An investigation into the 'Sawtooth Effect' in GCSE and AS / A level assessments

September 2016

Ofqual/16/6098

Contents

Tables

Figures

Glossary of terms

Assessment / Test – The exam, coursework, or controlled assessment used to assess students’ knowledge, skills, and understanding of a certain topic. Although the specific questions / tasks may change in each assessment series, the specification that outlines the content to be assessed, and the style of assessments to be used, will remain largely the same in each series between periods of reform.

Percentage outcomes – The percentage of a cohort exceeding the standard required for a certain grade.

Performance – The quality of work produced by a student for an assessment.

Test-specific performance – Performance in a particular assessment that is not necessarily representative of a student’s overall mastery of the subject (ie a level of performance that may not be replicated in other assessments).

Post-reform performance gains – Increases in test-specific performance caused by adjustments being made over time (by students and / or teachers) in response to the introduction of a new assessment.

Teaching to the test / measurement driven instruction – Teaching activities targeted towards the knowledge, skills, and understanding that is assessed, rather than the knowledge, skills, and understanding encompassed by the content of the subject overall.

1 Executive summary

The 'Sawtooth Effect' is a pattern of change caused by assessment reform. Specifically, performance on high stakes assessments is often adversely affected when that assessment undergoes reform, followed by improving performance over time as students and teachers gain familiarity with the new test. This pattern reflects changes in test-specific performance over time, whilst not necessarily reflecting changes in a cohort’s overall mastery of the subject (ie changes in performance that may not be replicated in other assessments or in employment).

Although some evidence for this effect can be found within the US literature, evidence is lacking for the UK, as are estimates of the duration and size of this effect. Given that secondary school assessments (GCSEs and AS / A levels) are currently undergoing reform in the UK, more evidence is needed to enhance our ability to predict how students' performance may be affected in the coming years. The purpose of this research was to gather such evidence.

In Study 1, changes in grade boundaries for over 1,100 AS / A level units and 450 GCSE units since the previous set of reforms (2010 for AS / A levels and 2011 for GCSEs) were used as a proxy measure for performance change. When averages were plotted over time, a general trend was observed whereby grade boundaries increased relatively rapidly over the first 3 years of the new assessments, and then changed less rapidly for the remainder of the specification lifespan (with the exception of AS / A level E-grades, which seemed to increase at a steady rate throughout the lifespan). When estimates of outcome change were calculated (using simulated outcome distributions), the size of these changes were relatively small, with estimated average outcomes changing by 2% each year for the first 3 years, and then by 0.5% per year thereafter.

A limitation of Study 1 was a reliance on the inference that changes in grade boundaries over time reflected underlying changes in cohort performance. A second study was therefore conducted to assess the validity of this claim. In Study 2, a comparative judgement methodology was employed, whereby examiners from 4 subject areas were asked to rank order packs of 3 scripts in terms of the performance demonstrated by the candidate. A triples comparison design was used to allow us to compare performance before, after, and at the point at which test familiarity appears to have been reached for each subject. Rasch analyses were used to produce estimates of quality for each script, based on these judgements. When plotted, most of the patterns of these estimates of perceived quality broadly matched the patterns of grade boundary change, thus supporting the idea that underlying performance had changed in the same manner suggested by the results of Study 1.

In conclusion, the results of this research suggest that it seems to take roughly 3 years for students and teachers to become familiar with the nature and requirements of new assessments, meaning that we can have greater confidence that any improvements in performance after this time were due to meaningful gains in that subject area, rather than just test familiarity. Comparisons across cohorts in the early years of a new assessment should therefore be made carefully, to avoid drawing unfair conclusions about a cohort’s performance simply because it was one of the first groups of students to take the test. These findings offer a novel contribution to our understanding of how quickly, and by how much, students and teachers are able to respond to education assessment reforms, and can be used to better predict changes in student performance following any future reforms. The report concludes with a discussion of limitations and suggestions for future research.

2 Introduction

Due to the high-stakes nature of general qualifications (eg GCSEs and AS / A levels), education policy makers continually strive to improve the validity of assessments through a series of revisions and reforms. A pattern has been identified within the US literature whereby outcomes for high-stakes assessments dip suddenly when those assessments undergo reform, followed by a period of improvements. One likely explanation for the ‘dip’ is student / teacher unfamiliarity with the particular style and focus of the new assessment (eg Koretz, 2005). As students and teachers become more familiar with the new style, outcomes begin to rise once more. Then as students and teachers begin to reach maximum familiarity, the rate of change begins to lessen. This pattern, illustrated in Figure 1, has become known as the “Sawtooth Effect”.

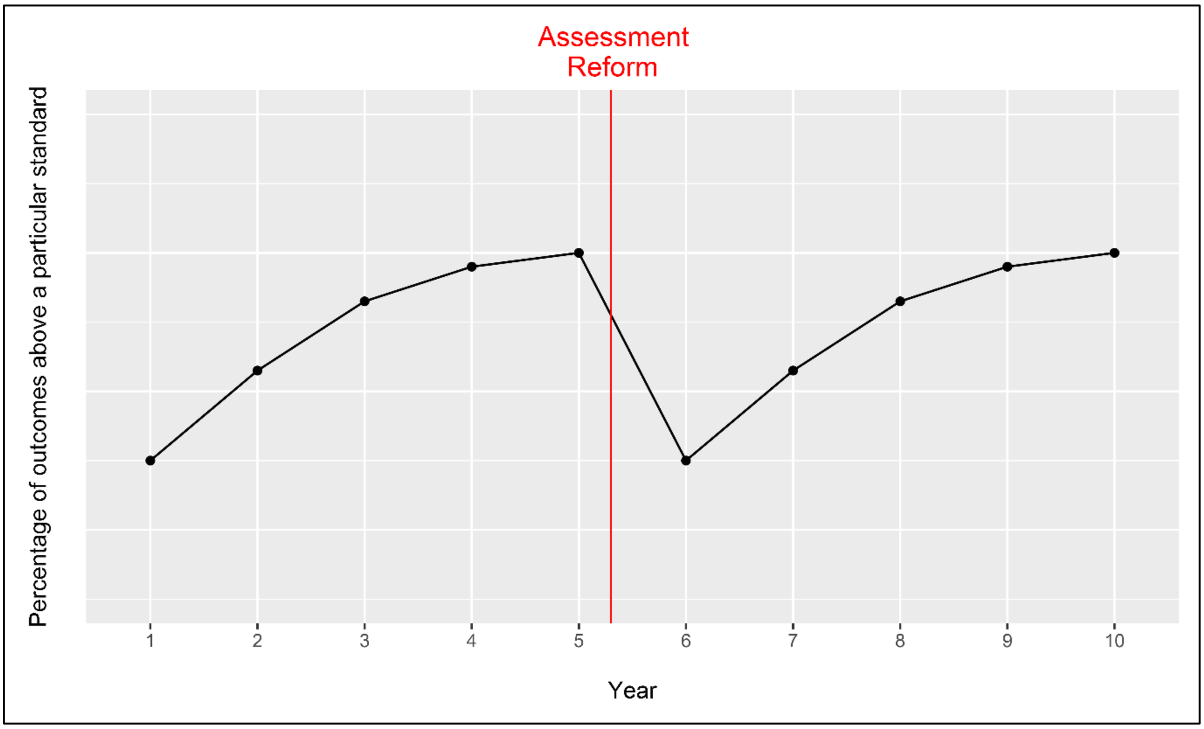

Figure 1. *Example sawtooth pattern caused by an assessment reform.*

Although evidence for this effect is fairly well documented in the US literature (eg Koretz & Barron, 1998; Koretz, Linn, Dunbar, & Shepard, 1991; Koretz, 2005; Linn, Graue, & Sanders, 1990; Linn, 1998, 2000), we are lacking investigations into its presence in UK education systems, despite the general acceptance that test-specific performance is affected in these ways following reform (eg Cresswell, 2003; Ofqual, 2015b). Gathering this information would allow us to better predict how performance and outcomes will be affected following the current set of reforms, which (at the time of writing) are underway for all GCSE and AS / A level assessments. This forms the rationale for the current research. This report will begin by outlining the possible causes of this pattern, before presenting 2 analyses of test-specific performance change following the last set of reforms (2010 for AS / A level and 2011 for GCSE).

