CAMBRIDGE ASSESSMENT

Volatility in exam results

Tom Bramley & Tom Benton

Cambridge Assessment Research Report

23rd April 2015

University of Cambridge
Local Examinations Syndicate

Author contact details:

ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG

Bramley.t@cambridgeassessment.org.uk
Benton.t@cambridgeassessment.org.uk

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

How to cite this publication:

Bramley, T. & Benton, T. (2015). *Volatility in exam results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Introduction

'Volatility' in exam results is a cause for concern in the eyes of many stakeholders in education. For example, the HMC in 2012 listed 'long-standing, year-on-year variations in grades awarded' as 'Failing 1' in a series of perceived failings of the examinations system:

> "What is the problem?
> Schools receive unexplained and very large variations in the percentages of grades given to successive annual cohorts of pupils from the same school in the same GCSE or A level subject. This is despite the subject being taught by stable teams of experienced staff to annual groups of students of largely similar ability…
>
> What specifically has gone wrong and what is the evidence?
> Between 2010 and 2011, variations of over 10% occurred in the award of GCSE A* or A*/A grades for both English and English Literature in at least one in five of those of our schools that enter candidates for GCSEs, but probably more…
>
> Problems were encountered with each of the main exam boards. Variances of 10%+ absolute are a serious concern. Variances of 20% are completely unacceptable and their causes require urgent attention." (HMC, 2012, p8-9).

The exams regulator in England, Ofqual, also saw it necessary to manage the expectations of schools about the possible impact of changes to the examination system in 2014. These included changing the rule about which results 'count' towards school league table performance, the removal of the January examination session, and the requirement for GCSE units to be taken 'linearly' – i.e. all in the same session. The quote from press coverage below is typical:

> "Parents and pupils should expect 'particularly volatile' GCSE and A level results this month as a result of sweeping changes to the exams system, the exams watchdog has said." (The Telegraph, 1st August 2014[1]).

The purpose of this report was to define volatility and investigate the extent to which volatility in exam results might be attributable to two aspects of the examinations system that the exam boards have some influence over or responsibility for: namely i) the quality of marking; and ii) the setting of grade boundaries.

[1] http://www.telegraph.co.uk/education/educationnews/11006140/GCSE-and-A-Level-results-will-be-particularly-volatile-this-year.html

1. To what extent is volatility related to reliability of marking?

To meet the concerns exemplified in the above quotes, it seems that volatility should be defined at the level of the school[2]. However, different schools will have examinees of different abilities and are therefore likely to experience volatility at different parts of the grade distribution. Also, some grade boundaries are more important than others for accountability purposes – for example the C boundary at GCSE. In order for our measure to be independent of these factors we chose in this section to define volatility in terms of the Kolmogorov-Smirnov distance (KSD) between a school's grade distributions in a particular examination in a pair of consecutive years. The KSD is a standard non-parametric statistical technique for comparing two distributions[3]. In this case it is simply the maximum difference across the grade range between two cumulative grade distributions, wherever this maximum occurs.

In order to get a sense of the typical amounts of volatility in exam results, and then to try to estimate the amount of volatility that might be attributable to unreliable marking, we chose to focus on two GCSE subjects, Mathematics and History[4]. We know that the reliability of marking in Maths is extremely high (less than 1% of variance in component scores can be attributed to the markers, see later), whereas the marking in subjects where the exam questions require extended answers or essays is less reliable. Therefore a comparison of volatility in similar schools' results in Maths and History should indicate the extent to which unreliable marking affects volatility.

We used the National Pupil Database (NPD) for the six years 2008 to 2013, which gave five comparisons: 2008-9, 2009-10, 2010-11, 2011-12 and 2012-13. The NPD for each year contains the exam results (all boards) of pupils in England taking their exams at the end of Year 11. We wanted to focus on schools where results might be expected to be stable (i.e. not volatile), so we selected schools with at least 50 examinees in each of the six years, and less than 20% difference in entry[5] for each of the five pairs of years. This produced 146 schools for History and 631 schools for Maths (not surprisingly since within schools more pupils in general do Maths GCSE than History). In order to make our sample of Maths schools as similar as possible to the sample of History schools we selected 146 of the 631 such that the number of examinees per school was matched reasonably closely. Table 1 below shows the average number of examinees per year and average change in entry per year-pair for these 146 schools.

