Performance in GCSE Mathematics topics pre- and post-reform

Joanna Williamson and Carmen Vidal Rodeiro

Research Division, Cambridge University Press & Assessment

Between 2015 and 2018, a major programme of reform in England replaced the GCSE qualifications studied by young people aged 14-16. This research analysed the performance in different mathematics topics of approximately 250,000 candidates from the final three years of pre-reform GCSE Mathematics (2014-2016) and the first three years of the post-reform GCSE (2017-2019). A particular contribution was analysing candidate performance on sets of similar items (i.e., families of near-identical items spanning pre- and post-reform GCSE Mathematics papers). The results confirmed that candidates achieved lower overall proportions of marks on post-reform GCSE assessments, but found no statistically significant variation across topics. Analysis of similar items showed that candidates at the grade C / grade 4 borderline demonstrated equal performance on these items pre- and post-GCSE reform.

Keywords: qualification reform; GCSE; assessment; secondary education

Introduction

A recent programme of reform replaced the GCSE (General Certificate of Secondary Education) qualifications that young people in England take at 16. Almost all students study GCSE Mathematics, and the stated aims for the reformed GCSE were ambitious: to ensure mastery of fundamental mathematics by all students, and improve preparation for further mathematical study and careers. There was a policy intention for the new qualification to “be more demanding” and “provide greater challenge” (Gove, 2013).

The motivation for the research reported here was to contribute new evidence on how GCSE reform affected students’ mathematics learning. Important early insights came from qualitative studies and surveys in the first years of the reformed GCSE (outlined below). The current research sought to complement these by exploring student performance in GCSE Mathematics assessments, specifically, analysing performance in different topics pre- and post-reform with the following research questions:

1. How well were pre- and post-reform students able to answer GCSE questions in different mathematics topics?
2. Is there any evidence that post-reform students had higher (or lower) levels of knowledge, skills and understanding in any mathematics topic?

Background

The reformed GCSE in mathematics, like other reformed GCSE subjects, is graded on a new 9 to 1 scale (replacing the previous A* to G scale). It remains a tiered qualification, with candidates entered for assessment at either Foundation (grades 1–5) or Higher tier (grades 4–9). The GCSE 9-1 syllabus (described as “much wider and deeper” (DfE,2013b)) includes new mathematical content, and also the re-assignment of some content from Higher tier to Foundation tier. Changes were also introduced to the relative weightings assigned to topics. Table 1 summarises the weightings as percentages of overall qualification marks, alongside the numbers of new and new-to-Foundation content statements. These give an idea of the extent of change across topics, and show that the largest number of new content statements were in Algebra. Most content in the new topic "Ratio, proportion & change" (hereafter "Ratio") was previously classified as Number or Geometry. For other content, topic classification was generally the same pre- and post-reform, except where smaller content statements were aggregated during GCSE reform into a larger statement, or vice versa.

Topic	**Legacy GCSE**		**GCSE (9-1)**		**New content**	**New to Foundation**
	F	**H**	**F**	**H**		
Number	30-33%	20-22%	25%	15%	6	2
Algebra	20-22%	30-33%	20%	30%	11	8
Ratio	-	-	25%	20%	7	-
Geometry	25-30%	25-30%	15%	20%	2	3
Probability & statistics	18-22%	18-22%	15%	15%	3	1

Table 1: Topic weightings (% overall marks) and new content statements.

Existing evidence on the impact of GCSE reform

One expected impact of GCSE reform was for schools to increase mathematics teaching hours (DfE, 2013b), and available evidence suggests this was implemented (Humphries et al, 2017; Neumann et al., 2016). Some schools reported changes to setting, teacher assignation, and exam focus, and there were some reports of 'strategic' rather than mathematical practices (Neumann et al., 2016, Pearson, 2019).

In their interview study with schools delivering the new GCSE, Humphries et al. (2017) found overall support for the GCSE reform principles, and an expectation that GCSE 9-1 students would gain more and deeper mathematical knowledge than previous students. Multiple studies, however, documented teacher concerns about whether all students would cope with the demands of GCSE 9-1 (Humphries et al., 2017; Neumann et al., 2016; Pearson, 2019). Both GCSE and A level teachers expected the benefits of the reformed GCSE to become more apparent as teachers gained familiarity with it. Some teachers reported a positive impact from the reformed GCSE on problem-solving skills. There were mixed views on algebra skills: some teachers reported an improvement due to the reformed GCSE, and others the opposite (Howard & Khan, 2019; Pearson, 2019). There was agreement that students' algebra fluency was still (on average) insufficient for the transition to A level.

