Education, *41*(5), 647–670. Available online at: http://dx.doi.org/10.1080/03054985.2015.1090967

Williams, J. (2011). Looking back, looking forward: valuing post-compulsory mathematics education. *Research in Mathematics Education*, *13*(2), 213–221. Available online at: http://dx.doi.org/10.1080/14794802.2011.585831

Williams, J., Hernandez-Martinez, P., & Harris, D. (2010). *Diagnostic testing in mathematics as a policy and practice in the transition to higher education*. Paper presented at the conference of the British Educational Research Association, University of Warwick, Coventry.

Wolf, A. (2002). *Does education matter? Myths about education and economic growth*. London: Penguin.

Wood, L. (2001). The secondary-tertiary interface. In D. Holton (Ed.), *The teaching and learning of mathematics at university level* (pp.87–98). London: Kluwer Academic Publishers.

Question selection and volatility in schools' Mathematics GCSE results

Cara Crawford Mosaic Data Science (The study was completed when the author was based in the Research Division)

Introduction

Exam-setters face a common problem: how to condense a year or more's worth of learning into a couple of hours of test-taking. In the end, they make choices, and some topics receive more coverage in examinations than others. As a result, students may do better on one version of the test than they would do on a hypothetical alternative. In other words, for students, there is always a bit of luck involved.

But what about schools? Certainly individual students have different strengths and weaknesses within a topic area. However, there is less reason to think that the choice of test questions would have a large impact on an entire school's[1] results. Schools have recently expressed concern that test scores vary considerably from year-to-year (Headmasters' and Headmistresses' Conference [HMC], 2012), and previous research has suggested that the questions selected for a test may have small influences on candidates' grades (Benton, 2013a, 2014). If schools are not large enough to be insulated from small question-related effects on their students' grades (because each student has a non-negligible effect on the school's performance), it is possible that question-level influences on students' achievement translate to increased variability in school-level outcomes.

This research estimated the extent to which volatility in schools' scores may be attributable to changes in the selection of questions on question papers by comparing candidates' performance on two halves of the same assessment. Once student grades had been calculated for each half-test, these were aggregated within each school to form school-level outcomes for each half-test (e.g., percentage of students with a grade of C or above). Comparing the variation in schools' outcomes for their students' performance on two parts of a single test should give us some idea of the amount of variation in actual year-to-year results that could be due to changes in test questions.

Data

Data was obtained from 54,167 students who took OCR's GCSE Mathematics B (J567) qualification in the June 2014 exam session. This was chosen because it had the largest entry of any OCR GCSE and also because it consisted of a large number of questions, leaving plenty of scope for looking at variations between them. The assessment was fully linear and consisted of two written question papers. Candidates could either enter for the two Foundation Tier papers (Papers 1 and 2), covering simpler material, or for the two Higher Tier papers (Papers 3 and 4), covering upper-level material. About 56 per cent (30,310 students) were entered for the Foundation Tier (Papers 1 and 2).

All four papers had a maximum possible mark of 100, and qualification grades were based on the sum of the marks achieved on the two completed question papers. This meant that the two papers had an equal impact on final grades for the qualification.

Table 1 shows the breakdown of items (part-questions) and questions across the papers for both tiers (e.g., on Paper 1, 59 item-level marks were combined into 20 question-level marks).

Table 1: Questions and items on OCR's GCSE Mathematics B (J567), June 2014

Foundation Tier	Paper 1	59 items	20 questions
	Paper 2	65 items	23 questions
Higher Tier	Paper 3	48 items	21 questions
	Paper 4	46 items	19 questions

Methods

Overview

This research compared how the same candidates performed on two halves of a single full-length assessment. First, question papers were split by tier, with all Higher Tier questions from Papers 3 and 4 in one set and all Foundation Tier questions from Papers 1 and 2 in a second set. Within each set, questions were split into two subgroups that were as similar as possible. Candidates' marks were calculated for both subgroups of questions completed, and then mapped onto the same mark scale as the complete qualification so that grade boundaries could be set for the subgroups, and subgroup marks could be converted into grades. Each subgroup of grades in one tier was then paired with a subgroup of grades in the other tier, resulting in two combined sets of half-qualification grades. Within each school, the percentage of students achieving grades A*-C and A*-A was calculated for each half-qualification, yielding two pairs of scores for each school. Finally, school-level outcomes on the two half-qualifications were compared.