Table 1. History centres and Maths matched centres. Distribution of average entry and average yearly % change in entry.

Source	Variable	N	Mean	Std Dev	Minimum	Maximum
History	av_N	146	99.18	30.20	54.30	197.30
	av_change	146	8.69	2.28	3.38	15.71
Maths	av_N	146	110.03	24.90	55.40	197.40
	av_change	146	5.90	3.06	0.73	15.38

Even after matching, the Maths schools had slightly larger numbers of examinees on average, and less average change in entry per pair of years, leading us to expect (all things being equal) slightly

[2] We use 'school' and the more accurate 'centre' interchangeably (not all examination entries come from schools).
[3] Strictly the distributions should be continuous rather than discrete, but our purpose is not to derive statistical significance tests.
[4] We chose History rather than English to avoid the problems associated with changes in entries between syllabuses titled 'English', 'English literature' and 'English language and literature'.
[5] That is, ((maximum entry size)/(minimum entry size))<1.2.

less volatility in the Maths school than the History schools even before marking reliability is taken into consideration.

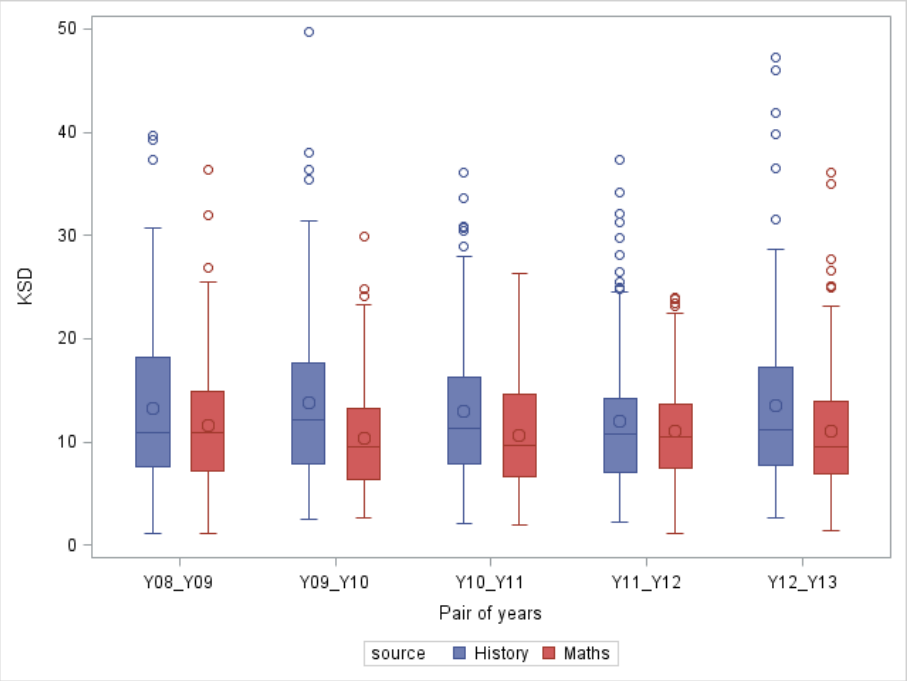


Figure 1. Distribution of KSD by subject and year-pair. (The circle in each box represents the mean, the line in each box represents the median, the box covers the interquartile range, circles beyond the 'whisker' represent outlying schools.)

Figure 1 shows that for each subject, the median KSD was around 10-12 percentage points in each pair of years. Because the KSD is the maximum difference across the grades, we know that the difference at any particular grade boundary (e.g. A or C) is less than or equal to the KSD. Figure 1 shows that even in a reliably marked subject like Maths, there is still considerable fluctuation at school level from one year to the next, even in schools with relatively large and stable entries. The average KSD is slightly higher for History by about 2-3 percentage points, and (perhaps significantly) History schools show more variability in KSD (wider boxes and more outliers).

Table 2. GCSE Maths: matched centres with N>50 and entry change < 20%. Average volatility (KSD) by pair of years.