2.1 Post-reform performance gains

This report focuses on the specific case of changes in test-specific performance following assessment change, so that we might better predict how performance on these assessments will be affected in the coming years (the UK is currently undergoing reforms to GCSE and AS / A level assessments). To avoid confusion with causes of increasing outcomes over longer periods of time (such as grade inflation caused by year-on-year leniencies in standard setting – see Pollitt, 1998), the term 'post-reform performance gains*'* shall henceforth be used to refer specifically to increases in test-specific performance caused by adjustments in teaching and learning being made over time in response to the introduction of a new assessment. Some examples of post-reform performance gains can be found in the US literature. For example, Koretz and Barron (1998) reported improving mean test scores over the first 2 years of a new reading assessment, but without any increases on a more established national reading assessment. Klein, Hamilton, McCaffrey, and Stecher (2000) reported similar findings. Although we lack concrete evidence to explain why these effects occur, there are some potential explanations.

Since the advent of high-stakes accountability testing in schools, teachers have become increasingly focused on ensuring high outcomes for their students, and as such, assessments have become 'instructional magnets' (Popham, 1987), leading to a narrowing focus of instruction (Stecher, 2002). This has the potential to produce inflated estimates of a student's mastery of the subject. Koretz (2005) provided the example of a vocabulary test, whereby it would not be feasible to test an individual's entire vocabulary range in a single assessment, and so one might construct an assessment to test knowledge of a narrow sample of a few select words. If that individual's education focuses on the tested materials, a result of 'teaching to the test' (otherwise known as *measurement driven instruction*[1]; Popham, 1987), learning a large number of the sampled words might lead to impressive test outcomes, but would only lead to very small gains in that individual's overall vocabulary size. The test therefore gives an inflated estimate of meaningful gains. Such an effect appears to be particularly problematic for 'high-stakes' assessments, where schools may face sanctions for failing to achieve goals (see Department for Education, 2015b, for

[1] Other terms such as 'washback', 'backwash', and 'curriculum alignment' have also been used to describe the influence of assessment on teaching and learning practices (see Cheng, 2000).

details on how this will work in the UK), as accountability systems are likely to encourage greater test-specific instruction, leading to inflated gains (Jacob, 2002). On the other hand, performance in a ‘low-stakes’ assessment is not associated with any rewards or sanctions, and so would normally have less impact on practice (see Stecher, 2002, for a lengthier discussion).

Koretz (2005) provided a more in depth discussion of the various methods used as part of measurement driven instruction but, in brief, teachers can reallocate teaching time to match the focus of the test (ie elements given more focus in the test are given the most teaching time, regardless of importance for broad domain mastery[2]), omit topics from teaching that are not included in the test, and coach students on the style of test items and test-taking heuristics (eg tactics for handling multiple choice questions). Each of these methods exist across a continuum in terms of severity and prevalence, ranging from legitimate instruction to outright cheating. When teachers are more familiar with the specifics of each test, they are more able to assist their students in these ways, which has the potential to elicit test-specific performance gains. When a new specification is introduced, this familiarity is lost, and so teachers lack the necessary information to appropriately direct their activities, leading to a ‘dip’ in their students’ test-specific performance for that series. As teachers become more familiar with the new test over time, performance once again begins to rise, explaining the sawtooth pattern shown in Figure 1. One would presume that this trend would eventually begin to plateau, as there are limits to the gains which can be made from familiarity with each test (this is why we have presented curved lines in Figure 1). However, some improvements can of course still be gained via relatively more gradual (ie relative to gains due to familiarity), general (ie not test-specific) improvements in schools and teaching.

In the UK, although questions do change series on series, important features of assessments, such as the overall structure of the test, the patterning of questions, and certain aspects of the mark schemes, can remain largely consistent from one series to the next. Thus, it is not uncommon for teachers in the UK to make use of the above methods (Baird, Chamberlain, Meadows, Royal-Dawson & Taylor, 2009; Baird, Daly, Tremain, & Meadows, 2009; Daly, Baird, Chamberlain & Meadows, 2012; Sturman, 2003 - the same has also be found abroad: eg Au, 2007; Shepard & Dougherty, 1991; Stecher, Chun, & Barron, 2004). As already discussed, the problem with such methods is that they lead to increases in test-specific knowledge, but weaker increases in the broader domain that each test is intended to measure (see Stecher, 2002). This can result in less preparation for further learning and education, and a fragmented understanding of the subject area (a concern voiced by the House of Commons, 2008, and Ofsted, 2008). Such factors can undermine the

[2] Koretz (2005) believed this to represent the strongest influence behind the Sawtooth Effect.

validity of assessments, and damage public confidence in qualifications. A narrow focused teaching plan can also place undue stress on students when teachers are wrong in their predictions of what will be in a future assessment (ie when they mistakenly focus their teaching on untested aspects of the specification – see Baird, Chamberlain, et al., 2009, p. 23). Despite these potential negative effects, the pressures placed on schools to make improvements elicited by high-stakes accountability testing can be a strong motivating force for teachers to teach to the test, as it is likely that improvements to test-specific knowledge can be made much more quickly than to broad domain knowledge (as suggested by Koretz & Barron, 1998). The reader is referred to Stecher (2002) and Cheng (2000) for more detailed reviews of the various positive and negative effects of high-stakes testing on teaching practices.

In addition, and in a related manner to the above point, increases in student familiarity with specific assessments may also lead to relatively rapid increases in students' test-specific performance in the first few years of a new specification. Baird, Ahmed, Hopfenbeck, Brown, and Elliott (2013) noted how students will do best in an assessment when they know what to expect, and are familiar with the ways in which questions are presented and marked, allowing students to focus on answering the questions without having to first concentrate on understanding what is required of them. The availability of past papers and marking schemes help with this (eg Elwood, Hopfenbeck, & Baird, 2015). Therefore, it may take time for cohorts of students to become familiar with a new test (eg with the increased availability of past papers and marking schemes), which may lead to improvements over time in test-specific performance as familiarity and preparedness improves (Ofqual, 2015b). Shepard (1988) provided some examples of these kinds of effects. In particular, she presented an example where percentage outcomes on an assessment were worse when questions were provided to students in an unfamiliar format, compared to when they were presented in a familiar format, despite placing the same academic demands on students.

Finally, increasing outcomes over time can also be explained by year-on-year leniencies during standard setting. For example, Ofqual (2015b) reported historical year-on-year increases in the percentage of students achieving A-grades at A level between 1996 and 2009, despite no meaningful gains in internationally standardised measures such as the TIMSS and PISA. Setting grade boundaries is notoriously difficult (eg Baird, 2007[3]) and when choosing between 2 marks in a given year, examiners have a tendency towards choosing the lower mark (eg Stringer, 2012). In the following year, examiners use the previous year's standards as a starting point to choose the grade boundary and, as before, are more likely to choose a more lenient

[3] Interested readers are also directed to Benton and Bramley (2015), who present some notable counterarguments to the idea that expert judgments are unreliable.

figure for this boundary. This is known as 'stepwise standards' (Pollitt, 1998), and can result in a year-on-year lowering of standards, which translates to inflated outcomes. The task of setting boundaries is made even more problematic when new specifications are introduced (Bramley, Dawson, & Newton, 2014), which may again contribute to sudden changes in outcomes following periods of reform. However, although these potential sources of inflation may possibly help to explain some instances of the Sawtooth Effect in the literature abroad, they are controlled for in the UK via the comparable outcomes approach to standard setting (see Ofqual, 2015c), and so are less relevant to the focus of the current report.

2.2 The 2016 reforms

The discussions thus far suggest that following assessment reform, one would expect to observe a trend where students' test-specific performance increases over time, as students and teachers become more familiar with new approaches, and then begins to plateau once the limits of familiarity are approached. As noted previously, this has particular relevance to the present climate, as the UK is currently undergoing reforms in all GCSE and AS / A level assessments (see Ofqual, 2015a). There are several reasons to expect that the Sawtooth Effect would occur as a result of these reforms. Firstly, although the presence of the effect in the American literature does not necessarily imply the same for the UK, the fact that core skills such as reading and mathematics seem to be affected in this manner (eg Koretz et al., 1991) may suggest that many UK assessments could also follow the same trends, as these core skills will be important for most, if not all, general assessments. Secondly, as discussed previously, high-stakes accountability testing (which applies to GCSEs and AS / A levels) means that students and teachers are increasingly motivated to achieve high outcomes, and are likely to direct their teaching / learning focus towards the specifics of each test to ensure the best possible outcomes. One would therefore expect performance to be relatively lower in the first year, followed by a period of rapid post-reform performance gains as students and teachers become more familiar with the new specifications, followed by a period of relative stability, or at least less rapid increases caused by more general improvements in teaching and schools.