1. In this article the term 'school' is used for ease of communication instead of the more generic 'centre'. The vast majority of GCSE candidates are in schools.

Splitting questions into half-tests

Questions were split in a way that maximised the covariance between the groups, using the technique developed for calculating Guttman's λ4 reliability coefficient (Guttman, 1945). Initially questions were split into those with odd and even numbers, and then swaps between the two groups that increased the covariance were applied until no further swaps could be found. After that, the same process was repeated using additional starting splits that were assigned for the first 12 questions according to a 12 × 12 Hadamard matrix (simply a matrix that provides lots of different ways of splitting 12 questions into 2 groups so that the splits are as different as possible [Benton, 2013b]). The split yielding the highest covariance between halves (from any starting split) was retained for analysis. Benton (2013b) showed that by first splitting questions in multiple ways (e.g., even-versus-odd numbered questions, first half versus second half) and then swapping individual questions between groups to maximise the covariance between them, an optimal split can be obtained that in theory should ensure a good balance of topic areas and skills between the two halves. By maximising covariance instead of maximising correlations, one should end up with two sets of questions that have similar scales and similar distributions of scores in addition to being highly correlated.

Equating question group marks with full qualification marks

Once questions were split into two groups, equipercentile equating was used to calculate the number of marks on each question group that would correspond to each certificate-level grade on the full qualification. This was done using the *equate* package in R 3.3.1 (R Core Team, 2016) with a single-group design. The single-group design compares two tests taken by a single set of individuals (see Kolen and Brennan, 2004, for a detailed discussion of this method). This method equates scores by calculating the cumulative percentage of candidates achieving different scores on the two mark scales being compared. The score on one scale that is denoted as corresponding to a particular score on the other mark scale is chosen in a way that makes their percentile distributions (the number of candidates achieving at or below each possible score) as close to equal as possible. The intuition behind the method is that if two tests are graded to be equally difficult, then if the same students were to take both tests, the same percentage would achieve grades at or below certain points on them, and the scores that included equivalent proportions of test-takers would represent equivalent levels of performance.

Grade boundaries were then selected for the question groups based on the equated question group mark for each grade boundary on the full qualification. For example, the minimum number of marks needed to achieve a grade A* on the full (Higher Tier) qualification was 166 marks; therefore, the mark equivalent to 166 within each Higher Tier subgroup (rounded to the nearest integer) would be used as the minimum number of marks for a candidate to have achieved a hypothetical grade A* on this subgroup's questions.

Combining grades across tiers

Next, one question group from the Foundation Tier was combined with one question group on the Higher Tier so that grades across all candidates could be easily compared. Figure 1 shows how the question groups within each tier were combined into 'half-qualification' groups, with the Foundation Tier subgroups labelled as Groups W and X, and the Higher Tier subgroups denoted as Groups Y and Z.

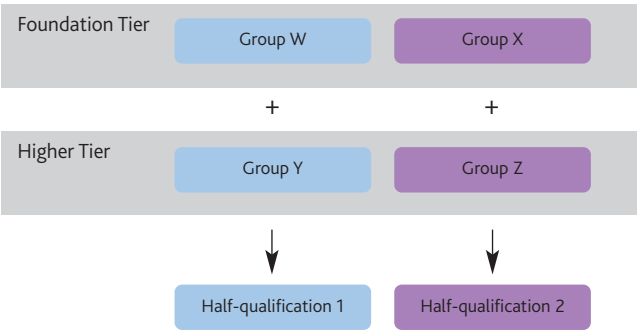

Figure 1: Combinations of question groups into half-qualifications

For each half-qualification, two school-level outcomes were computed: the percentage of students at the schools achieving grades A*-C on the half-qualification (percentage of C or above); and the percentage of students achieving grades A*-A on the half-qualification (percentage of A or above). To prevent individual students at small schools from having a disproportionately large influence on the school-level pattern of results, only schools with at least 10 students entered for the qualification were included.