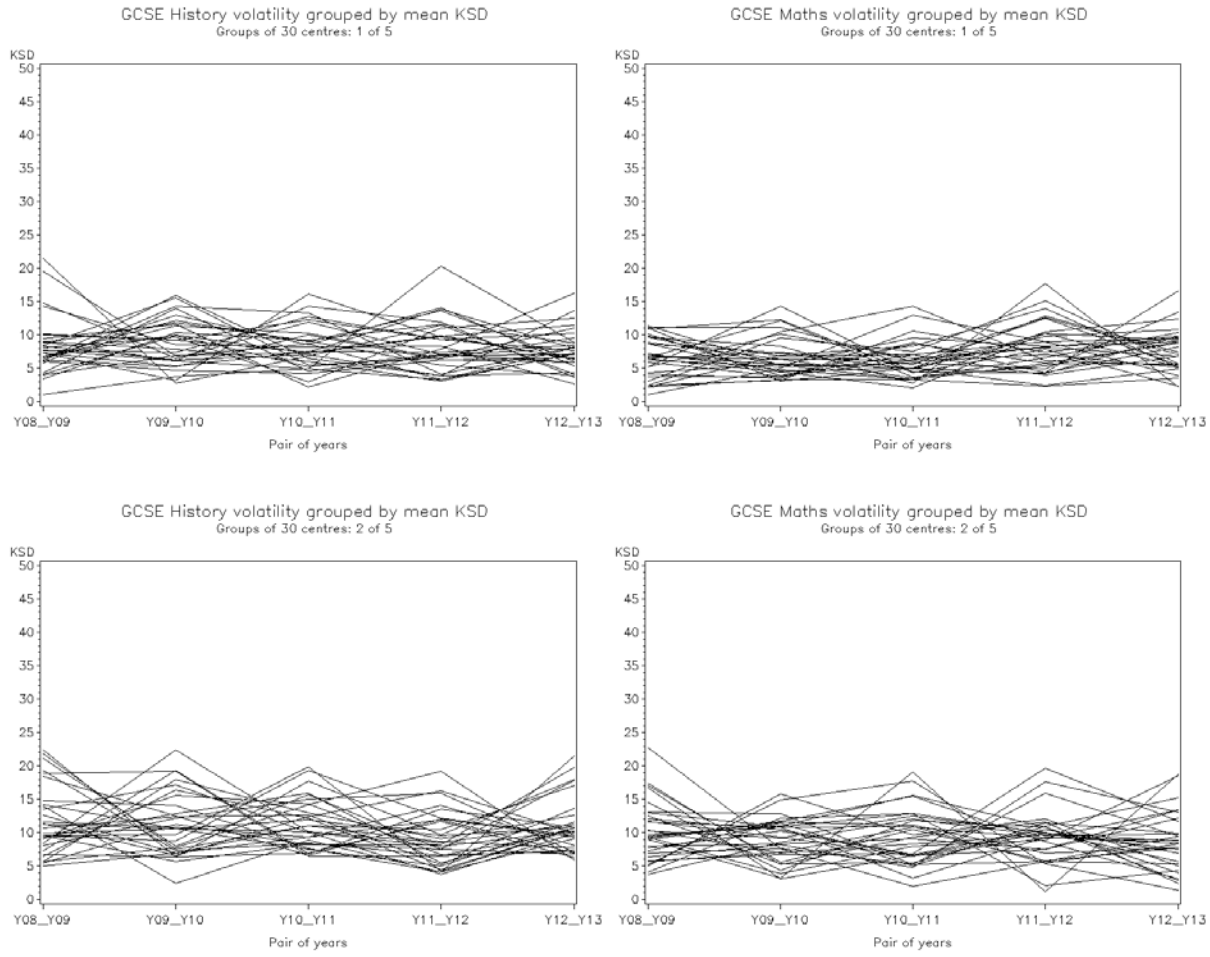
YearComp	N	Mean	Std Dev	Minimum	Maximum
Y08_Y09	146	11.57	6.07	1.07	36.34
Y09_Y10	146	10.39	5.35	2.60	29.89
Y10_Y11	146	10.59	5.03	2.00	26.26
Y11_Y12	146	11.08	4.83	1.18	24.02
Y12_Y13	146	11.01	6.27	1.35	36.11

Table 3. GCSE History: centres with N>50 and entry change < 20%. Average volatility (KSD) by pair of years.