2.3 Estimates of magnitude and duration

Although we can have a degree of confidence in our prediction that there are likely to be some improvements in test-specific performance for GCSEs and AS / A levels in the years following the current reforms, the size of such changes is difficult to predict, as there has been little evidence for the magnitude of the Sawtooth Effect in the literature. What evidence we do have from the UK is outdated, and may not translate well to modern teaching practices. For example, Cresswell (2003, citing data from the School Curriculum and Assessment Authority) provided evidence consistent with the expected pattern, showing that the percentage of students achieving a C grade or better in English increased by around 3% each year between 1988 (when GCSEs replaced the old-style O-levels) and 1991, with no test-specific improvements occurring between 1991 and 1995. Other evidence for effect-sizes comes from the US, although this does not translate well to UK awarding. For example, Koretz et al. (1991) reported a ‘dip’ of half an academic year in ‘Grade-Equivalent Scores’ following the introduction of new mathematics and reading assessments. Equivalent measures do not exist in the UK and so comparisons cannot be made.

Prior estimates for the length of time it takes to reach the assumed limits of familiarity allow for some comparisons, but are also somewhat inconsistent. For example, Linn et al. (1990) reported that percentile rank state means in reading and maths returned to previous levels in the fourth year after reform, but did not appear to stabilise during this time (possibly due to other factors such as stepwise standards). The results reported by Koretz et al. (1991), however, seemed to suggest that outcomes in reading and maths stabilised following the second testing series of a new assessment. The data provided by Cresswell (2003) suggested that outcomes in GCSE English may have stabilised following the fourth assessment series of the first GCSEs. Based on these few estimates, one might therefore loosely expect that the more rapidly changing portion of the Sawtooth Effect lasts somewhere between 2 and 4 assessment years.

Further limiting our predictive ability is the fact that the duration and magnitude of the Sawtooth Effect will likely differ between the US and UK, and between 1988 and the present day, limiting the generalisability of each of the aforementioned reports to the current UK assessment reforms, as schools may handle reforms in different ways, leading to different patterns of post reform change. In addition, it is unclear whether this issue is specific to assessments of English and maths, which all prior research has focused on (to our knowledge), or whether the Sawtooth Effect is a more general issue to all high-stakes assessments. There is a clear importance, therefore, to examine a recent post-reform period in the UK, to foster a more powerful prediction of how GCSE and AS / A level outcomes will be affected in the coming years, as centres and students adjust to the new assessments. Such knowledge would also allow us to improve our theoretical understanding of the size of the gains to performance / outcomes that can be made following assessment reform, reflecting how quickly teachers and students may tend to adjust to the introduction of a new assessment.

3 Study 1: Changes in grade boundaries

The changes in cohort performance in assessments following the last set of reforms (2010 for AS / A levels and 2011 for GCSEs) were examined. In the current case, it would have been unfruitful to extend the analyses prior to the 2010 / 2011 reforms in order to observe a ‘dip’ in performance. This is because one can only draw meaningful comparisons if outcomes were awarded in the same manner. Due to differences in approaches to standard setting and awarding between the specification periods of interest, any comparisons of grade boundaries across these periods would be difficult to interpret. Nevertheless, this method does allow one to observe the presence of any post-reform performance gains during the specification lifespan (ie to the right of the red line in Figure 1).

The intended task is also enabled by the ‘comparable outcomes’ approach to standard setting, which has been used in England to regulate standard setting since 2010[4]. Comparable outcomes were introduced as part of an ‘ethical imperative’ to ensure that cohorts did not receive low percentage outcomes (when compared to previous cohorts) simply because they were the first to sit the new series of assessments (Cresswell, 2003; Ofqual, 2015b), therefore deliberately aiming to prevent the occurrence of the sawtooth pattern in terms of GCSE and AS / A level outcomes (although specific reference to the Sawtooth Effect has not been made). This approach is also used to control grade inflation over time caused by stepwise standards shift. Maintaining percentage outcomes comparable with previous years is particularly important for fairness when one considers, for example, that the UCAS system treats grades as equivalent across cohorts when considering university admissions, even though the demands for achieving each grade may have changed (as argued by Black, 2008).

The specifics of the comparable outcomes approach have been dealt with in greater detail elsewhere (eg Benton & Lin, 2011; Benton & Sutch, 2014; Bramley & Vidal Rodeiro, 2014). In brief, it aims to produce predicted grade distributions for a subset of students that have been matched with previous test scores (Key Stage 2 assessments are used as a prior attainment match for GCSEs, and GCSEs are used as a prior attainment match for AS / A levels). The initial distribution from the first year of awarding (known as the ‘reference year’) becomes the basis for future awarding[5]. For example, if Key Stage 2 outcomes are the same for the cohort in the reference year compared to the current year, then GCSE percentage outcomes are

[4] In fact, comparable outcomes have been used as a guide in standard setting since 2002, but they have only been more strictly adhered to since 2010 (see Bramley et al., 2014, for a historical overview of its use).

[5] This is a rather simplified example. In actual UK awarding, the first 2 years of awarding are averaged to form the reference. See Ofqual (2015c) for further details.

also not expected to change, and grade boundaries for the current year will be set with this in mind. It is perhaps worth noting that although exam boards can submit evidence relating to the script performance of students to support an award away from prediction (out of tolerance), this is uncommon. Statistical predictions form the strongest piece of evidence to inform the setting of boundaries but other forms of evidence are also considered.

The comparable outcomes approach means grade boundaries for the period of interest will have been positioned so that outcomes remained stable (in relation to prior attainment). Just as grade boundaries in the first year after reform are set in order to preserve outcomes and discount the drop in performance, subsequent series might see a raising of grade boundaries in order to discount test-specific improvements in performance (ie those that are not supported by improvements in matched Key Stage 2 / GCSE scores). However, this approach does allow for the analysis of performance change by proxy through analyses of changes in grade boundaries over time. If test-specific performance improved during the post-reform period (eg due to student / teacher familiarity), one would expect to observe rising grade boundaries to smooth out this effect. If test-specific performance decreased, grade boundaries would also have decreased. This proposition forms the foundation for following analyses.

3.1 Hypotheses

Taken together, the previous discussions suggest that test-specific performance (and by proxy, grade boundaries) should have increased following the last set of reforms, mirroring the increasing familiarity and preparedness of teachers and students with the new specifications. Assuming that these changes were caused mainly by teacher/student test familiarity, one would expect to see any rapid gains in test-specific performance to lessen after a few years, as presumably there are limits to the gains which can be made from familiarity with each test.

3.2 Data collection

A full table of the units (assessments)[6] investigated, along with the relevant inclusion / exclusion criteria, is available from the authors upon request.

Raw unit grade boundaries for the just-ended specification period (2010 to 2015 for AS / A levels and 2011 to 2015 for GCSEs) were sourced from 4 exam boards: AQA,

[6] GCSE / AS / A level qualifications are made up of a number of units. Each unit is assessed by an exam or piece of coursework / controlled assessment, and unit marks are aggregated to produce outcomes for the qualification.

OCR, Pearson Edexcel, and WJEC. These raw boundaries were standardised by converting them to percentages of the maximum marks per unit. Units were included if they were present throughout the whole period of interest to avoid bias in earlier or later years, caused by discontinued or late-starting units. For the same reason, only summer examination series were included, as January series (and November and March series, where they existed) were discontinued in 2014 (a process known as linearisation[7]). Units from small entry subjects (*n* < 500 in any year) were also removed as these are not subject to the comparable outcomes thresholds set by Ofqual (2015c).

As performance, and therefore grade boundaries, are sensitive to curriculum change, it was necessary to take into account the various smaller reforms that occurred during this period. These changes are outlined in Figure 2. In order to avoid any biasing influence of these reforms, affected units were also removed from the analyses to follow. Any specifications that were introduced after 2010 / 2011 were already excluded because they did not span the entire period of interest. GCSE units that were awarded extra marks in spelling and grammar from 2012 onwards were also excluded as this reform was significant enough to expect changes in boundaries in these subjects.

A summary of all the inclusion / exclusion criteria can be found in Table 1. Although these criteria allowed tighter control over potential sources of bias in the data (and so we should have greater confidence in the results gleaned from non-excluded units), it is worth noting that when analyses were conducted on all units (ie with no exclusion criteria in place), this produced approximately the same trends as when units were excluded according to Table 1. The fact that these trends were broadly similar to the trends observed in the analyses to follow (despite the various changes depicted in Figure 2) lends support to the representativeness and generalisability of the following analyses to the whole range of GCSEs and AS / A levels. For comparison, grade boundary graphs for **all** units within the dataset can be found in Appendix A (Figures 35 and 36).

