254	451,010	68%	21%
2012	3,162	372,105	68%	20%
2013	3,260	386,635	69%	19%
2014	3,280	479,619	71%	17%
2015	3,189	483,499	74%	20%
Total	3,547	2,172,868	70%	19%

Table 4. Description of English Literature GCSE data

			Student achievement (percent of students nationally)	
Year	Schools	Students	Grade C or above	Grade A or above
2011	3,130	401,380	80%	25%
2012	3,137	400,431	78%	24%
2013	3,123	397,344	79%	24%
2014	2,927	381,724	79%	24%
2015	2,554	348,883	78%	23%
Total	3,410	1,929,762	79%	24%

Table 5. Description of Additional Science GCSE data

			Student achievement (percent of students nationally)	
Year	Schools	Students	Grade C or above	Grade A or above
2011	2,886	239,531	70%	16%
2012	2,944	236,641	69%	14%
2013	2,866	231,758	66%	12%
2014	2,953	275,506	67%	13%
2015	2,884	303,185	66%	13%
Total	3,315	1,286,621	68%	13%

Table 6. Description of History GCSE data

			Student achievement (percent of students nationally)	
Year	Schools	Students	Grade C or above	Grade A or above
2011	2,854	178,109	72%	32%
2012	2,924	180,965	72%	31%
2013	3,075	214,684	70%	29%
2014	3,030	216,657	69%	29%
2015	2,888	206,106	69%	29%
Total	3,365	996,521	70%	30%

Table 7. Description of Biology GCSE data

			Student achievement (percent of students nationally)	
Year	Schools	Students	Grade C or above	Grade A or above
2011	2,267	122,951	94%	50%
2012	2,504	140,116	94%	49%
2013	2,375	139,126	92%	43%
2014	1,945	114,245	92%	44%
2015	1,876	109,881	93%	44%
Total	2,813	626,319	93%	46%

Model specifications

For each subject, two sets of models were specified, reflecting the likelihood of students achieving grades C or above, and separately, the likelihood of students achieving grade A or above. The models are given for grade C or above in Equations 1 through 5.

Let y_{ijk} equal an indicator of whether student i in cohort j at school k achieved a C or above, such that

$$y_{ijk} = \begin{cases} 1 & for \quad Grade \geq C \\ 0 & for \quad Grade < C \end{cases} \tag{1}$$

For the simplest possible model, we begin by ignoring the influence of background characteristics and assume the probability that student *i* gets a grade at or above a C is given by the formula:

$$\Pr(y_{ijk} = 1) = \frac{e^{\beta_0 + v_k + u_{jk}}}{1 + e^{\beta_0 + v_k + u_{jk}}} \tag{2}$$

where *v* and *u* represent independent and identically distributed random error terms specific to the school and cohort and β_0 relates to the probability of candidates achieving C or above nationally. A convenient way to rewrite this formula is to take the natural log of both sides so that we can estimate a linear function of the log of the ratio of probabilities of the possible outcomes. This function can be re-written as follows:

$$logit(y_{ijk}) = \beta_0 + v_k + u_{jk} \tag{3}$$

To develop the model, we added 4 binary variables for the specific year in which the exam was taken to account for year-specific differences in the exams that affected all students in all schools. Each binary variable was used to account for the difference between taking an exam in 2011 and each subsequent year. These year-variables were included as fixed effects (one parameter estimated for each, no school or cohort-specific adjustments).

$$logit(y_{ijk}) = \beta_0 + v_k + u_{jk} + \beta_1 * year_{2012} + \beta_2 * year_{2013} + \beta_3 * year_{2014} + \beta_4 * year_{2015} \tag{4}$$

Note that the year 2011 is identified by the other year variables all being set equal to 0.

To develop the model further, and to begin to account for the influence of student characteristics, we added an additional student-level fixed effect, the student's mean GCSE grade (on a 0-8 scale) on all other exams taken in the same year as Maths. This captures student ability: if some common intellectual aptitude is necessary to achieve high grades in any subject, then by averaging across other subjects' grades, we should get an indicator of this common ability-level that is not specific to any single subject. In theory, because we expect students' ability (or general intelligence) to be the most influential characteristic on their likely grade, we expect this model to capture the bulk of the important information about candidates. The model looks like this:

$$logit(y_{ijk}) = \beta_0 + v_k + u_{jk} + \beta_1 * year_{2012} + \beta_2 * year_{2013} + \beta_3 * year_{2014} + \beta_4 * year_{2015} + \beta_5 * meangcse_{ijk} \tag{5}$$

The models for grade A or above differ only in Equation 1, where instead of y_{ijk} indicating whether student i in cohort j at school k achieved a C or above, y_{ijk} indicates whether student i in cohort j at school k achieved an A or above, such that

$$y_{ijk} = \begin{cases} 1 & for \quad Grade \geq A \\ 0 & for \quad Grade < A \end{cases} \tag{6}$$

Having fitted each of the models above, for the purposes of our research, the main focus is on the extent of variation between schools (term v_k in Equations 3-5), and between cohorts within schools (term u_{jk} in Equations 3-5) that remains unexplained by the variables included within each model. In contrast to many other research studies, the model coefficients for the individual variables listed above (i.e., $\beta_0 - \beta_5$ in Equation 5) are of less interest and are not reported in this paper. However, more detailed results are available from the authors upon request.

Results

Schools differ more than cohorts

The first model was run to identify where, and not why, scores varied between students. To do this, we constructed a model of student performance that would partition variance into changes between schools, between cohorts within schools, and between students within cohorts and schools. The results are listed in Table 8. The table shows the variances at school and cohort level (which are read directly from the model results) as well as the total variance across all of the predictors on the logit scale due to characteristics (e.g., exam year) included in the model. Since this first model did not contain any explanatory variables, the variance across predictors is 0. Finally, as suggested by Mood (2010), to account for the uncertainty in the grades that will be achieved by individual students a fixed level of residual variance (of 3.29) is added to each model. This fixed variance uses the properties of the logistic scale to account for the uncertainty in student results after accounting for all other factors. To aid in interpretation, the table also lists each variance component as a percentage of the total variance under the “standardised variance” column. This column shows the percentage of variation in students’ results that is attributable to observable characteristics (predicted), schools, cohorts within schools, and individual students within cohorts. Standardising variances on a percentage scale has the additional benefit of allowing comparisons across models: whilst unstandardized coefficients cannot be compared between logistic models, these standardised variances can be compared directly (Mood, 2010).

When we look at the results from the baseline model, what stands out is the relative percentage of variance occurring at the “school” versus the “cohort” level. The results show that just less than one-third of the difference in performance seen across all of the students can be explained just by identifying which school a student attends, compared to only 4% occurring across cohorts within schools. The large school-level variance component could indicate genuine differences between schools, or it could be due to the differences in the types of pupils that attend different schools. For now, all we can say is that the changes between students in different cohorts within each school (4% of the total variance) are extremely small compared to the differences between students in different schools (29% of the total variance). In other words, what we are calling volatility is just the 4% of performance variability that is associated with a student’s cohort within a school.