Comparing schools

Correlations were computed to determine how closely a school's performance on half-qualification 1 predicted its performance on half-qualification 2. A high correlation between school outcomes on the two half-qualifications would suggest a low impact of question selection on volatility in schools' results.

Results

Table 2 examines how the questions and marks from each paper were distributed across the groups. In addition, the rightmost column shows (in bold text) the total number of questions and marks in each question group.

Table 2: Number of questions and marks in each question group by question paper

		Paper 1	*Paper 2*	*Paper 3*	*Paper 4*	*Total*
Group W	Questions	9	9			**18**
	Marks	57	40			**97**
Group X	Questions	11	14			**25**
	Marks	43	60			**103**
Group Y	Questions		8	8		**16**
	Marks		48	50		**98**
Group Z	Questions		13	11		**24**
	Marks		52	50		**102**

Table 2 shows that all question groups contained questions from more than one paper. Looking at the totals in the rightmost column of Table 2, we can see that despite differences in the number of questions in each group, they had similar numbers of marks available. This is most relevant within each tier. For example, it is good to see that even though Group Z had eight more questions than Group Y, this amounted to only four additional marks available from those questions.

Table 3: Equated scores (minimum number of marks needed to achieve each letter grade)

	Foundation Tier	*Group W*	*Group X*	*Higher Tier*	*Group Y*	*Group Z*
Range of marks	*0–200*	*0–97*	*0–103*	*0–200*	*0–98*	*0–102*
A*	-	-	-	166	82	85
A	-	-	-	133	65	68
B	-	-	-	96	46	50
C	110	53	57	59	27	31
D	91	43	48	29	12	16
E	72	33	38	14	6	7
F	54	24	29	-	-	-
G	36	16	20	-	-	-
U	0	0	0	0	0	0

Equated marks

The grade boundaries for each tier of the full qualification and the equated scores on each question group are presented in Table 3. Note that for the grades that can be obtained in both tiers (grades C, D, and E), fewer marks are needed on the Higher Tier papers than the Foundation papers. This is because the Higher Tier papers are harder, so fewer marks are needed to demonstrate the same level of mathematical knowledge.

Figure 2 compares the distribution of marks on each question group to the distribution of marks for the full qualification from which the questions were selected. The plots in the top row of Figure 2 compare the distribution of marks on the full Foundation Tier qualification against the distribution of marks in Group W (top left) and X (top right). The plots in the bottom row of Figure 2 compare the distribution of marks on the full Higher Tier qualification against the distribution of marks in Group Y (bottom left) and Z (bottom right). The main scatterplot in each figure shows the marks obtained on the