There was little existing quantitative evidence on the impact of GCSE reform. England's mathematics scores from PISA showed marked improvements between 2015 and 2018, driven mostly by increasing scores among lower-attainers (Sizmur et al., 2019). The National Reference Test was available only for post-reform cohorts, but results showed an increase in the percentage of students working at grade 7 and above between 2017 and 2019 (Whetton et al., 2019).

Data and methods

The research analysed data on six cohorts of candidates from one awarding organisation: the final three cohorts (2014-2016) of the legacy linear mathematics GCSE, and the first three cohorts (2017-2019) of the reformed GCSE. The analysis was restricted to candidates aged 16 who took their GCSE exams in the usual May/June summer session, a total of approximately 250,000 students. Most were from comprehensive schools (87.9% pre-reform, 86.1% post-reform). The candidate results data included GCSE grade, overall mark, and marks awarded for each item. Awarding organisation assessment grids recorded the specification content assessed by each item, and this was used to classify items into mathematical topics (according to the classification of content in the GCSE subject criteria, e.g., DfE, 2013a).

To answer the first research question, we produced descriptive statistics on the proportions of marks obtained by topic and grade each year, then carried out regression modelling of item facility values (where item facility is the average score on the item as a proportion of the maximum mark) with topic and year as predictors. To answer the second research question, we compared student performances on so-called 'similar items' from pre- and post-reform GCSE papers. An experienced Principal Examiner (PE) reviewed all Foundation and Higher tier papers from 2014 to 2019, and identified sets of items that were either identical, or "similar enough … for it to be reasonable to expect performance on them to be identical" (Bramley & Wilson, 2016). The item sets identified by the PE were reviewed by two researchers with mathematics backgrounds. We first compared item facility values by grade, for items within each similar item set. We then investigated whether the standard of a grade C/grade 4 borderline candidate had changed with GCSE reform. To do that, we analysed all item responses for candidates taking the same tier in a given year using partial credit models (Masters, 1982), then equated the Foundation and Higher tier models within years to put all candidates and items from the same year onto a common scale. Next, we found the ability estimate (in logits) for a candidate at the grade C/grade 4 boundary, and generated their expected score on each item. Finally, we converted the expected scores into expected facilities, and compared the values for pre- and post-reform items using multilevel regression models (outlined in the results section).

Results

The initial descriptive statistics showed that across all mathematics topics, GCSE 9-1 candidates tended to achieve lower proportions of marks correct than candidates in pre-reform GCSEs. Similarly, considering mean item facility by topic and year showed lower item facilities after GCSE reform (Figure 1). To investigate this more precisely, linear regression models of item facility were estimated for each tier, with topic, year and their interaction as predictors[1]. These models showed statistically significant overall effects for topic and year, but not their interactions (Table 2).

For each topic and overall, we tested the difference between the average "effect" of pre-reform and post-reform years (Table 3). The results showed a statistically significant decrease in item facility overall, for Algebra, and in Foundation tier Geometry. However, since there were no statistically significant interactions between topic and year, *relative* topic performance may have changed no more than expected given the numbers of items involved. That is, it could be by chance that post-reform candidates found Algebra items relatively more difficult.

The lower overall facility values in post-reform GCSE assessments were reflected in lower grade boundaries. As context, Figure 2 plots the published GCSE grade boundaries from the main awarding organizations (AQA, Edexcel and OCR).

[1] The preferred structure was multilevel with a fixed effect for reform status (binary), and a random effect for year, to account for the clustering of items within years. However, it was not possible to satisfactorily estimate both effects with the available data. Models with reform status as a fixed effect (ignoring year) were also estimated, for comparison, and produced the same conclusions.

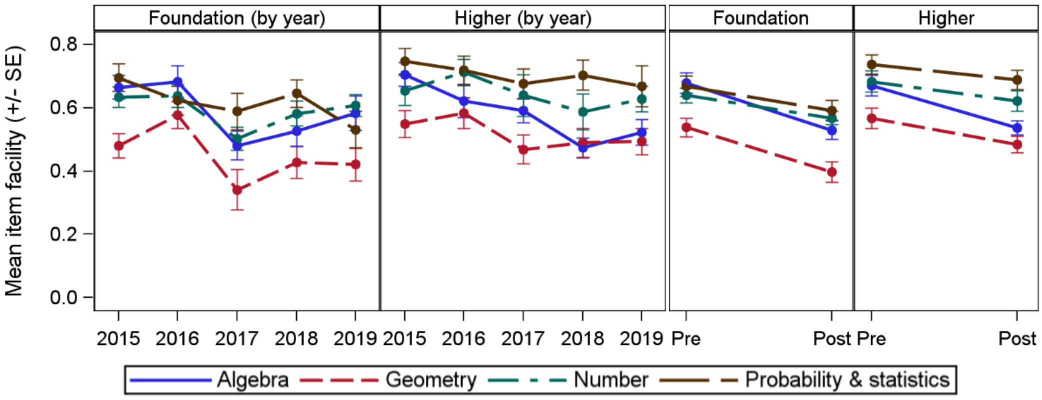

Figure 1: Mean item facility (with standard errors) by topic and year (left), and reform status (right).