Table 8. Variance decomposition for baseline multi-level logistic model of Maths Grade C or above.

	Variance Component	Variance	Standardised Variance
Model 0: No fixed effects, intercept-only (Mathematics)	Total	4.92	100%
	Predictors	0.00	0.0%
	School	1.43	29.1%
	Cohort (within school)	0.20	4.2%
	Student (within cohort)	3.29	66.8%

Looking across subjects at the standardised variances, it is clear that although in some cases, the cohort-specific variance component is higher than others, it is always only a small proportion of the total variation in scores across students. Figure 6 shows the standardised variance breakdown for each subject, with the panel on the left representing the standardised variance in grade A or above, and the panel on the right representing variance in grade C or above. The exact values shown in the figure are reproduced below it in Table 9 (grade A or above) and Table 10 (grade C or above). Looking at grade C or above, we can see that variation between cohorts within schools ranges from 3.1% for History to a maximum of 12.6% for Biology. Turning to Grade A or above, the year-level variance is similar to grade C for Mathematics, Additional Science, and History, with Maths and History consistently having fairly low volatility (3-5% of total standardised variance) and Additional Science having a relatively large volatility at both grade levels (about 11%). Interestingly, English Literature and Biology both have large (compared to History and Maths) volatility at grade C, around 10-12% of total standardised variance, yet these subjects have much lower volatility at grade A.

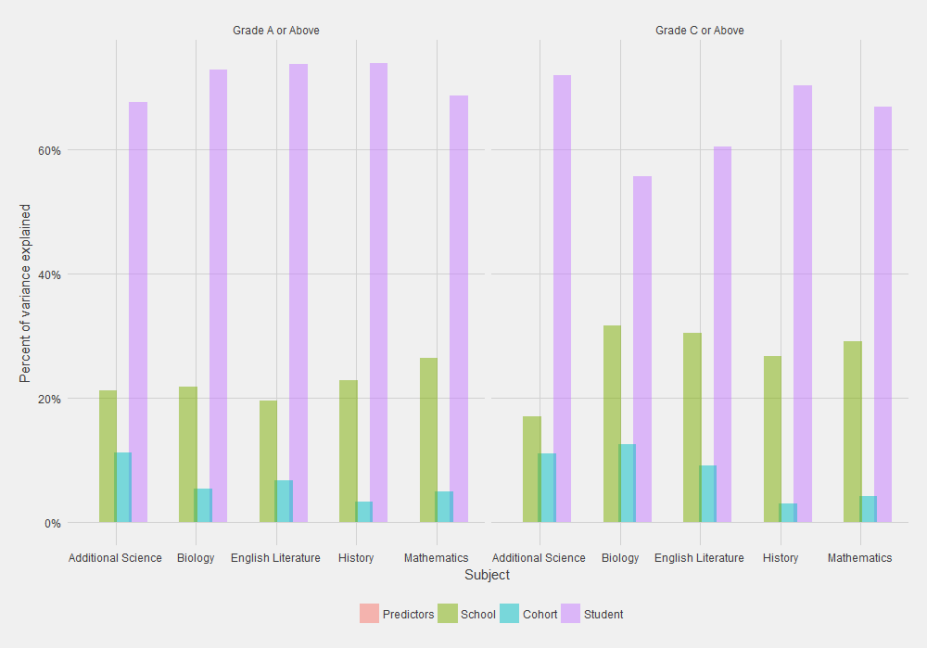

Figure 6. Standardised variance by component for each GCSE subject (base models with no fixed effects).

Table 9. Standardised variance components for Grade A or Above models with no fixed effects.

		Predictors	School	Cohort	Student
Grade A or Above	Mathematics	0.0%	26.4%	4.9%	68.6%
	English Literature	0.0%	19.6%	6.8%	73.6%
	Additional Science	0.0%	21.2%	11.2%	67.6%
	History	0.0%	22.9%	3.3%	73.8%
	Biology	0.0%	21.8%	5.4%	72.8%

Table 10. Standardised variance components for Grade C or Above models with no fixed effects.

		Predictors	School	Cohort	Student
Grade C or Above	Mathematics	0.0%	29.1%	4.2%	66.8%
	English Literature	0.0%	30.4%	9.2%	60.4%
	Additional Science	0.0%	17.1%	11.1%	71.8%
	History	0.0%	26.7%	3.1%	70.2%
	Biology	0.0%	31.7%	12.6%	55.7%

Next, we added in variables to account for the year in which an exam was taken, so that cohorts at different schools could be compared within each year (because the variance between schools could be due to the fact that school-level variance includes differences between cohorts in different schools in the same year as well as differences between cohorts in different schools in different years). However, when the exam year is included, only 0.5% of performance differences between students is predicted, and the amount of variance occurring between schools stays exactly the same, at 29% (see Table 11).

Within schools, the variance between cohorts drops slightly, from 4.3% to 3.6% of the overall performance differences observed in the data. This 16% decrease in cohort-level variance accounts for the differences in performance that occurred because of nationwide changes in the way the test was constructed or marked in different years. For example, the specific questions asked each year change, and this could result in slight differences in the content of the exam overall. The remaining year-level variance estimates the differences in performance that would still occur between cohorts (at the same school) even if the exact same test were given to both groups of students.

Table 11. Variance decomposition for multi-level logistic model of Maths grade C or above with year fixed effects.

	Variance Component	Variance	Standardised Variance
Model 1: Year fixed-effects (Mathematics)	Total	4.92	100%
	Predicted	0.02	0.5%
	School	1.43	29.0%
	Cohort	0.18	3.6%
	Student	3.29	66.9%

These results are not shown for the remaining subjects and models, as the variance accounted for by exam year is never greater than 0.5%, and in most cases is very close to 0.

Different test-takers achieve different test results

In the third model, we added a variable to account for each student's ability level, as estimated by his or her average grade on all other (non-Maths) GCSE's taken in the same year (Table 12). If changes in performance between cohorts are due to changes in the ability levels of the students from one year to the next, then adding in this student ability measure should result in a large drop in the unexplained variance in results. When we look at the predicted standardised variance for model 2, we can see that 63% of difference in the probability of obtaining at least a C between students is predictable if the exam year and a student's average academic performance are known. The predicted variance removed unexplained variance at the other levels; although the largest drop was seen in the standardised variance at the school level (down from 29% to 2.7%), the predictions also reduced the variance at the cohort level, from 4.3% in Model 0 to 2.3% in Model 2.

Table 12. Variance decomposition for multi-level logistic model of Maths Grade C or above with year and mean GCSE fixed effects.