[7] For students that certified (ie completed) their GCSEs in 2013 or earlier, at least 40% of the assessments that made up each qualification must have been taken in the final June series (known as the 40% terminal rule). This meant that up to 60% of assessments could have been taken in an earlier series during the course of study (eg the previous June or January, or even January of the preceding year). However, from 2014 onwards all units had to be taken at the end of the course, meaning that all exams were sat in the final June series.

The linearisation process for AS / A levels was slightly different. Similar to GCSEs, January series were no longer available from 2014 onwards. However, no terminal entry rules were introduced, meaning that students still sat assessments in 2 consecutive June series. For example, in most cases students would have sat their AS level exams in June 2014 and their A2 exams in June 2015.

	Academic Year					
	2009/2010	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015
AS / A level	New specifications launched				January exams discontinued (linearisation)	
GCSE		Most new specifications launched	New specifications for English, maths, ICT, and science	New specifications for additional science, biology, chemistry, and physics Marks now awarded for SPAG in English lit., history, geography, and religious studies New early entry rules introduced	January exams discontinued (linearisation) Speaking and listening components removed from English and English language	New history and English literature exams Standards raised for English and geography

Figure 2. *Timeline of reforms.*

Note. "SPAG" = Spelling and grammar

Table 1. *Summary of inclusion / exclusion criteria*

Criterion	Included if:	Excluded if:
Lifespan of the unit	The unit existed across the entire period of interest.	The unit did **not** exist across the entire period of interest.
Entry size of the subject	Entry size was greater than 500 in each year.	Entry size was less than 500 in any given year.
The introduction of marks for spelling and grammar in 2013	The unit was **not** awarded marks for spelling and grammar.	The unit was awarded marks for spelling and grammar.

3.3 Analyses and results

3.3.1 AS / A level grade boundaries

Mean A- and E-grade boundaries[8] (with 95% confidence intervals) for all AS / A level units (not excluded by the criteria set in Table 1) were plotted over time (Figure 3). Upwards trends can be seen for both boundaries. Also displayed in Figure 3 are the differences between years (the values in bold). The rate of change for both grades between 2010 and 2011 was lower than expected, as one would expect the greatest improvements to be made in these first few years. The A-grade difference between 2013 and 2014 was also lower than expected (possibly because of the effects of linearisation; ie the discontinuation of January exams; see Figure 2). These factors make it somewhat difficult to determine when the rate of change became less rapid (ie when the limits of familiarity began to be approached). The rate of change did seem to lessen for A-grades following either the third or fourth assessment year (ie 2012 or 2013), but the rate of change for E-grades continued largely at the same rate throughout the specification lifespan (other than a relatively large change occurring between 2011 and 2012).

[8] These particular boundaries were chosen as the comparable outcomes approach is used only to set certain standards for AS / A levels. These are known as 'judgemental boundaries'. Other boundaries (ie B – D), are known as 'arithmetic boundaries' and are set evenly between the A- and E-grades. See Ofqual (2011, pp. 63–68) for a more detailed description of the specifics of setting grade boundaries.

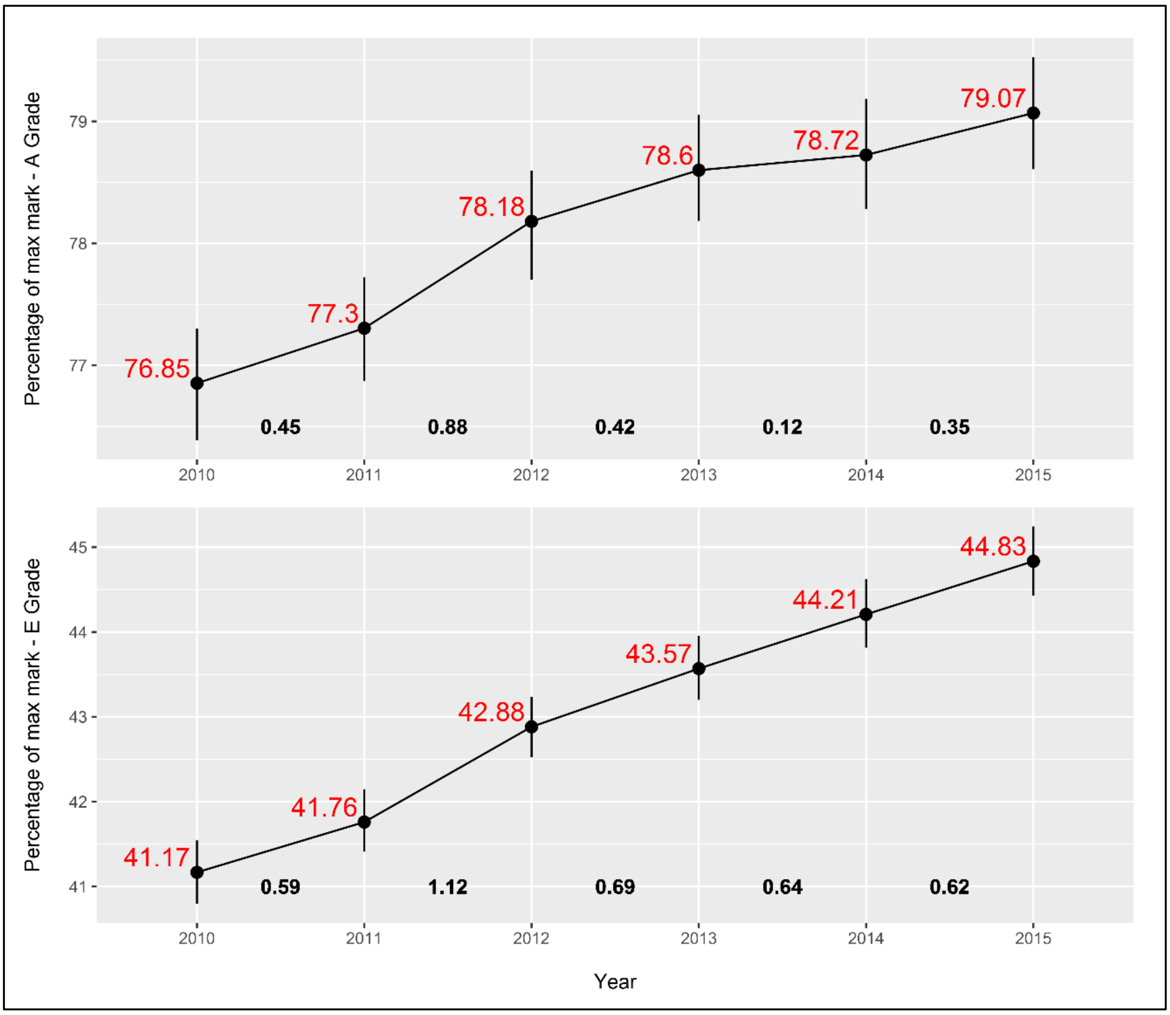

Figure 3. *AS / A level unit boundaries*

Note. n = 1,141 A-grade boundaries and 1,135 E-grade boundaries[9]. Points indicate the mean grade boundary, and bars indicate the 95% confidence interval. Values in bold indicate the size of the difference in mean boundaries between adjacent years.

[9] The difference in these numbers is due to some missing data.

It was important to take into account the variation of percentage outcomes over time (ie the cumulative percentage of students achieving an A / E-grade in each year compared to that predicted). Although the comparable outcomes approach is used to stabilise outcomes against prior attainment between cohorts, percentage outcomes can vary within a certain tolerance threshold[10]. Should outcomes vary in a pattern opposite to that found in Figure 3 (ie by gradually decreasing over time) then this variation may explain why boundaries have risen over time, rather than because of any change in underlying performance.

As unit level outcome data is not recorded, we were unable to match boundaries with percentage outcomes for each unit. Mean subject level outcomes were instead plotted, as this method still allows one to identify any systematic changes in averages over time. Subjects were only included if there were more than 500 students entered for each assessment series, because comparable outcomes thresholds are only applied to entries of this size (see Footnote 10). Figure 4 shows the actual-versus-predicted outcomes at each grade, averaged over all subjects[11]. The 2 confidence interval spikes appeared to be due to single influential outliers (ie one A-grade outcome in 2010 and one E-grade outcome in 2013). When these outliers were removed from the sample, the values shown in Figure 5 were found.