YearComp	N	Mean	Std Dev	Minimum	Maximum
Y08_Y09	146	13.23	7.62	1.09	39.68
Y09_Y10	146	13.74	7.68	2.44	49.74
Y10_Y11	146	12.98	6.89	2.15	36.04
Y11_Y12	146	12.05	6.86	2.18	37.38
Y12_Y13	146	13.57	8.29	2.68	47.22

The graphs in Figure 2 show the KSD values in each pair of years for each school, with schools grouped into blocks of 30 in order of mean KSD. Because KSD is a difference between years, a fluctuating line on the graph does not indicate a school with fluctuating (volatile) results: rather a consistently high horizontal line indicates such a school. A zig-zag line shows a school where a big change was followed by a small change, then a big change etc. The difference between Maths and History does not become particularly pronounced until the most volatile 30 schools are considered (final pair of graphs in Figure 2).

This suggests that marking unreliability has a small effect on volatility on average, but may have a larger effect in some centres, although, of course, differences between the two subjects other than marking may also explain these results.



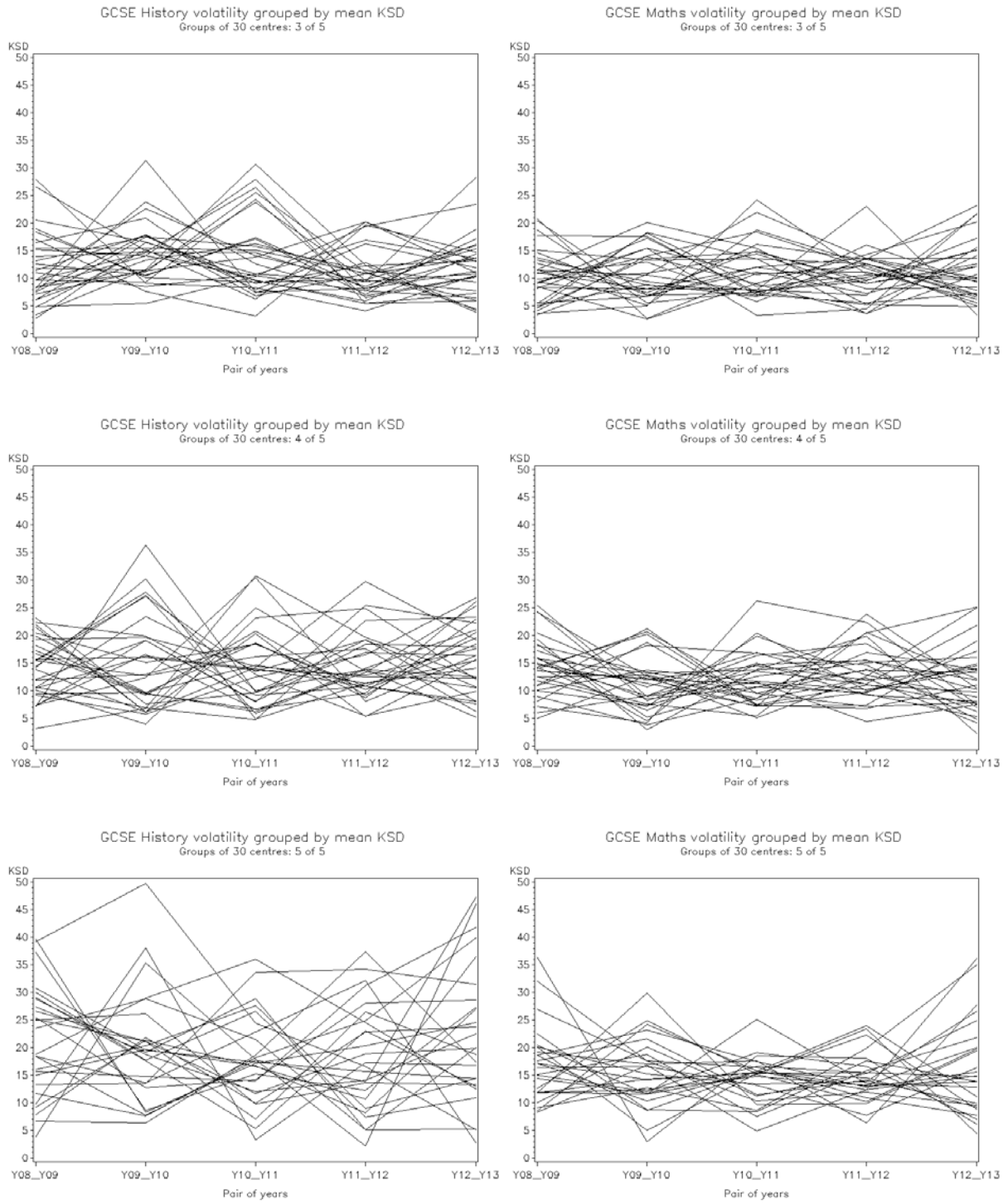


Figure 2: KSD values for each school in each pair of years, grouped by average KSD.

2. To what extent does volatility depend on where the grade boundaries are set?

In this section we will seek to demonstrate that in fact there is a lot of volatility even in subjects where marking is accurate *no matter where we put the grade boundaries*. This implies that the detailed statistical work used to inform grade boundaries should not be blamed for volatility in results. It further implies that most volatility is likely to be due to causes beyond the way papers are marked and grade boundaries are set.

To demonstrate these points we examine changes in schools' results between June 2013 and June 2014 in OCR's largest Mathematics qualification: specification J567 (GCSE Mathematics B). Assessment for this qualification requires candidates to take two 100 mark papers either at the Foundation tier or the Higher tier. Analysis will focus on volatility in results at grade C – a grade that can be achieved by either route.

For each of 2013 and 2014, we restricted analysis to examinees in year 11[6] for whom we had scores recorded in both papers. The numbers of such examinees available for analysis in each year are shown in Table 4. In order to examine volatility, it was necessary to further restrict analysis to centres with large and stable entries in both years. Specifically we restricted analysis to centres with at least 100 year 11 candidates taking J567 in each year where the size of the entry changed by less than 20 per cent between years. Details of candidates in these "benchmark" centres are also included in Table 4.

Table 4: Descriptive information for J567