Effect	Foundation tier model				Higher tier model			
	Num DF	Den DF	F Value	Pr > F	Num DF	Den DF	F Value	Pr > F
Topic	3	546	11.86	<.0001	3	411	14.07	<.0001
Year	4	546	7.58	<.0001	4	411	3.97	0.0036
Topic*Year	12	546	0.97	0.4734	12	411	0.84	0.6083

Table 2: Tests of fixed effects, item facility models (Foundation tier N=566 items; Higher tier N=431). ‘Num DF’ = numerator degrees of freedom and ‘Den DF’ = denominator degrees of freedom.

Topic	Foundation tier model					Higher tier model				
	Est.	SE	DF	t Value	Pr > |t|	Est.	SE	DF	t Value	Pr > |t|
Algebra	-0.14	0.04	546	-3.49	0.0005	-0.13	0.04	411	-3.63	0.0003
Geometry	-0.13	0.04	546	-3.02	0.0026	-0.08	0.04	411	-1.90	0.0578
Probability & statistics	-0.07	0.04	546	-1.61	0.1073	-0.05	0.04	411	-1.15	0.2521
Number	-0.07	0.04	546	-1.98	0.0488	-0.07	0.05	411	-1.39	0.1647
Overall	-0.10	0.02	546	-5.06	<.0001	-0.08	0.02	411	-3.88	0.0001

Table 3: Estimated reform effect (post-reform– pre-reform years) by topic.

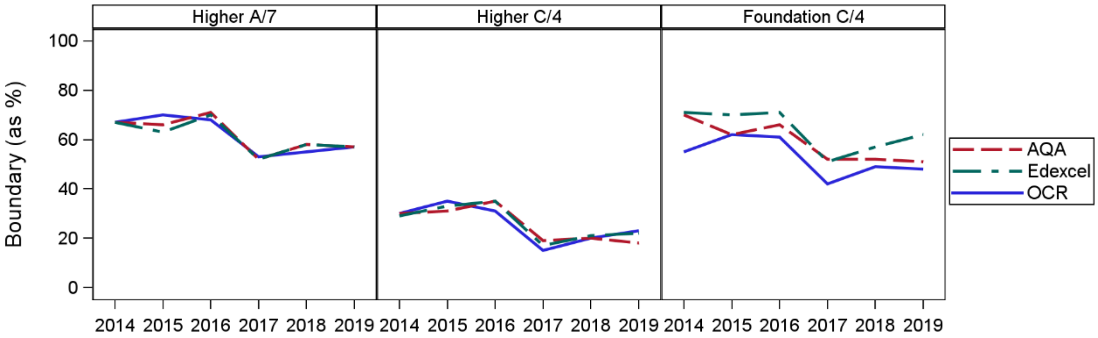

Figure 2: GCSE Mathematics grade boundaries pre- and post-reform.

Similar items analysis

Similar items were first compared by calculating facility values for candidates at each grade. These were plotted as item characteristic curves, with numerical grade equivalent[2] on the x axis to permit legacy and GCSE 9-1 items to be shown on the same scale. Figure 3 shows a typical example. In most item sets, the curves of different items were close or overlapping, indicating similar levels of performance. For one pair of

[2] A* = 8.5, A = 7, B = 5.5, C = 4, D = 3, E = 2, F = 1.5, G = 1 (DfE, 2016, p. 3).

items (identically structured percentage calculations, but using different numbers), the facility curves were very different, indicating that the judgement of item 'similarity' was not correct; this item pair was excluded from the remaining analyses.

The similar items were next analysed using the method previously described. The purpose of this analysis was to investigate whether GCSE reform resulted in 'mid-level' candidates (with an ability estimate corresponding to the grade C/4 boundary) finding items in particular topics easier or more difficult than pre-reform. The multilevel regression model used was as follows: $y_{ij} = \beta_0 + \beta_1 X1_{ij} + \cdots + \beta_k Xk_{ij} + u_j + e_{ij}$, where y_{ij} is the expected facility of item i in similar item set j; $X1$ to Xk are independent variables capturing item topic, GCSE reform status (0 = pre-reform item, 1 = post-reform item) and the interaction between topic and reform status; β_1 to β_k are the regression coefficients; u_j is a random variable at similar item set level (to account for the nesting of items within item sets), and e_{ij} is a random variable at item level. The regression model showed no statistically significant effect on estimated facility from reform status or topic (Table 4). The model was used to create predicted pre- and post-reform facility values (i.e., least squares means), both within topics and overall, and then significance tests were carried out on the pre-post reform differences (Table 5). These results indicated that the level of performance from mid-ability candidates on post-reform items did not differ from performance on similar pre-reform items.