[10] In UK awarding, awarded outcomes can depart from predicted outcomes up to 1% for matched entries (ie students with matching KS2 scores for GCSE, or matching GCSE scores for AS / A levels) of 3000 or more, up to 2% for matched entries of 1001-3000, and up to 3% for matched entries of 501-1000. There is no threshold tolerance for matched entries of 500 or less, which is why small subjects have been excluded. See Ofqual (2015c).

[11] The data used for this analysis consisted of the pre-award outcomes. Various post-award adjustments may have occurred, such as to account for partial absences, additional matching of candidates, and changes due to enquiries about results.

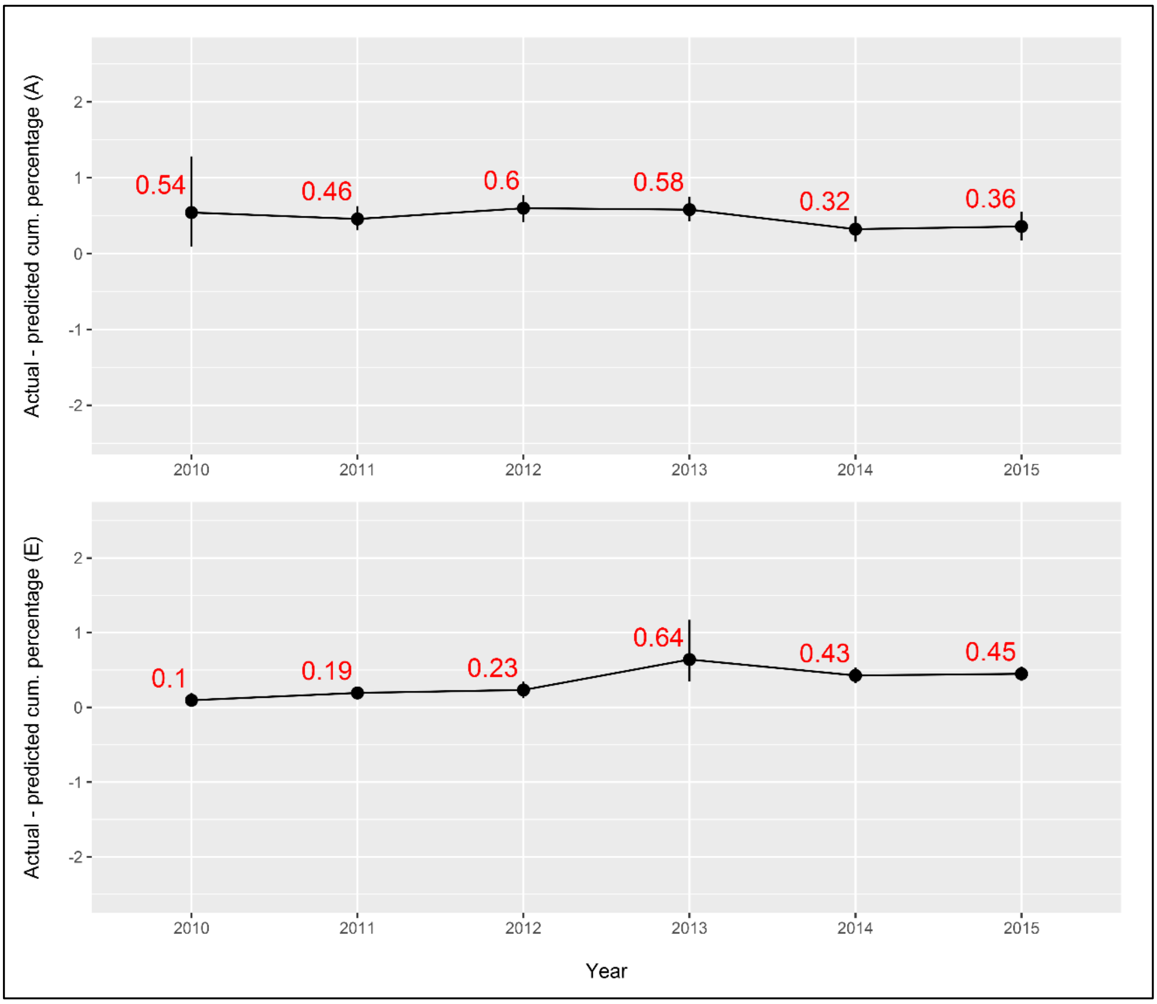

Figure 4. *Deviation of actual AS / A level outcomes from predicted outcomes over time*

Note. Number of subjects = 118.

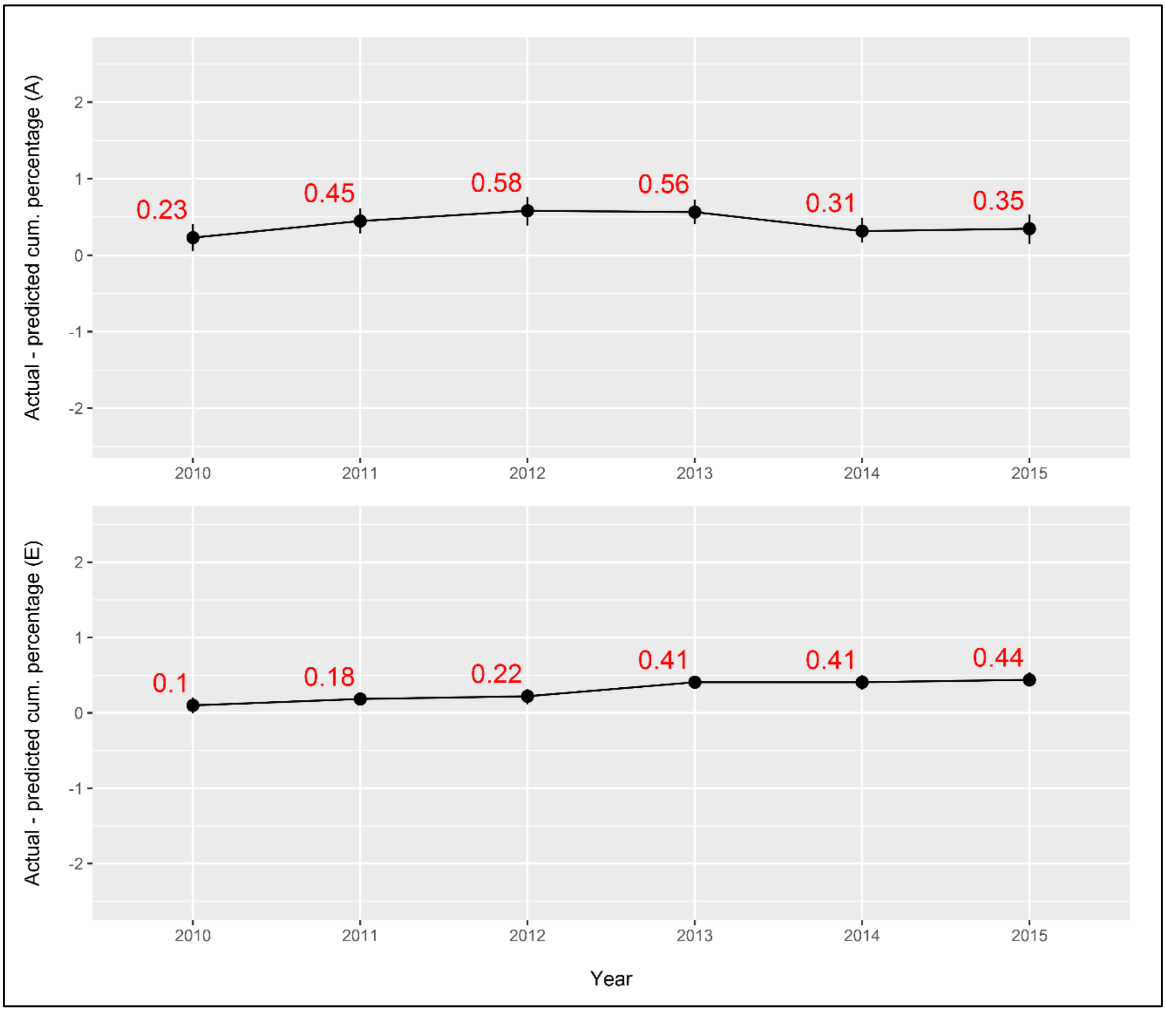

Figure 5. *Deviation of actual AS / A level outcomes from predicted outcomes over time*

Note. Number of subjects = 117. One outlier has been removed from each grade.