	June 2013	June 2014
Overall (year 11)		
Number of candidates	29,203	47,895
% achieving grade A or above	13.4%	14.0%
% achieving grade C or above	53.7%	62.7%
Number of centres	435	569
Benchmark centres (year 11)		
Number of candidates	9,511	9,729
% achieving grade A or above	22.3%	22.1%
% achieving grade C or above	66.6%	71.9%
Number of centres	55	55

Before beginning analysis of this data, it was first necessary to confirm that marking was extremely accurate for these components. In order to confirm this, we analysed the marks awarded by each live marker to *seed scripts*. Seed scripts are exam papers that are marked by all examiners as part of the quality control process to allow us to monitor the ongoing accuracy of marking. Data from these seed scripts was analysed using generalizability theory to estimate the percentage of the variance in total scores awarded to scripts that could be attributed to: i) the severity of different markers; ii) erratic marking; and iii) the quality of the script. Ideally all of the variance in the total scores that are awarded should be associated with which seed is being marked and none of it should be associated with which marker is marking. The results of this analysis are shown in Table 5. As can be seen, for every one of the components in J567 at least 99 per cent of the variance in scores is related to which seed is being marked rather than who is marking it. This provides strong evidence that marking is reliable for this qualification.

[6] As defined by their year and month of birth.

Table 5: Marking reliability for components within J567

Code	Component Name	Number of instances of marking analysed	Number of seeds	Number of markers	% of variance in scores attributable to…		
					Marker Bias	Erratic marking	Seed quality
J567_01	Paper 1 (Found.)	1567	19	96	0.00%	0.35%	99.64%
J567_02	Paper 2 (Found.)	1560	19	94	0.01%	0.40%	99.59%
J567_03	Paper 3 (Higher)	1241	18	82	0.00%	1.00%	99.00%
J567_04	Paper 4 (Higher)	1117	16	77	0.01%	0.23%	99.76%

Having established that marking is highly reliable for this unit (and thus cannot be responsible for any volatility) we now move on to examine both how much volatility there is in results and the extent to which this could be reduced if we chose different grade boundaries.

For this analysis we defined volatility in terms of the size of the change between 2013 and 2014 in the percentage of candidates achieving grade C or above in each centre. This is because the C boundary is the 'key' boundary at GCSE for accountability purposes. Also, the boundary-setting procedure requires only 3 boundaries to be 'set' based on the available evidence – the other boundaries follow automatically by application of interpolation rules (see the Code of Practice, Ofqual, 2011).