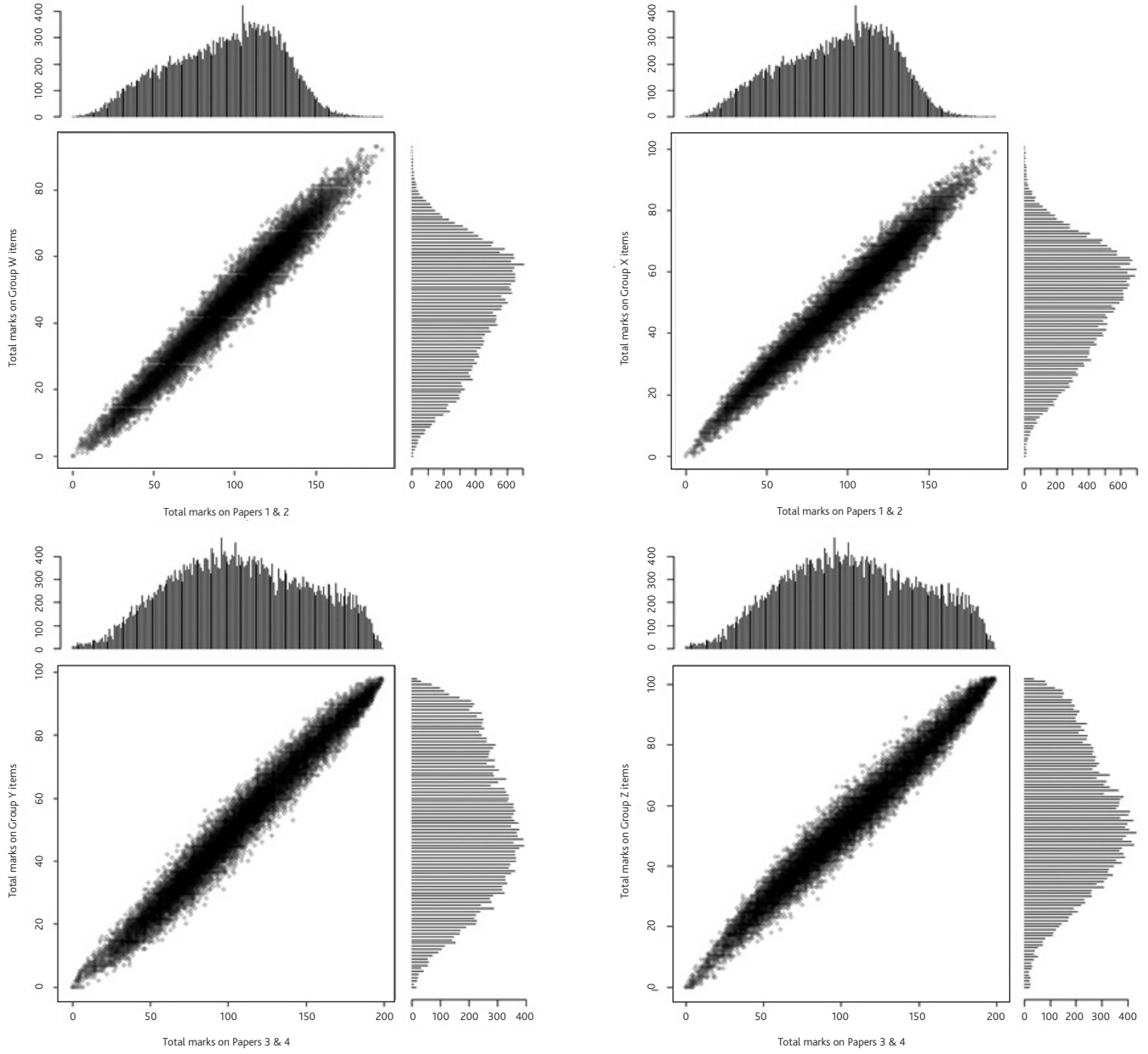

Figure 2: Marks on subgroup questions compared to total marks in each tier: Group W vs. Foundation Tier total marks (top left); Group X vs. Foundation Tier total marks (top right); Group Y vs. Higher Tier total marks (bottom left); Group Z questions vs. Higher Tier total marks (bottom right)

full papers on the x-axis and the marks obtained by the same individual on a question group on the y-axis. The fact that all four scatterplots show a positive linear relationship suggests that the question groups are all representative of the content covered in the papers they were selected from, such that higher performance on the subset of questions in each group is correlated with higher performance on the full set of questions (correlations between question groups and full qualification: Group W = .983; Group X = .982; Group Y = .986; Group Z = .986). The figures also show the distribution of marks obtained on the full qualification for each tier (above the scatterplot) and the distribution of marks on the question groups (to the right of each scatterplot). These histograms allow a comparison of the shapes of the distributions of marks between the full qualifications and the questions in each group. If the distributions have similar shapes, it suggests that a question group contains questions of a similar range of difficulty to the full qualification that its questions were selected from. Of course, these correlations are going to be positively sloped and somewhat similar because the full qualification marks include all of the marks in the question subgroups. Nonetheless, it is reassuring to see the similar patterns as these confirm that the subgroups are representative of the full set of questions.