Results suggested that average percentage outcomes did not vary substantially from what was predicted, thus comparable outcomes were achieved (well within the thresholds outlined by Ofqual, 2015c). In fact, the most likely explanation for these small changes is because grade boundaries must be set to the nearest integer, and so there is often no single mark upon which the predictions can be met exactly. The fact that outcomes have remained stable against a context of rising grade boundaries supports the conclusion that test-specific performance did improve during this time, as successive students have had to achieve more marks to achieve the same grade (thus reflecting better performance in later cohorts). However, the proposition that variations in grade boundaries equate to variations in underlying performance relies on the assumption that awarded outcomes match predicted outcomes perfectly (ie the lines in Figure 5 should ideally be flat along 0%). At present, at least some of the variation in grade boundaries (as presented in Figure 3) could be explained by differences in year-on-year deviations from predicted outcomes, rather than necessarily a change in underlying performance. We therefore decided to adjust the findings relating to grade boundaries to take into account the deviation from predicted outcomes and, as we were not restricted by rounded integers in the current report, the use of decimals could be used to refine these values. For example, it seems as though there were, on average, 0.23% (of the total entry per subject) more students achieving A-grades in 2010 than were predicted, and so we needed to raise the grade boundary to account for this fact (else this affects our interpretation of the grade boundary trends).

Unfortunately, this was no easy task, as it is impossible to establish how much one would need to adjust individual unit boundaries to return actual subject-level outcomes to what was predicted without having candidate level data for all units (which were not available to us). This is because there are a large number of possible combinations of units that can be aggregated to arrive at subject level outcomes, making it difficult to know how much adjusting each individual unit boundary will have an effect on outcomes (eg see Bramley & Dhawan, 2012). However, by making observations at the average level, one can make an estimation of how much the **average** boundary should be changed to return to **average** predicted outcomes. For example, by looking at the subject level data we know that, on average, 25% of students got an A-grade in 2010 AS / A levels. Based on this, if we assume that on average 25% of students also got an A in each unit*,* then changes in grade boundaries at unit level will roughly translate to changes in average outcomes at subject level. There will certainly be a margin of error here, but this method should nevertheless provide a rough approximation of how much these deviations from predicted outcomes can explain the changes in grade boundaries observed.

To determine how many students would be affected by a change in grade boundaries, a frequency distribution of outcomes was needed. Candidate level data was not available for all subjects and so in order to determine the ‘average distribution’ across the current sample of subjects a data simulation procedure was used, with a number of constraints in place to make the simulated dataset as representative of the sampled subjects as possible. The process of how this data was simulated is detailed in Appendix B.1. In brief, the position of the distribution on the x-axis was determined by calculating the grand mean of subject-level outcomes. The shape of the distribution was determined by calculating the mean standard deviation of subject-level outcomes (correcting the kurtosis), and by the mean cumulative frequency of students achieving each grade (correcting the skew).

Grade boundaries were adjusted to reduce or increase the number of simulated students achieving each boundary (see Appendix B.2 for more detail). For example, increasing the mean 2010 A-grade boundary by 0.08% reduced the number of simulated students achieving an A-grade by 0.23% of the sample (returning the 2010

A-grade to predicted outcomes). Figure 6 shows how the values displayed in Figure 3 changed after these adjustments were made. As one can see, this had little effect on the trends over time, suggesting that a variation in actual outcomes did not explain the changes in grade boundaries that were observed.

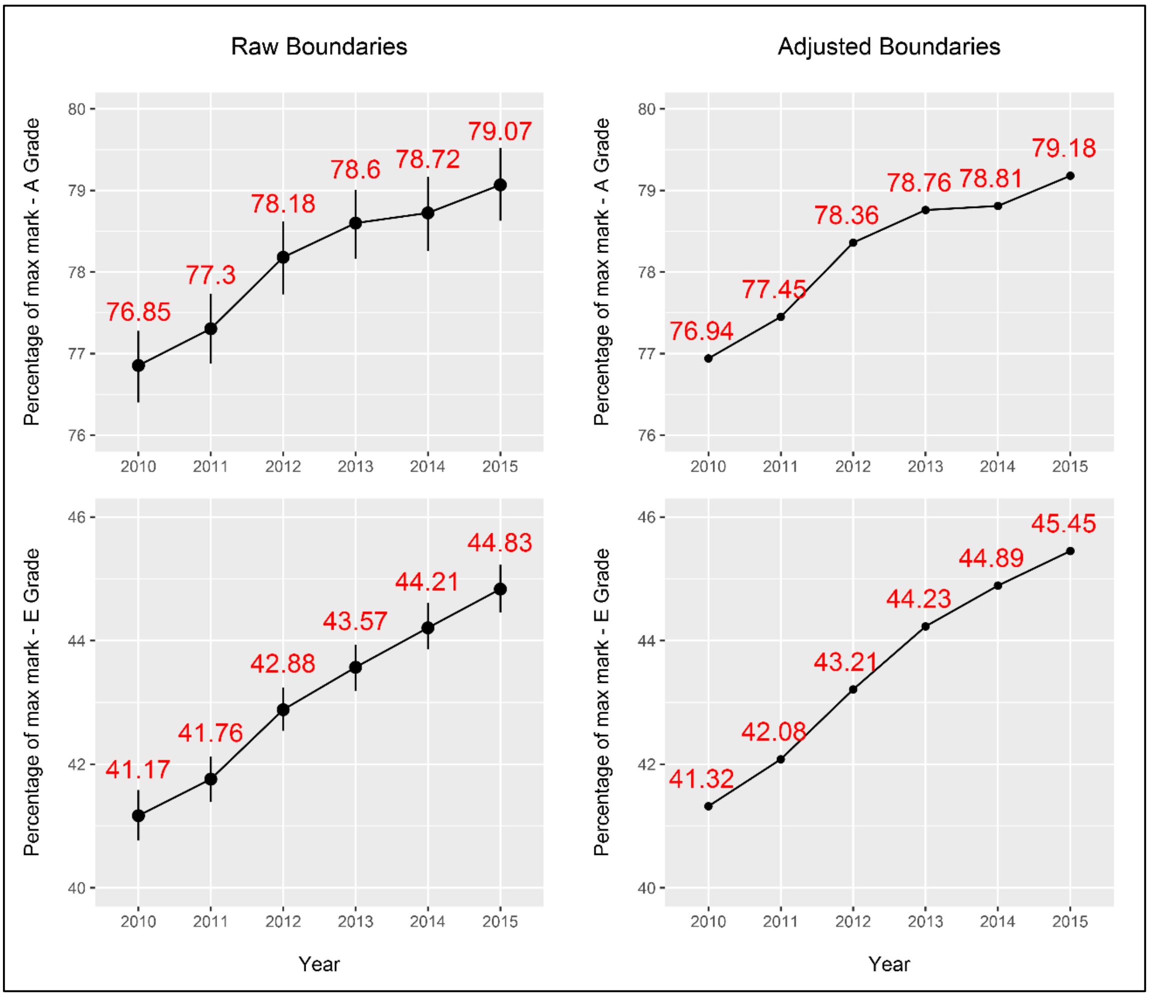

Figure 6. *Grade boundaries adjusted for deviation from predicted outcomes.*

During unit aggregation (ie when calculating overall subject grades from multiple units), outcomes on coursework units can affect the awarding of examined units (eg see Bramley et al., 2014), because coursework marks increase year-on-year to a greater extent than exam marks (eg see Ofqual, 2013). Analyses of raw boundary changes were therefore repeated using only subjects that were assessed entirely by examination, to determine whether the inclusion of units assessed by coursework provided a distorted picture of grade boundary change. However, the overall trend for examined only subjects (Figure 7) did not appear to depart substantially from the trends presented in Figure 3, meaning that examined-only subjects were affected by the reforms in a similar manner to that reported previously.