On average, benchmark centres experienced a change of 7.3 percentage points in their pass rate. According to the HMC (see previous quote), for centres with more than 100 candidates "Variances of 10%+ absolute are a serious concern" and "Variances of 20% are completely unacceptable". With this in mind we should note that 19 of the 55 benchmark centres saw a change of 10 or more percentage points in their pass rate and for three centres the pass rate changed by more than 20 percentage points.

To determine whether these large fluctuations could have been avoided if different grade boundaries had been set in June 2014, we calculated what the average level of volatility would have been for every possible combination of C grade boundaries on the Foundation and Higher tier. The mean volatility for each possible set of grade boundaries between 100 and 130 for the foundation tier and between 40 and 70 for the higher tier is shown in Figure 3. The black dot shows the mean volatility (7.3 percentage points) for the actual grade boundaries (110 on the Foundation tier and 59 on the Higher tier). As can be seen, either raising or lowering both boundaries simultaneously tends to be associated with increased volatility. However, raising the grade boundary on the Foundation tier, whilst lowering it on the higher tier, is associated with lower volatility. The optimal position for the grade boundaries in terms of reducing volatility is shown by the green dot; dramatically changed boundaries to 125 marks and 45 marks on the Foundation and Higher tiers respectively reducing the average volatility in the pass rates from 7.3 percentage points to 5.9 percentage points.

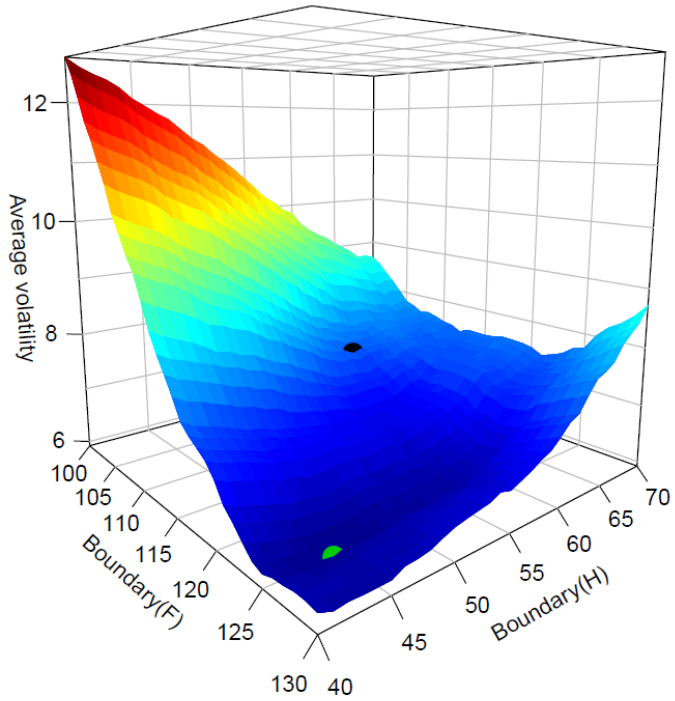


Figure 3: Average volatility for each possible set of grade boundaries

Despite focussing entirely on evidence from the 55 benchmark centres, and despite the large changes made to grade boundaries, the associated reductions in volatility are rather slight. Further visual comparisons of volatility depending on where grade boundaries are positioned are shown in Figures 4 and 5. When examined in this way it can be seen that the change in volatility is not really substantial. The "optimal" definition of grade boundaries results in fewer centres seeing large increases in their pass rates but this is balanced out by a greater number of centres seeing large decreases.

The small extent of the improvement in average volatility using the "optimal" grade boundaries is further reflected in the fact that 12 centres would still see changes of more than 10 percentage points in their pass rate and one would still see a "completely unacceptable" change of more than 20 percentage points[7].

[7] Another three would see changes in their pass rate of between 15 and 20 percentage points.