	Variance Component	Variance	Standardised Variance
Model 2: Year and mean GCSE fixed effects (Mathematics)	Total	10.31	100%
	Predicted	6.50	63.1%
	School	0.28	2.7%
	Cohort	0.24	2.3%
	Student	3.29	31.9%

In other words, knowing students' abilities as measured by other GCSEs removes almost half of the unexplained volatility between years. We also constructed a model with additional student (e.g., prior attainment, age) and school (e.g., independent or selective school) characteristics to see if these would capture any of the remaining unexplained variance in the data. Because the additional variables did not add substantially to the models' ability to predict student outcomes, these results are not reported. However, an example of the expanded model used for Mathematics grade C or above can be found in the appendix. Overall, the results show that student ability, as measured by concurrent attainment, explains the majority of variation in student performance on Maths, including those differences between students in different cohorts at the same school.[1]

This remains true for additional GCSE subjects. Figure 7 shows the standardised variance by component for each subject. The left panel shows these results for the models predicting attainment of grade A or above; the right panel shows the same for the models predicting attainment of grade C or above. This information is repeated (with exact percentages given) in Table 13 (grade A or above) and Table 14 (grade C or above). It is notable that just by accounting for students' average GCSE grade in other subjects and the exam year, we are routinely able to explain nearly two-thirds of the variation in scores in any given model. In the figures, the variance labelled "Predictors" is that accounted for by these two fixed effects (year and mean GCSE score). Student variance includes any leftover variation between students not otherwise accounted for in the models.

When the variables for the exam year and students' mean GCSE grade are included, the cohort-level variation in results drops substantially for all subjects except History. The History standardised variance at the cohort level is very low, at around 3% for both grade C and grade A, and this value stays constant even with the additional fixed effects. The results are particularly strong for the Additional Sciences GCSE, where cohort-level variance drops

1 The appendix contains results from a model with additional variables to account for other student characteristics; the total between-cohort variance explained is very similar to that in Model 2.

from just over 11% of total variance at both grade levels to 4.7% for grade C, and to just 2.8% at grade A.

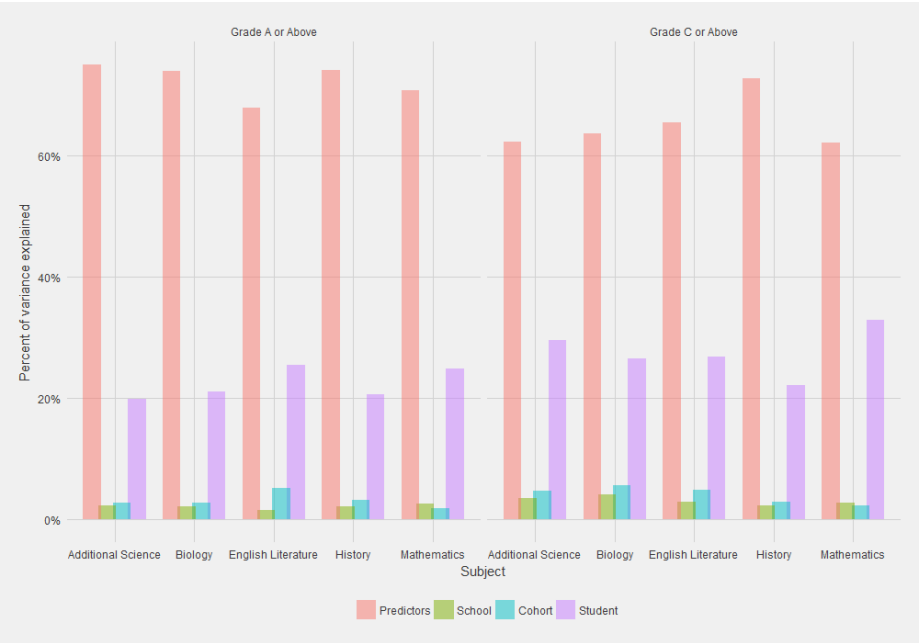

Figure 7. Standardised variance by component for each GCSE subject (full models with fixed effects for the exam year and each student's mean GCSE grade).

Table 13. Standardised variance components for Grade A or Above models with fixed effects for exam year and mean GCSE grade.

		Predictors	School	Cohort	Student
Grade A or Above	Mathematics	70.7%	2.7%	1.8%	24.8%
	English Literature	67.8%	1.5%	5.2%	25.5%
	Additional Science	74.9%	2.4%	2.8%	19.9%
	History	74.0%	2.2%	3.2%	20.6%
	Biology	73.9%	2.2%	2.8%	21.1%

Table 14. Standardised variance components for Grade C or Above models with fixed effects for exam year and mean GCSE grade.

		Predictors	School	Cohort	Student
Grade C or Above	Mathematics	62.1%	2.8%	2.3%	32.9%
	English Literature	65.4%	3.0%	4.9%	26.8%
	Additional Science	62.2%	3.5%	4.7%	29.6%
	History	72.7%	2.3%	2.9%	22.1%
	Biology	63.6%	4.2%	5.7%	26.5%

How predictable is the extent of volatility?

Finally, we examined the *accuracy* of our predictions. First, we looked at individual students: how does uncertainty in a student's performance relative to grade C vary across ability levels? We looked across all years at the predicted probability of a student achieving a C or higher in Mathematics, and compared this to the student's actual outcome relative to that grade (Figure 8). A good model should be able to cleanly differentiate candidates above and below a C-grade-level. If our model does this, then the predicted probability spread should be bimodal, with those scoring D's and lower assigned probabilities close to zero and those scoring at least a C assigned probabilities close to one. Given that it is more common for students to achieve a grade in the C-or-above range than to achieve grades D and under, it was harder to predict grades below a C than at or above this level. For those that did get at least a C, though, the majority had a predicted probability of this outcome of greater than 0.5. Looking at the figure, the model does appear to correctly categorise the outcomes of most students.