Student half-qualification grades

A comparison of grades obtained by students on the two question groups within their tier showed that around two-thirds of students had identical grades. Specifically, 66% of Foundation Tier students had the same grade on Group W and Group X questions, and 69% of Higher Tier students had the same grade on Group Y and Group Z questions. These low-sounding levels of classification consistency[2] demonstrate how even highly correlated assessments (see top half of Table 4) can appear unreliable when analysed in this way. Although the level of absolute classification consistency does not sound particularly high, when we look at the number of grades that were either identical or just one letter grade apart (e.g., an A* and an A, or an A and a B), the figures look much better. For the Foundation Tier students, 2.4% had non-consecutive grades on the two question groups (e.g., a C on Group W's questions and an E on Group X's questions). For the Higher Tier students, the likelihood of non-consecutive grades was less than one-tenth of this size, with just 0.2% of students achieving grades on Group Y's questions that were more than one letter apart from their grade on Group Z's questions.

School-level half-qualification results

Next, half-qualifications were aggregated to school level. This resulted in two alternative sets of (half-qualification) GCSE results for each of 487 schools[3].

Correlations were computed to determine how closely a school's performance on half-qualification 1 predicted its performance on half-qualification 2. If these correlations were low, it might suggest that a good deal of school-level volatility in assessment results may be due to differences between the questions used in different exam years. However, the correlation in the percentage of grade C or above grades across schools was 0.98 and the correlation in the percentage of grade A or above grades across schools was about 0.99 (see bottom half of Table 4). In other words, looking at the variation in grade C or above results between schools, 96% of the variation in schools' half-qualification 2 results was explained by variation in half-qualification 1 results[4]. Similarly, 98% of the variation in schools' percentage of grade A or above on half-qualification 2 was explained by variation in half-qualification 1 percentages of grade A or above. This means that despite individual students sometimes receiving different scores for different groups of questions, at the school level question selection appears to have had little effect on outcomes in Mathematics.

Table 4: Correlations between half-qualification results

	Correlation coefficient between half-qualification outcomes
Student-level correlation	
Total marks (Foundation Tier)	0.944
Total marks (Higher Tier)	0.930
Grade (both tiers combined)	0.942
School-level correlations	
% grade C or above (both tiers combined)	0.978
% grade A or above (both tiers combined)	0.989

Scatterplots were created to further explore these relationships, as an overall correlation coefficient can mask variation in certain parts of a dataset. These are shown in Figure 3, with results for schools' percentage of grade A or above plotted on the left, and results for schools' percentage of grade C or above plotted on the right. On the plots, each point represents a school. Each point's position on the x-axis reflects one school's performance on half-qualification 1, and its position on the y-axis reflects the same school's performance on half-qualification 2. In both plots, a blue line shows the predicted percentage (or the most common percentage across all schools) of grade A*-A/A*-C on half-qualification 2 for each percentage of the same grades for half-qualification 1. Points are scaled by school size, with larger schools represented by larger dots on the graphs.

It appears that school size may influence differences in outcomes between the two half-qualifications, as the dots farthest from the blue lines in Figure 3 are very small (i.e., represent schools with very few students). This makes sense because a one grade change (e.g., from a C to a D) for a single student makes a larger difference in the percentage of grade C at a smaller school. To examine this potential cause of differences in half-qualification scores across schools, the absolute value of the difference in the percentage of grade A*-C and grade A*-A for half-qualification 1 versus 2 at each school was plotted against the number of students entered for the exam. These results are shown in Figure 4, with differences in the percentage of grade C or above in the plot on the left, and differences in the percentage of grade A or above in the plot on the right. The points on the graphs are semi-transparent, so a darker point indicates overlapping values for multiple schools (i.e., their points are stacked on top of each other). In both graphs, we can see that as the number of students increases, the difference in schools' achievement on the two halves diminishes. Note that the differences form lines on the graphs because results on half-qualifications differ in whole numbers of students, corresponding to a limited number of possible percentage point variations for each school.