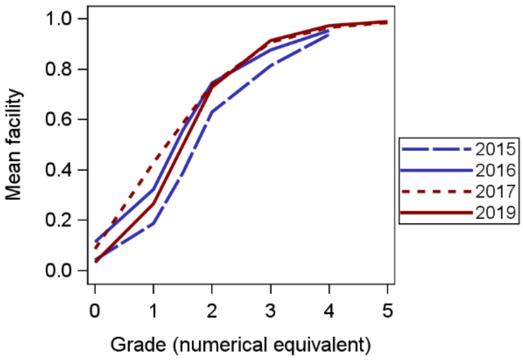

Figure 3: Item facility by grade, for Foundation tier items in a similar item set.

Effect	Num DF	Den DF	F Value	Pr > F
Reform	1	68	3.14	0.0811
Topic	4	68	0.94	0.4449
Reform * Topic	4	68	1.14	0.3448

Table 4: Tests of fixed effects, similar items model (N=117).

Topic	Estimate	Std Err	DF	t Value	Pr > |t|
Geometry	0.04	0.04	68	1.15	0.2528
Number	0.04	0.03	68	1.25	0.2162
Number/Ratio	0.09	0.05	68	1.77	0.0806
Probability & statistics	-0.03	0.04	68	-0.63	0.5292
Algebra	0.00	0.02	68	0.19	0.8486
Overall	0.03	0.02	68	1.77	0.0811

Table 5: Estimated reform effect (post-reform– pre-reform years) by topic.

Conclusions

The research confirmed that, on average, post-reform GCSE Mathematics candidates achieved lower proportions of marks on their assessments than pre-reform candidates, and were likely to have experienced their assessments as more challenging. The results highlighted some variations by topic, including a larger decrease in the proportions of marks achieved in Algebra than in other topics. Combined with changes (by design) to topic content and topic weightings, these variations may have influenced candidates' experiences of different topics, and teachers' early perceptions of the reformed GCSE. However, the analyses indicated that the pre- to post-reform change in Algebra performance was not statistically different from the change in other topics, indicating that the observed differences may simply have been due to the usual variation in difficulty between different items rather than a fundamental feature of GCSE reform. In particular, it is possible that such differences will not persist in future years.

The similar item sets did not include sufficient item overlap to equate the entire grading scale from each year, but we were able to use performances on similar items to make a link between standards from year to year. Reassuringly, our analyses indicated that pre- and post-reform candidates at the grade C / grade 4 boundary gave performances of equal standard on similar items.

The above findings relate only to the first few years of GCSE 9-1 performances, and for this reason may not reflect the impact of the GCSE reform in the longer term. In particular, teachers' ability to prepare candidates for GCSE 9-1 assessments would be expected to improve with time and practice, and consequently we might expect to see improved performances in later cohorts.

References

Bramley, T., & Wilson, F. (2016). *Maintaining test standards by expert judgement of item difficulty*. Research Matters, 21, 48-54.

DfE (2013a). *Mathematics: GCSE subject content and assessment objectives.* (DFE-00233-2013). UK Government, Department for Education.

DfE (2013b). *Reformed GCSE subject content consultation: Government response*. UK Government, Department for Education.

DfE. (2016). *Progress 8: How Progress 8 and Attainment 8 measures are calculated* (DFE-00252-2016). UK Government, Department for Education.

Gove, M. (2013). *Reformed GCSEs in English and mathematics*. [Written statement to Parliament]. UK Government, Department for Education.

Howard, E. & Khan, A. (2019). *GCSE reform in schools: The impact of GCSE reforms on students' preparedness for A level maths and English literature.* (Ofqual/19/6556). Ofqual.

Humphries, S., Cotton, W., Khan, A., & Taylor, R. (2017). *GCSE mathematics: understanding schools' approaches to tiering.* (Ofqual/17/6155). Ofqual.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/bf02296272

Neumann, E., Towers, E., Gewirtz, S. J., & Maguire, M. (2016). *A Curriculum for All?* National Union of Teachers.

Pearson (2019). *GCSE Mathematics Qualification – UK. Regulated qualification efficacy report.* Pearson Education.

Sizmur, J., Ager, R., Bradshaw, J., Classick, R., Galvis, M., Packer, J., Thomas, D., & Wheater, R. (2019). *Achievement of 15 year-olds in England: PISA 2018 results.* (DFE-RR961). UK Government, Department for Education.

Whetton, C., Hopkins, A., & Benson, L. (2019). *National Reference Test Results Digest 2019*. National Foundation for Educational Research.