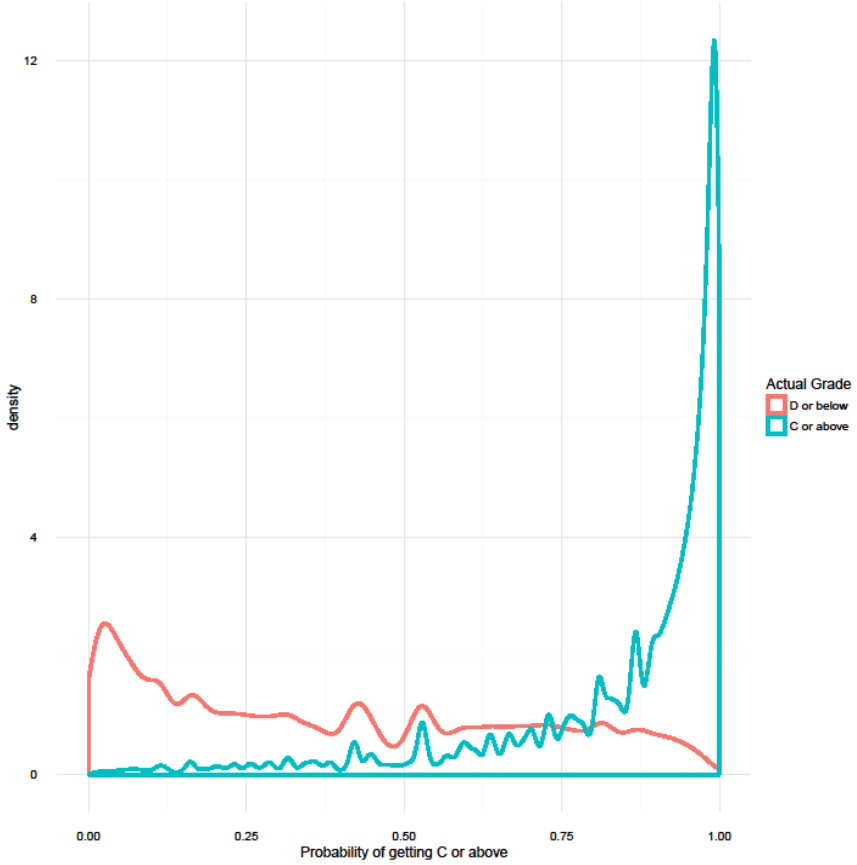

Figure 8. Predicted probability of C-or-above by actual student outcomes for Mathematics.

Although most of the predictions are clustered toward one end of the probability space, there are some (especially those scoring below grade C) that were assigned probabilities near the middle of the scale. Probabilities close to 0.5 indicate uncertainty; this shows that for some candidates concurrent attainment was less helpful in estimating Maths scores than for other candidates. It is likely that these candidates are those whose ability level is concentrated near the C/D grade boundary: if their knowledge is right on the border of the level required to get a C, then a small shift in the test-taking environment, or the chance variation in a student's ability to recall the right information from memory at the right time, could push the student's score from just above the C-cutoff to just below, or from the high end of the D range up to a low-C. A similar size shift for a candidate at a different ability level might be

equally likely, but would not be visible on a dichotomous scale that splits candidates on whether they passed the C-threshold or not.

It is reassuring to see that the model is able to classify students into those that received grades A*-C and those that did not, but our primary goal is to predict variation among school cohorts and not just individuals. To see whether accurate student-level predictions translate into accurate cohort-level estimates, we estimated the expected distribution of changes in the proportion of students achieving grades A*-C for the 2015 cohort compared to the 2014 cohort across schools. This was done using the following method.

Because cohorts are treated as random samples of students from a population, we can estimate characteristics of a cohort using the same methods that we use to estimate characteristics of the distribution of samples of a given size from any larger population. Specifically, we compute the expected proportion to achieve a C or above ($\bar{x}$) as the average of all predicted probabilities within a particular year within a school. We then need to calculate the uncertainty in this prediction – that is, its variance ($\sigma_{\bar{x}}^2$). Specifically, because the cohort can be treated as a sample of binomially distributed random variables that each predict a student's likelihood of a C or above grade, we use standard formulae for the mean and variance of the mean:

Let p_i = the predicted probability of student i getting a grade of C or above.

$$\bar{x} = \frac{\sum p_i}{n} \tag{7}$$

$$\sigma_{\bar{x}}^2 = \frac{\sum p_i(1-p_i)}{n^2} \tag{8}$$

The expected change between two cohorts is then the difference between the random variables representing each cohort's performance in each year. The same method was used to calculate the expected change between 2014 and 2015 cohorts using our other performance measure of interest, achievement of grade A or above. The distribution of these variables is summarised using the mean and variance as calculated above. To calculate the distribution of the difference between random variables we subtract one mean from the other to calculate the expected change and add the variances to calculate the uncertainty in this estimate (Papoulis, 1965). This calculation is performed for each individual school to give an idea of how much we expect their results to change and how uncertain we are in this prediction. We then took the expected values for the change in a school's performance and used the normal approximation[2] to estimate the likelihood that the actual change would exceed various amounts. Finally, we plotted the expected distribution of changes between 2014 and 2015 nationally (given by averaging normal density curves with mean $\bar{x}_{2015} - \bar{x}_{2014}$ and variance $\sigma_{\bar{x}_{2015}}^2 + \sigma_{\bar{x}_{2014}}^2$ across all schools) on top of a histogram of the actual distribution of changes in the percentage of grades C-or-above or A-or-above for each school (Figures 9 and 10).

This is done in two ways. In each graph in Figures 9 and 10, the red dotted line is based on calculations starting from predicted probabilities based on all the terms used in the model – crucially including the effects of individual schools on students' attainment in different years.

2 Other work by the authors explored using more complex techniques based on simulation. These were found to yield almost identical results so the approximation to the normal distribution was considered acceptable.

As such, it is no surprise to see the extremely close agreement of this prediction to the actual distribution. In contrast, the black line is based on predicted probabilities assuming that the effect of a school on its students' performance is constant every year. In other words, the black line shows the level of volatility we would expect even if the influence of individual schools on attainment never changed. As such, the only causes of fluctuations in schools' results are: changes in the specific test questions, changes in the observable characteristics of students (mean GCSE), and the aggregate effect of uncertainty for individual students on their school's results.

The crucial point from Figure 9 is that the majority of the variation in schools' performance as measured by student achievement of grade C or above between cohorts can be predicted by observable differences in the students that attend a given school in different years and the use of probability theory to incorporate the aggregate influence of uncertainty for individuals—after all, even with extensive data on a student's typical performance, we never know how he or she will perform on a particular test taken on a particular day.