2. The values are not low compared to those typically found in individual units or components of GCSEs and A levels – see Wheadon and Stockford (2012).
3. A total of 505 students were excluded from further analyses because they were at very small schools (entering fewer than 10 students).
4. Calculated by the fact that 0.98 (the correlation) squared is equal to 0.96.

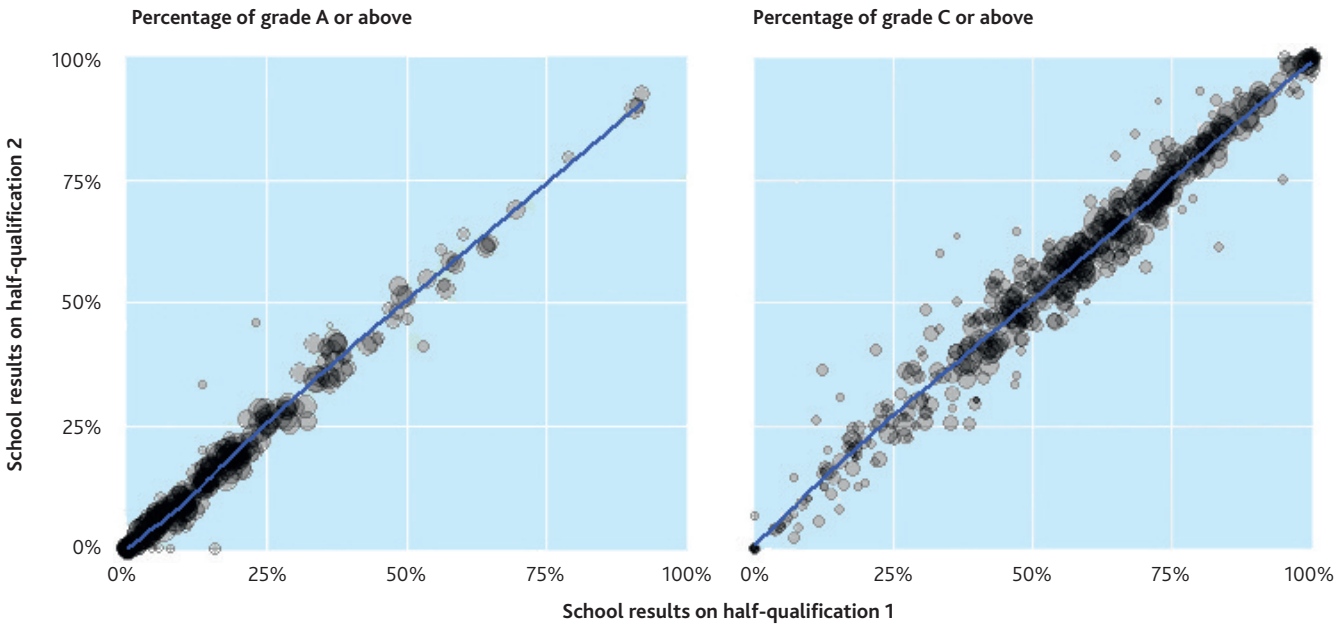

Figure 3: School-level comparison of grade A*-A (left) and grade A*-C (right) on Mathematics half-qualifications (bigger dots indicate larger schools)