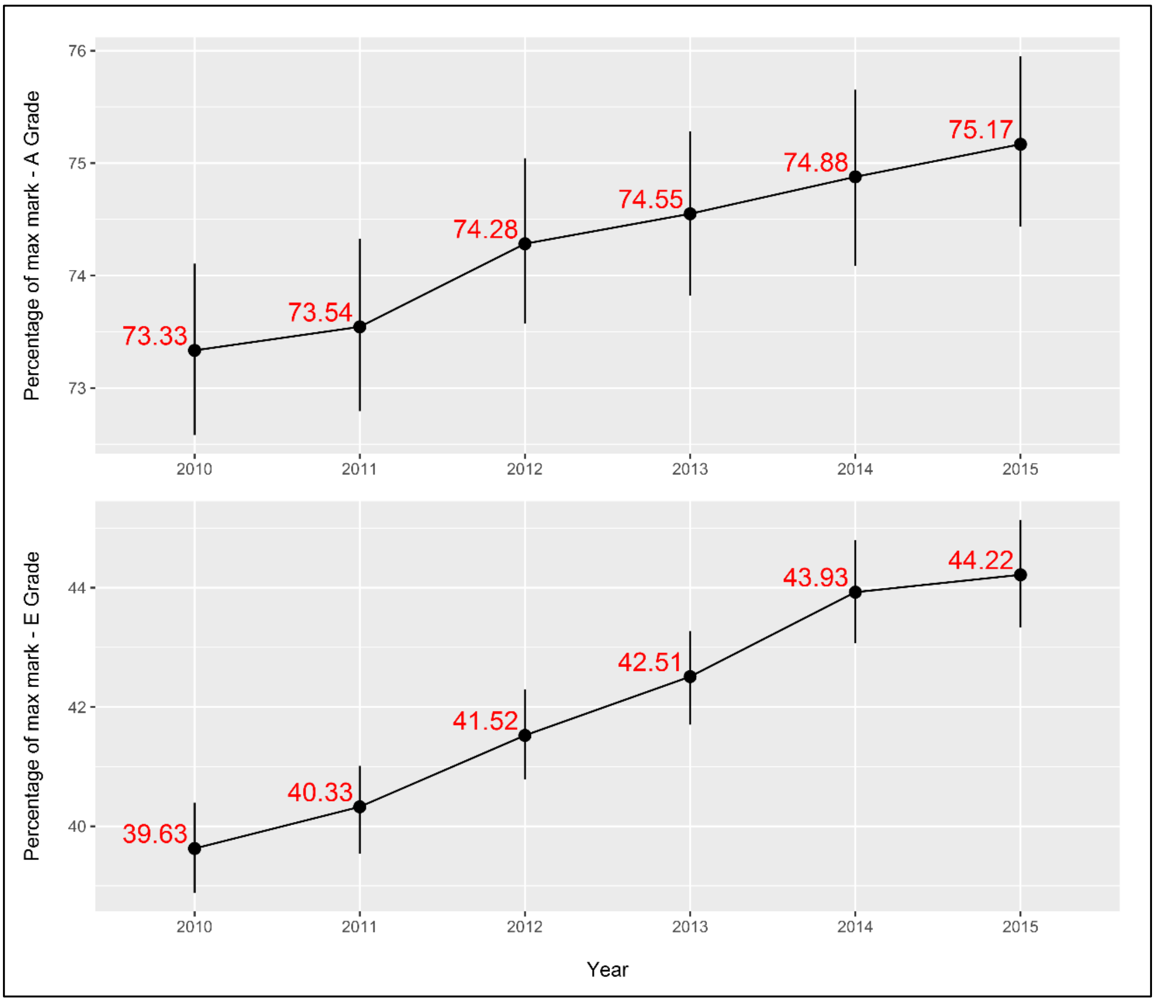

Figure 7. *AS / A level boundaries: Subjects that were assessed entirely by exam*

Note. n = 296 A-grade boundaries and 293 E-grade boundaries.

For the sake of completeness, boundaries were also plotted for units from subjects that were not assessed entirely by exam (ie from subjects that were either assessed entirely by coursework or a mix of coursework and exam) (Figure 8). The overall patterns for both A- and E-grades were consistent with that found for all units (ie Figure 3); most likely because these units made up the majority of the total AS / A level sample.

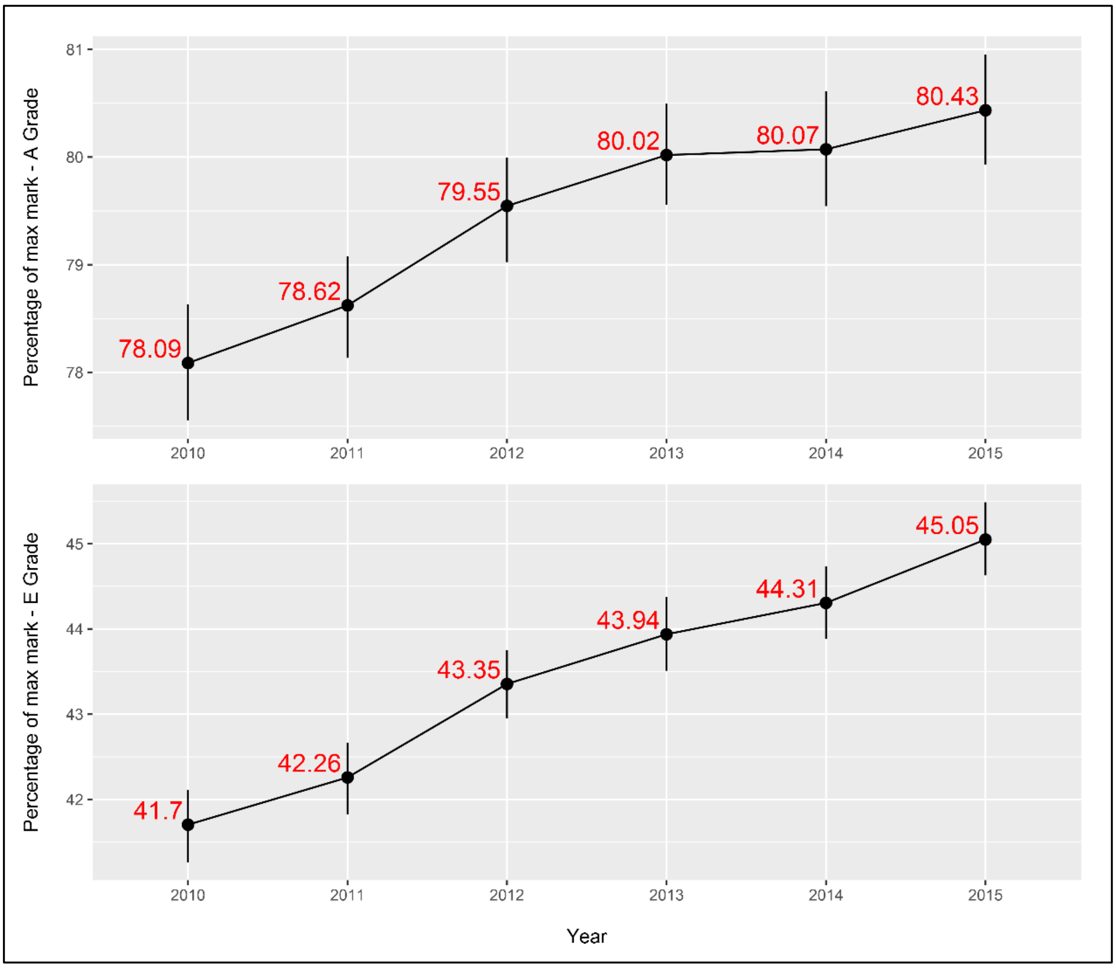

Figure 8. *AS / A level boundaries. Subjects that were **not** assessed entirely by exam*

Note. n = 845 A-grade boundaries and 842 E-grade boundaries.

3.3.2 GCSE grade boundaries

The same graphs were plotted as in the previous section. Mean grade boundaries (with 95% confidence intervals) for all units were plotted over time for A-, C-, and F-grade boundaries[12] (Figure 9). The differences between years are also displayed in bold. These results were much clearer (in terms of support for the predicted pattern) than that for AS / A levels, and suggested that the rate of boundary change was quite rapid between the first 3 assessment years (2011 to 2013), with the rate of change between the latter years (2013 to 2015) being much more gradual (and slightly in reverse). This is again consistent with the pattern expected to be caused by post-reform performance gains, and suggests that the limits of familiarity may have approached in the third assessment year of the new specifications.

As before, the deviation of actual outcomes from predicted outcomes was also considered, as these may account for some of the variation in grade boundaries over time. Outliers again affected our interpretation of results (Figure 10), and so 2 influential outlier subjects were removed (Figure 11). Similar to before, all actual outcomes were within 0.5% of prediction (being most likely explained by integer rounding), and therefore unlikely to explain the degree of changes over time observed in Figure 9. This was again supported by the results of the data simulation method (see Appendix B.1 and B.2), which showed that adjusting the values of Figure 9 to take into account the deviation from predicted outcomes in Figure 11 would have little effect on one's interpretation of the results (Figure 12).