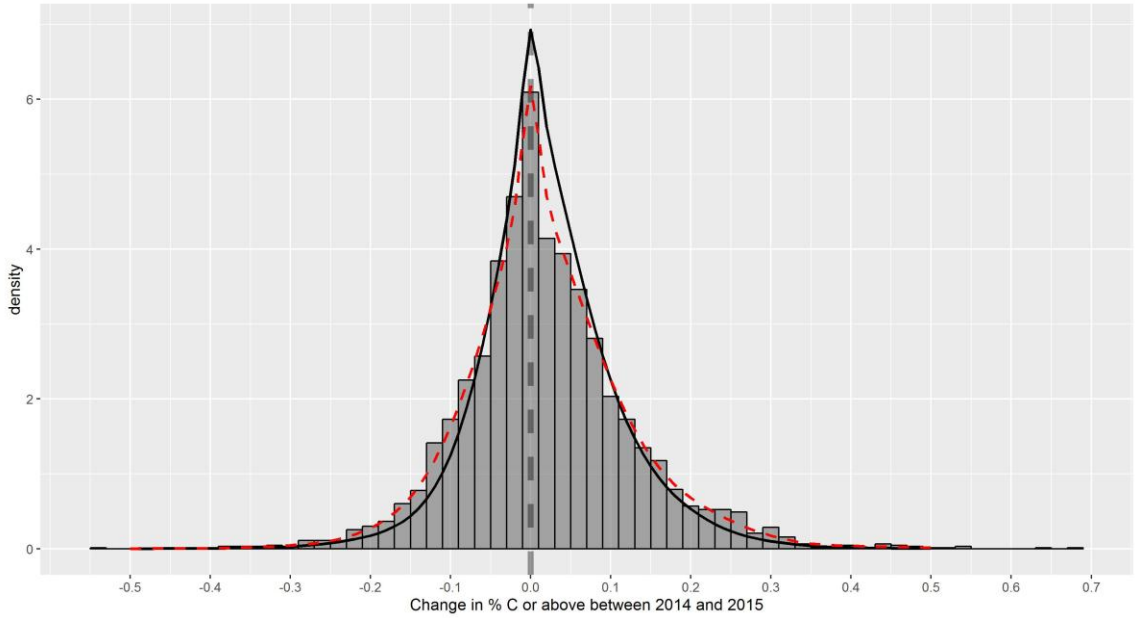

Figure 9. Change in the percentage of C or above grades in Mathematics for 2015 vs. 2014.

Figure 10 shows this same result—the change in grades for the 2015 compared to the 2014 cohort at each school—for the other subjects. In Figure 10, the left panel shows results for grade C or above; the right panel shows results for grade A or above. The results for Mathematics grade C or above (Figure 9) are replicated in Figure 10 for ease of comparison. These figures show that, similar to Mathematics, much of the change in cohort-level results was predicted. Biology stands out as having particularly few schools with large changes, and for being particularly predictable as well. This is likely due to self-selection of GCSEs whereby the only students taking individual science GCSEs (such as Biology) are extremely good at science and likely to score well above the C-grade boundary. The History results for grade C appear similar to Maths; there seem to be slightly more schools with large changes than in Maths, but changes in both subjects are well-predicted by the multi-level models. Additional Science seems to both have large changes in cohort attainment between 2014 and 2015, and the model seems less able to capture (and predict) these changes than in other subjects. This is seen by the black line, representing predicted outcomes for schools, having a larger peak around 0 and a narrower spread than the grey bars, which represent the true results.

Looking at the results for grade A or above, the results are largely similar, with a few noteworthy exceptions. The Mathematics and History results for grade A are similar to those for grade C, with some variation in schools' results between the two years, but most of this predicted by our models. For Additional Science, which had very high and relatively un-predictable volatility at grade C, the results are both less variable and more predictable for grade A. In contrast, English Literature and Biology show much greater variation in different cohorts' attainment of grade A than grade C. For Biology, there is a particularly noticeable increase in the variation in proportion of A-or-above grades than C-or-above grades, which is in line with our earlier speculation that the students entering the Biology GCSE are high-achieving pupils (who are much more likely to have ability levels on the A/B boundary than the C/D boundary). However, the predictions are fairly well-matched with these results, with a good deal of the variation being predicted. English Literature shows a somewhat different pattern, whereby there appear to be more schools with large changes between 2014 and 2015 than our models anticipated. It is possible that this is due to the small number of items on the English Literature question papers; if there are just 4 questions to be answered, then small differences in the questions between years could result in larger-than-expected variation in performance.

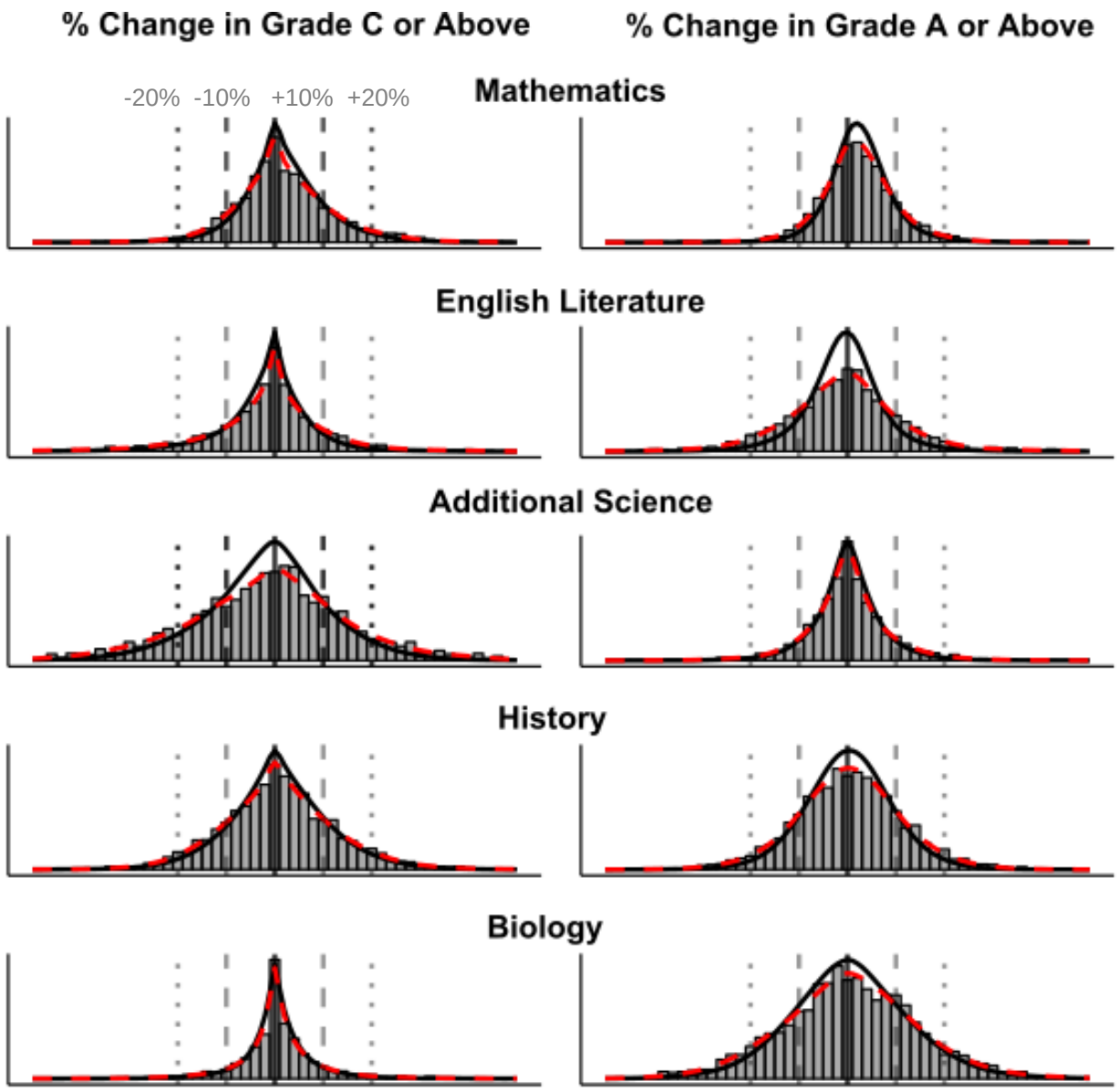

Figure 10. Change in the percentage of C-or-above and A-or-above grades in 2015 vs. 2014. As shown in the top left panel, the dashed lines on each histogram indicate a change of ±10 percentage points in cohort-level achievement; the dotted lines indicate a change of ±20 percentage points.