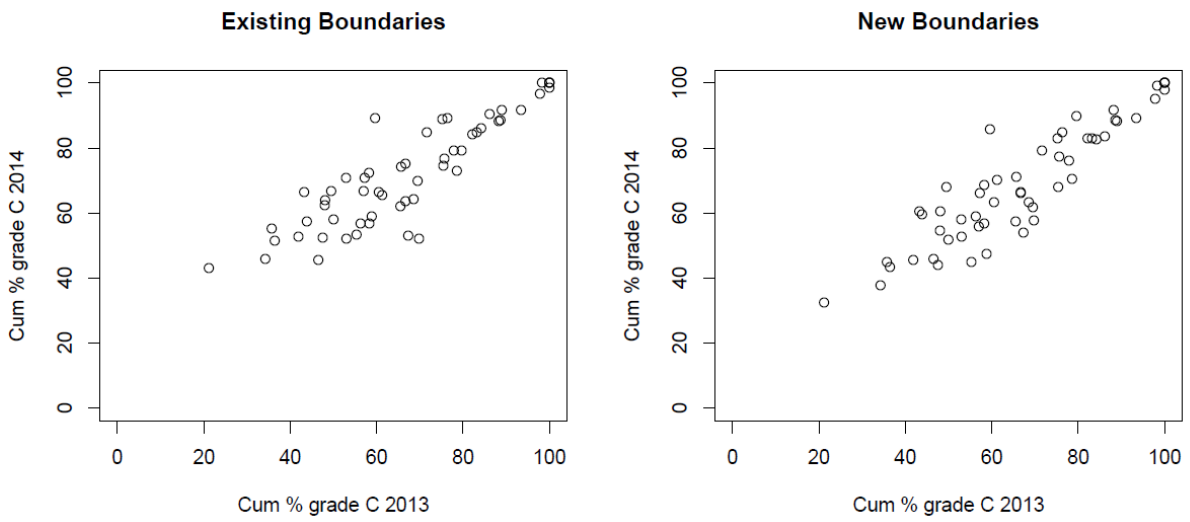


Figure 4: A comparison of grade C pass rates for benchmark centres in June 2013 and 2014.

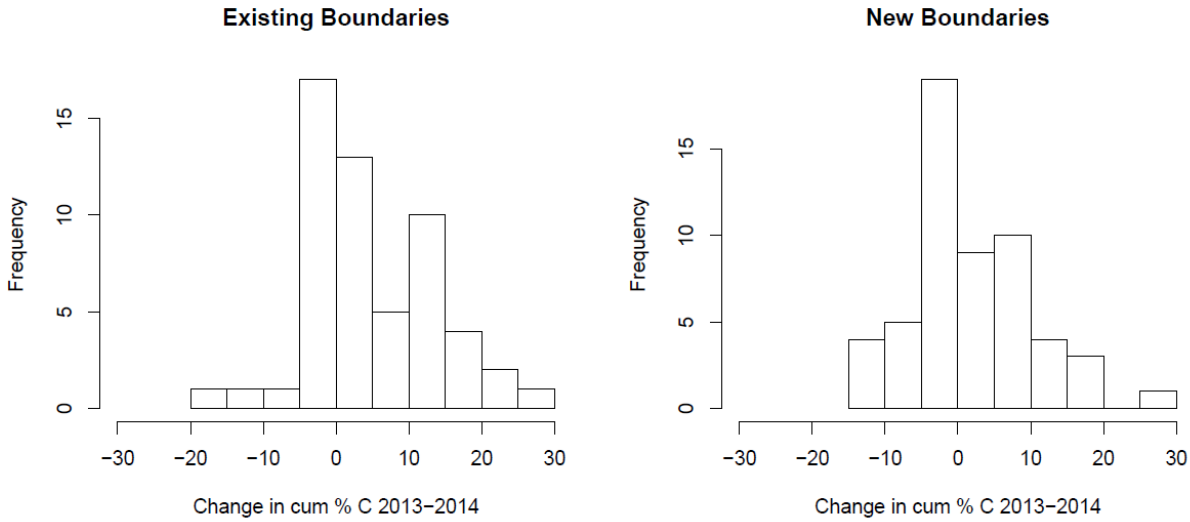


Figure 5: Histograms of changes in pass rates for benchmark centres.

Discussion

It is interesting that, over the period of time considered in this report, there has been very little volatility in the grade distributions of the *overall* cohorts in GCSE Mathematics and History. Figure 6 shows the year-on-year volatility using both definitions of volatility used in this report – namely the KSD, and the change in cumulative percentage at grade C. Ironically, this *lack* of volatility at the cohort level can also attract the criticism of stakeholders, with terms like 'statistical fix' or 'norm-referencing' being used in media coverage (e.g. BBC, 2012; Mansell, 2012).