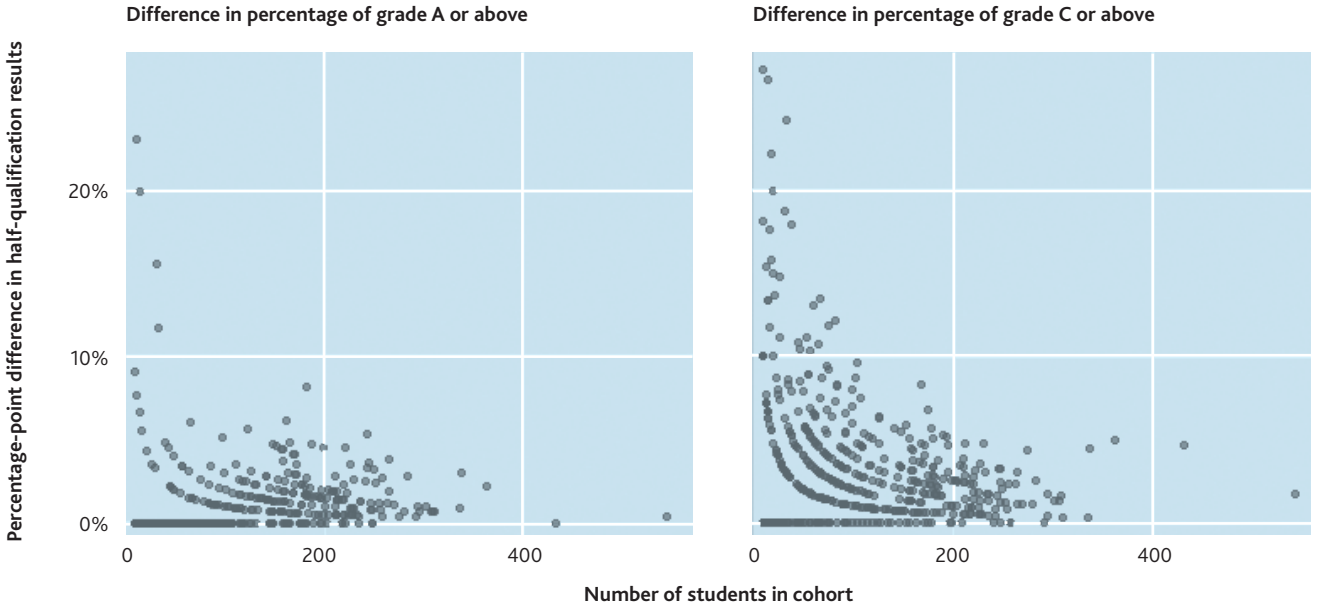

Figure 4: Absolute difference in Mathematics half-qualification results by school size

Together, these two figures indicate that as a school gets larger, volatility due to question selection decreases. To better understand this relationship, Table 5 shows descriptive statistics for the differences in percentage of grade C or above and grade A or above for schools of different sizes.

It appears that school size explains some but not all of the question-selection volatility in Mathematics results; however, even for small schools, the overall differences in performance are quite small. Because the effect of each individual question increases as the length of a test decreases, the values in Table 5 will overestimate the true amount of question-specific volatility that would occur on two full-length Mathematics qualifications. Overall, given that recent reports have considered schools to have relatively stable results when year-to-year variation is less than 10 percentage points (Ofqual, 2016), it seems that the particular selection of Mathematics questions on a given examination does not make a meaningful contribution to volatility in schools' results.

Table 5: Absolute percentage point differences in school grades on half-qualifications

	% C or above		*% A or above*		*Number of schools*
	Mean	*Max*	*Mean*	*Max*	
All schools	3.47	27.27	1.01	23.08	487
At least 50 students	2.79	13.43	1.03	8.24	356
At least 100 students	2.23	9.62	1.36	8.24	228
At least 150 students	2.04	8.33	1.53	8.24	150
At least 200 students	1.76	4.97	1.34	5.35	74

Discussion

This research investigated the potential effect of changes in questions on the same assessments in different years on volatility in schools' results over time. We did this by splitting the assessments in a single year into two shorter 'half-qualifications' and compared schools' outcomes had their students taken one of the half-qualifications instead of the entire assessment. We were interested in the extent to which schools' outcomes changed based on which of the two half-qualifications was used to determine students' grades. Our hypothesis was that if questions are comparable on two versions of an assessment (as they are supposed to be between years and as they were selected to be between halves), then students – and as a result, schools – would get similar results on both halves. Furthermore, we predicted that even if students were likely to have small differences in performance on different questions, if the two sets of questions were sufficiently alike, then these differences would not translate to differences at the school level.