Another way to look at these results would be to compare the absolute change (ignoring whether the 2015 cohort performed better or worse than the 2014 cohort) and to examine how many schools we would expect to have changes of certain sizes between the two years—after all, volatility describes the amount of change over time and not the direction of change. In Figure 11, the red line represents that same data as the histogram for Mathematics grade C or above, but in terms of absolute change. For example, twenty eight per cent of schools saw a change of at least 10 percentage points in their Mathematics grade C results between 2014 and 2015. The remaining lines explore the extent to which this could have been expected. The green line illustrates the same information as the black line in Figure 9. That is, it shows the percentage of schools we would expect to see changes of at least each given amount if these were only due to differences in observable student characteristics and the effect of student-level uncertainty. Using these calculations, for example, we would actually expect 22 per cent of schools to have experienced a change of at least 10 percentage points in the percentage of the 2014 vs. 2015 cohort achieving grade

C or above in Mathematics--even if the actual influence of schools remained completely constant over time. In other words, out of 884 schools with a change of more than 10 percentage points (in Maths grade C or above), a change of at least this size was expected for around four-fifths of this number. It is hardly surprising the actual variation is slightly greater than expected when assuming schools' influence *never* changes—schools do make changes to staff and teaching approaches, and these changes might affect student performance.

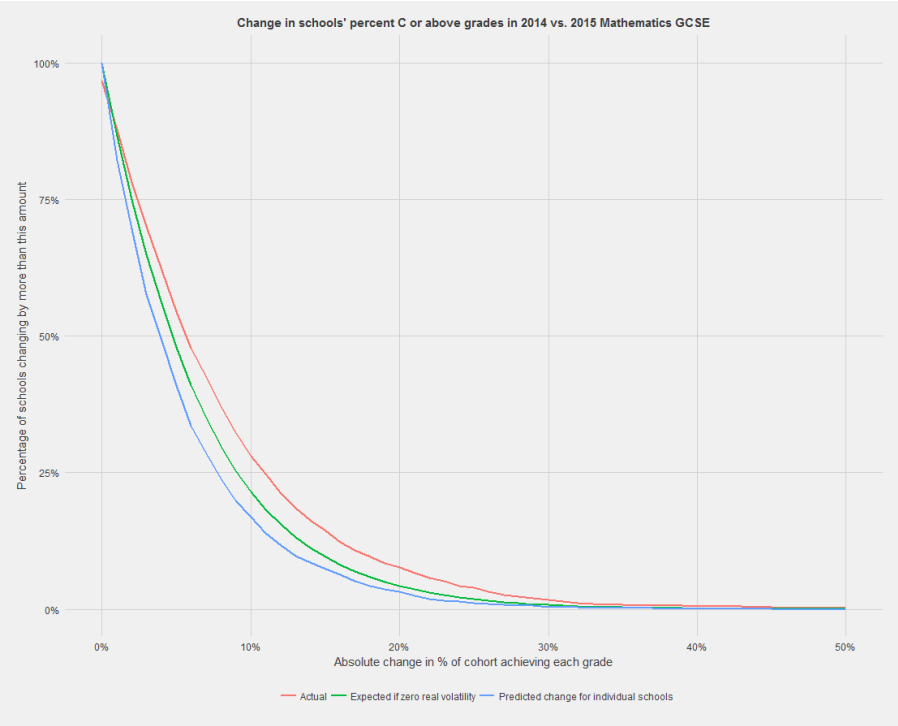

Figure 11. Absolute change in cohorts' percentage of grades C or Above for the 2015 vs. 2014 Mathematics GCSE

Finally, we wished to examine the relative contribution of changes in students' characteristics as opposed to uncertainty in students' outcomes (despite known characteristics) to school-level volatility. This is explored by the blue line in Figure 11. This simply shows the percentage of schools where the difference in the predictions of cohort performance in 2014 and 2015 differ by each given amount. In other words, it shows how much change would occur between the two years if students' performance were always in line with expectations given their mean GCSE grades. The fact that this line is so close to our earlier estimate suggests that the majority of changes in school performance are a direct result of changes in students.

Figure 12 shows the same results for the other subjects for grade C and grade A. The graph shown in Figure 11 is replicated in Figure 12 for ease of comparison. As in Figure 10, the left panel shows results for grade A or above; the right panel shows results for grade C or above. As previously discussed, it is notable that Additional Science had a lot of unexplained volatility in grade C results, whereas English Literature had greater unexplained volatility in the results for grade A. Still, even in these subjects, the predictions were not terrible. For instance, for grade A or above results in English Literature, there were 881 schools with changes in results of more than 10 percentage points between their 2014 and 2015 cohorts. This is equal to 36% of the schools included in the analyses for English Literature that had data available for both years. Based on changes in the 2014 and 2015 students' mean GCSE scores alone for each school, we estimated that 269 schools would experience a

change of at least 10 percentage points in their cohorts' A-or-above English Literature GCSE results. When we account for the error in our measurements, and the uncertainty introduced when schools had many pupils with mean GCSE grades very close to the A/B boundary, our estimate increased to 421 schools. In other words, while we did not predict all of the schools with greater than 10 percentage point changes, we did predict almost half of these. And when we compare the number of schools with unexpectedly large changes in English Literature cohort-level performance, it is just 19% of the total number of schools with entries for this subject in both years. Of course this is not to say that those 19% of schools do not matter, or that their results are irrelevant, just that a good proportion of these changes were not necessarily "true volatility" so much as expected differences in scores as a result of observed changes in pupils.

The findings in Figures 11 and 12 can be summarised as three key points. First, a single measurement of the observable differences in students' ability across cohorts is sufficient to explain most of the differences in school performance from year to year. Second, we should not be concerned that the model did not yield perfect predictions about cohort changes, because it only accounts for a single student characteristic—mean GCSE grade—which may not be sensitive enough to capture all of the relevant student-specific traits that affect performance. Third, and most importantly, the fact that the actual and expected changes are so close despite these limitations of the model is encouraging: it suggests that volatility is much more predictable than people think. In other words, if a lot of volatility is predictable from just a few pieces of information (i.e., the exam year and typical attainment of pupils taking the exam), then schools may be able to use alternative measurements of students' ability to make their own estimates of cohort performance. For example, even though GCSE grades are released at the same time for all qualifications taken in a single session, such that teachers could not use concurrent attainment as measured by GCSEs to predict volatility in advance, it is possible that schools hold sufficient information through informal internal assessments that would serve equally well as a measure of students' ability. Our results suggest that schools may want to use such informal performance evaluations to estimate GCSE performance for a specific cohort if they are not already doing so. This approach may be particularly useful for schools with small cohorts or wide variation in student ability levels from year to year.