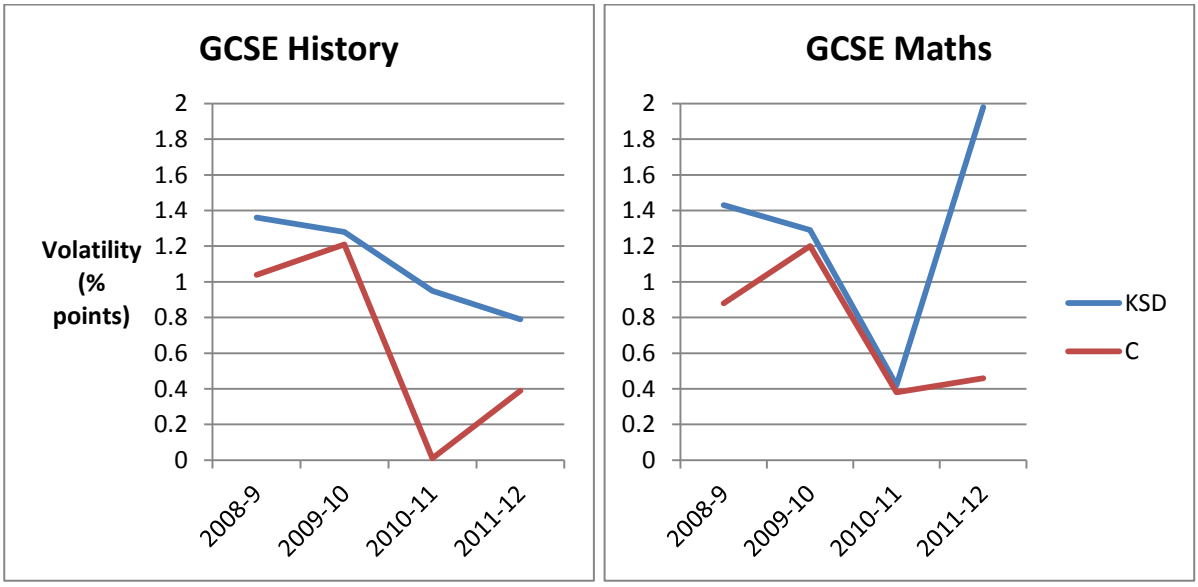


Figure 6: Volatility in GCSE History and Mathematics for the entire cohort, 2008 to 2012[8].

However, at the level of the individual schools, we have shown that there is considerably more volatility in the system. The analysis we have presented suggests that even if marking is accurate, and even if we deliberately choose grade boundaries purely to minimise volatility[9] (to the exclusion of all other sources of evidence) volatility in schools' results would remain. Indeed more than a fifth of schools would still experience levels of volatility that, according to the HMC, should be seen as a "serious concern". Whether or not this level of volatility is concerning remains an open question, and one that cannot be answered without far more detail about the individual circumstances surrounding particular schools. However, what is clear is that volatility alone cannot be taken to imply that either marking or setting of grade boundaries has been performed incorrectly.

[8] Data from the Joint Council for Qualifications (JCQ) for all UK examinees. Not available for 2013 and 2014 at time of writing.
[9] And we already know that this would not provide a strong form of evidence. See Benton (2014).

References

BBC (2012). *GCSE English grades 'statistical fix', High Court told.* Retrieved 10/12/14 from http://www.bbc.co.uk/news/education-20664793

Benton, T. (2014) *Formalising and evaluating the benchmark centres methodology for setting GCSE standards*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online: http://www.cambridgeassessment.org.uk/Images/181083-formalising-and-evaluating-the-benchmark-centres-methodology-for-setting-gcse-standards.pdf.

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.

Mansell, W. (2012). *Ofqual's apparent use of norm referencing at GCSE.* Retrieved 10/12/14 from http://www.naht.org.uk/welcome/news-and-media/key-topics/assessment/ofquals-apparent-use-of-norm-referencing-at-gcse/

Ofqual. (2011). *GCSE, GCE, Principal Learning and Project Code of Practice.* Coventry: Ofqual.