The results were consistent with our predictions. For the Mathematics GCSE, it seems that little of the volatility in schools' results can be explained by differences in the questions on different versions of the tests. When students' grades were computed-based on different subsets of questions from the same question papers, the school-level outcomes were extremely similar; correlations between half-qualification percentages were extremely high, at 0.98 for the percentages of grade C or above, and 0.99 for the percentages of grade A or above.

Despite the overall pattern of results, it is not possible to determine how question selection would affect particular individual schools in particular years, other than adding an additional component of 'measurement error' to any attempts to evaluate schools based on students' test scores. Like other sources of volatility, question-selection variation will affect some schools more than others: the more students with ability levels close to the grade boundaries used to evaluate a school (e.g., borderline A/B-ability students when looking at schools' percentage of grade A and above, and C/D-ability students when looking at schools' percentage of C and above grades), the more uncertainty there will be in how that school will perform on a particular set of questions.

Caution should be used in generalising these results to other subjects. It is possible that question selection would play a larger part in the variability of schools' results in subjects that require candidates to complete fewer total questions on each question paper, or where the assessments cover a smaller range of the total taught material.

Although in general the volatility in results that occurs between exam years – and that is not explainable by differences in student ability – is quite low (Crawford & Benton, 2017), it was possible at the outset of this research that any existing volatility could be due to question selection, whereby questions on one version of an exam emphasise slightly different skills relative to another version of the same exam. Looking at question-level results for Mathematics, it appears that this explanation does not hold; for this subject (and possibly others) we must look elsewhere for explanations of volatility.

References

Benton, T. (2013a). Exploring equivalent forms reliability using a key stage 2 reading test, *Research Papers in Education*, *28*(1), 57–74. Available online at: http://dx.doi.org/10.1080/02671522.2012.754227

Benton, T. (2013b). *An empirical assessment of Guttman's Lambda 4 reliability coefficient*. Paper presented at the 78th Annual Meeting of the Psychometric Society, July 2013. Available online at: http://www.cambridgeassessment.org.uk/Images/141299-an-empirical-assessment-of-guttman-s-lambda-4-reliability-coefficient.pdf

Benton, T. (2014). Calculating the reliability of complex qualifications. *Research Matters: A Cambridge Assessment publication*, *18*, 48–52. Available online at: http://www.cambridgeassessment.org.uk/Images/174492-research-matters-18-summer-2014.pdf

Crawford, C., & Benton, T. (2017). *Volatility happens: Understanding variation in schools' GCSE results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at: http://www.cambridgeassessment.org.uk/Images/372751-volatility-happens-understanding-variation-in-schools-gcse-results.pdf

Guttman, L. (1945). A basis for analysing test-retest reliability. *Psychometrika*, *10*, 255–282. Available online at: https://link.springer.com/article/10.1007%2FBF02288892?LI=true

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*. Available online at: http://www.hmc.org.uk/wp-content/uploads/2012/09/HMC-Report-on-English-Exams-9-12-v-13.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking*. (2nd ed.), New York: Springer.

Ofqual. (2016). *What causes variability in school-level GCSE results year-on-year?* Ofqual/16/5956. Coventry: Ofqual. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/518409/Variability_in_Individual_Schools_and_Colleges_2016.docx_-_FINAL.pdf

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/

Wheadon, C., & Stockford, I. (2012). Classification accuracy and consistency in GCSE and A level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp.107–139). Coventry: Office of Qualifications and Examinations Regulation.