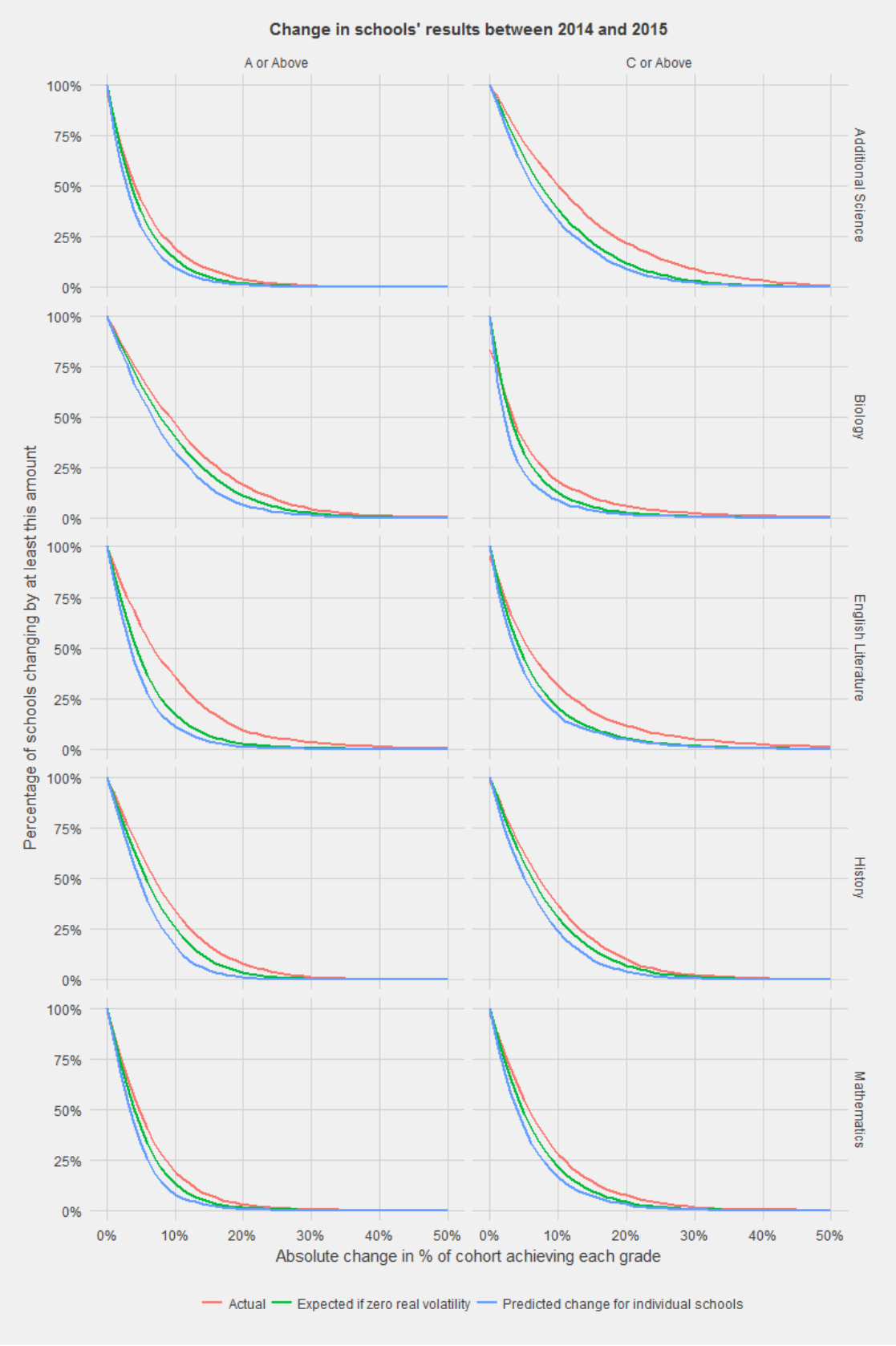

Figure 12. Percentage of schools by minimum change in cohort attainment, 2014 vs. 2015 (e.g., percentage of schools with more than a 10% change in the percentage of candidates achieving a C or above)

General Discussion

A recent report by the UK exams regulator concluded that “any attempt to explain a centre’s variability must consider the individual circumstances surrounding each centre in far more detail” (Ofqual, 2016, p. 21). However, our research shows that knowing just the number of pupils in different cohorts and each pupil’s mean GCSE grade allows us to consistently predict more than 60% of variation in results in a particular subject.

With respect to variation, one can think of cohorts and of schools as little more than groups of students that should, if similar, perform equally well on average. There is no higher-level regularity governing the variation of cohorts from other cohorts at a particular school, and therefore there is no reason to expect consistency in grades despite variation in pupils. This is evident in our results: when we know nothing about individuals other than the year they took an exam and which school they attended in the exam year, 67% of the variation in students’ outcomes in Mathematics (relative to grade C) is unexplainable, as seen in the student-level standardised variance component of Table 8. When we include the exam year and a variable for how students performed on other qualifications taken in the same exam season—a proxy for both ability and environment—we can predict 63% of performance variation just from these. The remaining variation in performance relative to grade C that is correlated with which school one attends is left at 2.7% of the total variance, and given attendance at a particular school, the remaining variation correlated with which cohort the student was part of at that school is equal to 2.3% of the total variance. This means that once student ability is taken into account, the likely performance of a cohort in a particular year at one school is no more or less predictable from that school’s past performance than it is from other similar schools’ past or current performance. And this is not unique to Maths: across five GCSE subjects, just knowing the exam year and mean GCSE grade in other subjects allows us to predict an average of 65% of variation in attainment relative to grade C. This number is even higher when we look at performance relative to grade A: on average, 72% of the variation in student performance relative to grade A is predictable from knowledge of the exam year and mean GCSE grades. This was not driven by a strong model in one subject compensating for weaker model performance in others; in all five subjects, the minimum predicted standardised variance across both the C-or-above and A-or-above models was 62.1%. For Additional Science, the subject with the greatest initial volatility (at 11.2% for Grade A and above outcomes), the predicted variance in A or above attainment is 74.9%, with remaining year-to-year volatility down to 2.8%.

This does not mean that large changes do not, or should not, exist. In contrast, it means that large changes do occur, but we are generally able to predict them based on known changes in the students from one year to the next. When we estimated the change between 2014 and 2015 scores using concurrent attainment, we found that we were able to predict a lot of the variation between years in all five subjects using both C and A as the grade boundary. In some cases we found large differences in cohort performance between the two years; in many of these cases, the large differences were expected.

As discussed in the results sections, we were able to predict a greater proportion of variation in some subjects and at some grade boundaries than others. Specifically, our models underestimated the extent to which scores for 2014 compared to 2015 cohorts differed in their percentages of C-or-above grades for Additional Science. In addition, they underestimated the extent to which 2014 and 2015 cohorts different in their percentages of A-or-above grades in English Literature and Biology. In English Literature, as mentioned earlier, it is possible that the small number of items on the question papers contributed to the greater-than-expected level of variation. Future research could investigate the extent to which the variation in items, and the number of items, on different versions of assessments contributes to volatility in students’ performance over time.

The observed volatility in cohort attainment is not only due to the variation in students within cohorts, but also to the inherent uncertainty in the outcome of any individual pupil on a specific exam. In some years pupils will perform better than expected, whereas, in other years they will perform worse than expected. It would be astonishing if in all schools, all pupils fell firmly in line with expectations every single year. In this way, uncertainty, and hence volatility, in schools' results is a direct consequence of uncertainty for individual students. Furthermore, where we have uncertainty we cannot expect it to immediately balance out for every individual school. This truth has been understood from the beginnings of the study of probability theory. To quote one of the pioneers of this subject:

> But suppose that gain and loss were so fluctuating, as always to be distributed equally, whereby Luck would certainly be annihilated; would it be reasonable in this case to attribute the events of Play to Chance alone? I think, on the contrary, it would be quite otherwise, for then there would be more reason to suspect that some unaccountable Fatality did rule in it… (de Moivre, 1718, p. v)

In other words, volatility happens, and it happens because students are not machines that can precisely access the same information from memory every single time it is required. It happens also because each year, different students enter Year 11 at each school and therefore they are not identical to the group of students that sat for the previous year's GCSEs at that school. Because there is chance involved, and because schools and their students change over time, it is natural that schools would see different results from one year to the next. It would be extremely worrisome if schools' results were too stable, because it would mean one of two things: either that the tests are not sensitive enough to differences in ability to tell us anything meaningful, or worse, that the relationship between a student's performance and his or her grade on an assessment are not as tightly correlated as they should be.

Other Cambridge Assessment research relating to volatility

This report is one of a number of pieces of research into volatility Cambridge Assessment have undertaken over the years. Two recent examples include:

- Bramley & Benton (2015). This earlier report investigated the extent to which volatility in exam results may be caused by the processes of marking and setting grade boundaries. It showed that, even when these two factors are controlled to have minimal impact, schools still see substantial variation in their results between years.

- Benton (2015). This report examined whether the move to on-screen marking was associated with any change in volatility in schools' results. It was found that the move to on-screen marking tended to be associated with increased stability in results with the likely reason being the fact that each school's scripts are distributed across many different markers within an on-screen system.

References

Benton, T. (2015). Examining the impact of moving to on-screen marking on the stability of centres' results. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. http://www.cambridgeassessment.org.uk/Images/268653-examining-the-impact-of-moving-to-on-screen-marking-on-the-stability-of-centres-results.pdf.

Bramley, T. & Benton, T. (2015). *Volatility in exam results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. http://www.cambridgeassessment.org.uk/Images/208853-volatility-in-exam-results.pdf.

de Moivre, A. (1718). *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play*. Pearson: London. Accessed via https://books.google.co.uk/books/about/The_Doctrine_of_Chances.html?id=3EPac6QpbuMC&redir_esc=y.

HMC (2012). *England's 'examinations industry': deterioration and decay. A report from HMC on endemic problems with marking, awarding, re-marks and appeals at GCSE and A level, 2007-12*. http://www.hmc.org.uk/wp-content/uploads/2012/09/HMC-Report-on-English-Exams-9-12-v-13.pdf.

Mason, P. (2016). *Cambridge International Examinations (CIE) IGCSE First Language English (0500) results in GSA & HMC schools*. http://www.hmc.org.uk/wp-content/uploads/2016/04/FINAL-VERSION-HMC-CIE-2015-REPORT.pdf.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, *26*(1), 67-82.

Ofqual. (2016). *What causes variability in school-level GCSE results year-on-year?* Coventry: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/518409/Variability_in_Individual_Schools_and_Colleges_2016.docx_-_FINAL.pdf.

Ofqual (2015). *Variability in GCSE Results for Individual Schools and Colleges: 2012 to 2015*. Coventry: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/454912/2015-08-20-variability-in-gcse-results-for-individual-schools-and-colleges.pdf.

Papoulis, A. (1965). *Probability, random variables, and stochastic processes.* New York: McGraw Hill.

Appendix

To test whether any other students characteristics (beyond mean GCSE) were important in explaining variation between students, cohorts and schools, the final model included all previous effects but added the following variables to account for additional student characteristics:

- Age in years (range=14-18 years, median = 15 years)
- Month of birth (integers 1 through 12, where 1 = September and 12 = August)
- Eligibility for free school meals (binary; 1 indicates eligibility)
- IDACI (Income Deprivation Affecting Children Indices) score (number between 0 and 1, with higher numbers indicating that the student lives in an area with lower socio-economic status)
- KS2 English (performance on Key Stage 2 English test; integer from 1-5[3])
- KS2 Maths (performance on Key Stage 2 Maths test; integer from 1-5, in addition to a binary indicator variable for missing values[4])
- KS2 missing indictor (binary, 1 = KS2 score missing and estimated as global average score[5])
- Gender (female = 1, male = 0)
- Number of non-Maths GCSEs taken (integer; range = 4-21, median = 8)
- Special needs statement (binary, 1 = has a statement for extra support in learning)
- Other special needs (binary, 1 = some form of special accommodation is required, but does not have an official statement of support)

And the following school characteristics:

- Independent school (binary, 1 = pupil attends an independent school)
- Selective school (binary, 1 = pupil attends a non-independent selective school)

The model looked like this:

$$\text{Logit}(y_{ijk}) = \beta_0 + v_k + u_{jk} + \beta_1(year_{2012}) + \beta_2(year_{2013}) + \beta_3(year_{2014}) + \beta_4(year_{2015}) + \beta_5 meangcse_{ijk} + \gamma_{ijk}$$

$$\text{Where } \gamma_{ijk} = \sum_{n=1 \text{ to } 11}\beta_{(n+5)}(x_{nijk}) + \sum_{n=1 \text{ to } 2}\beta_{(n+16)}(z_{nk})$$

Where x_{nijk} denotes the nth student characteristic of the ith students in the jth cohort in the kth year and z_{nk} denotes the nth school characteristic of the kth school.

	Variance Component	Variance	Standardised Variance
Model 3: Many fixed effects	Total	14.80	100%
	Predicted	10.92	73.8%
	School	0.28	1.9%
	Cohort	0.32	2.1%
	Student	3.29	22.2%

3 Pupils with missing information on this variable had the mean value imputed.
4 Pupils with missing information on this variable had the mean value imputed.