[12] As in the eighth footnote, these are the judgmental boundaries for GCSEs.

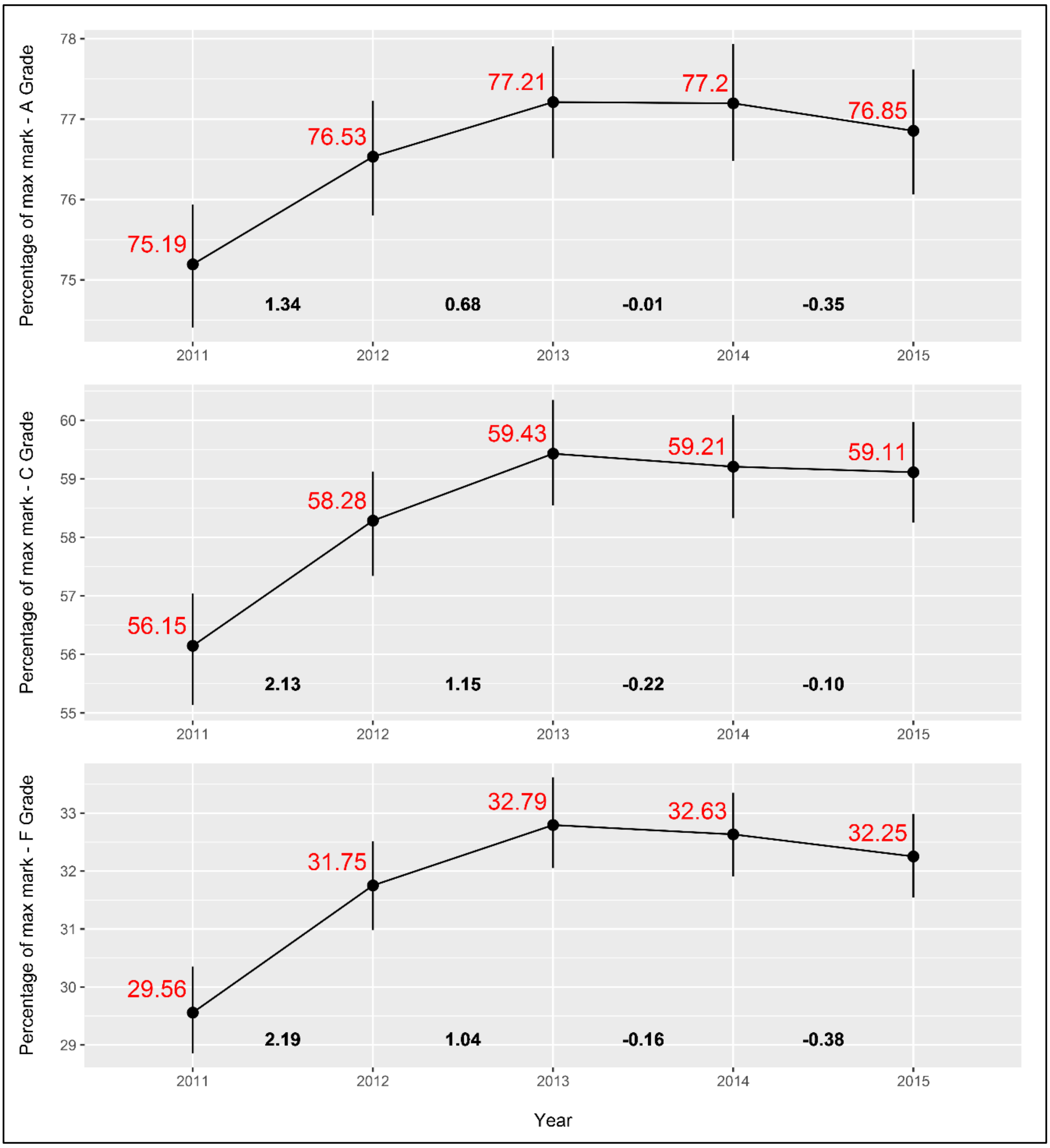

Figure 9. *GCSE unit boundaries.*

Note. n = 466, 550, and 466[13] for A-, C-, & F- grades respectively. Points indicate the mean grade boundary, and bars indicate the 95% confidence interval. Values in bold indicate the size of the difference in mean boundaries between adjacent years.

[13] The difference in these values is due to tiering. For higher tier assessments, there is a grade range of A*-D, while foundation tier assessments have a grade range of C-G. Thus, there are more C-grade boundaries in the sample due to the overlap between the grade range of higher and foundation tier assessments.

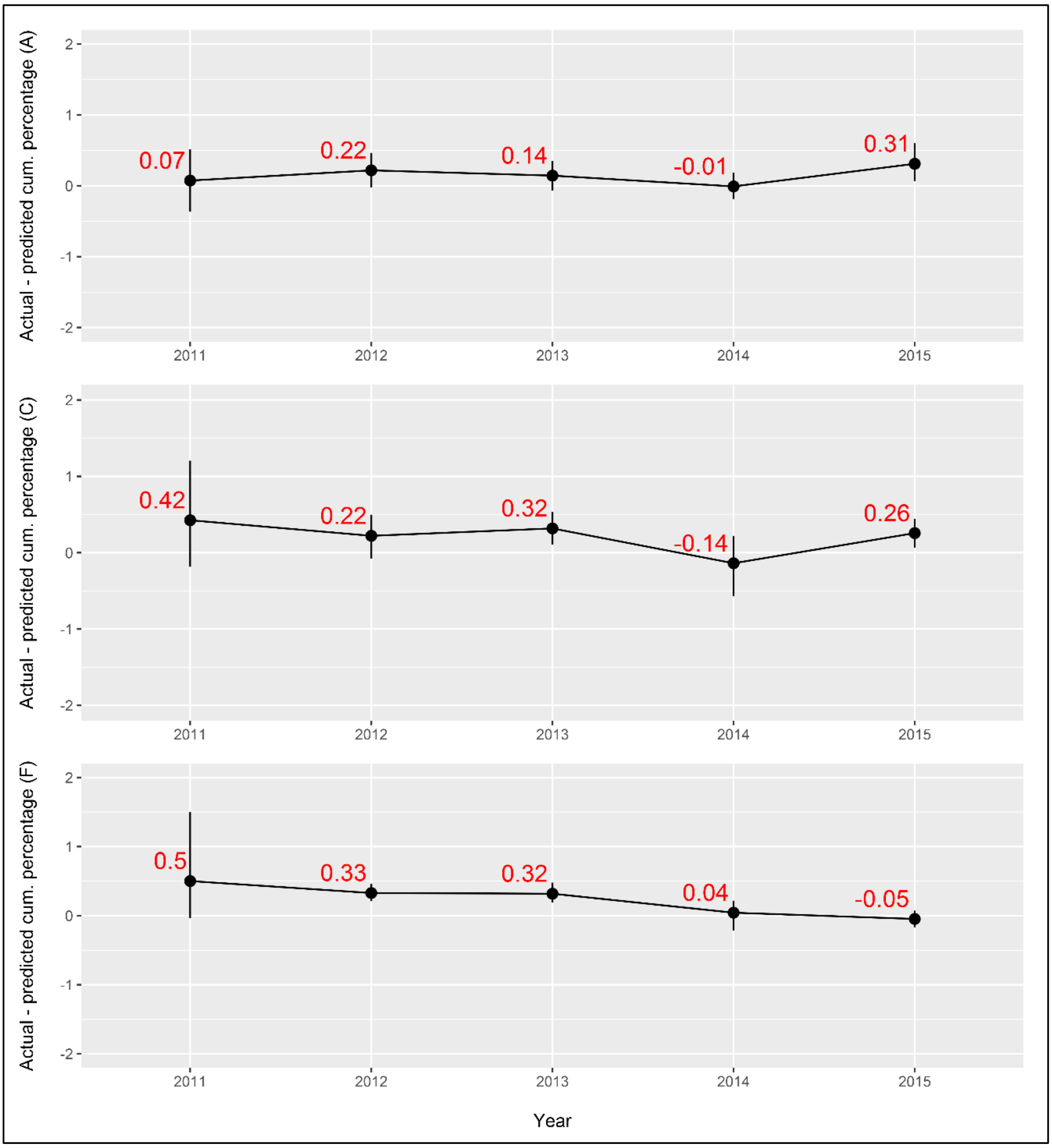

Figure 10. *Deviation of actual GCSE outcomes from predicted outcomes over time*

Note. Number of subjects = 96.

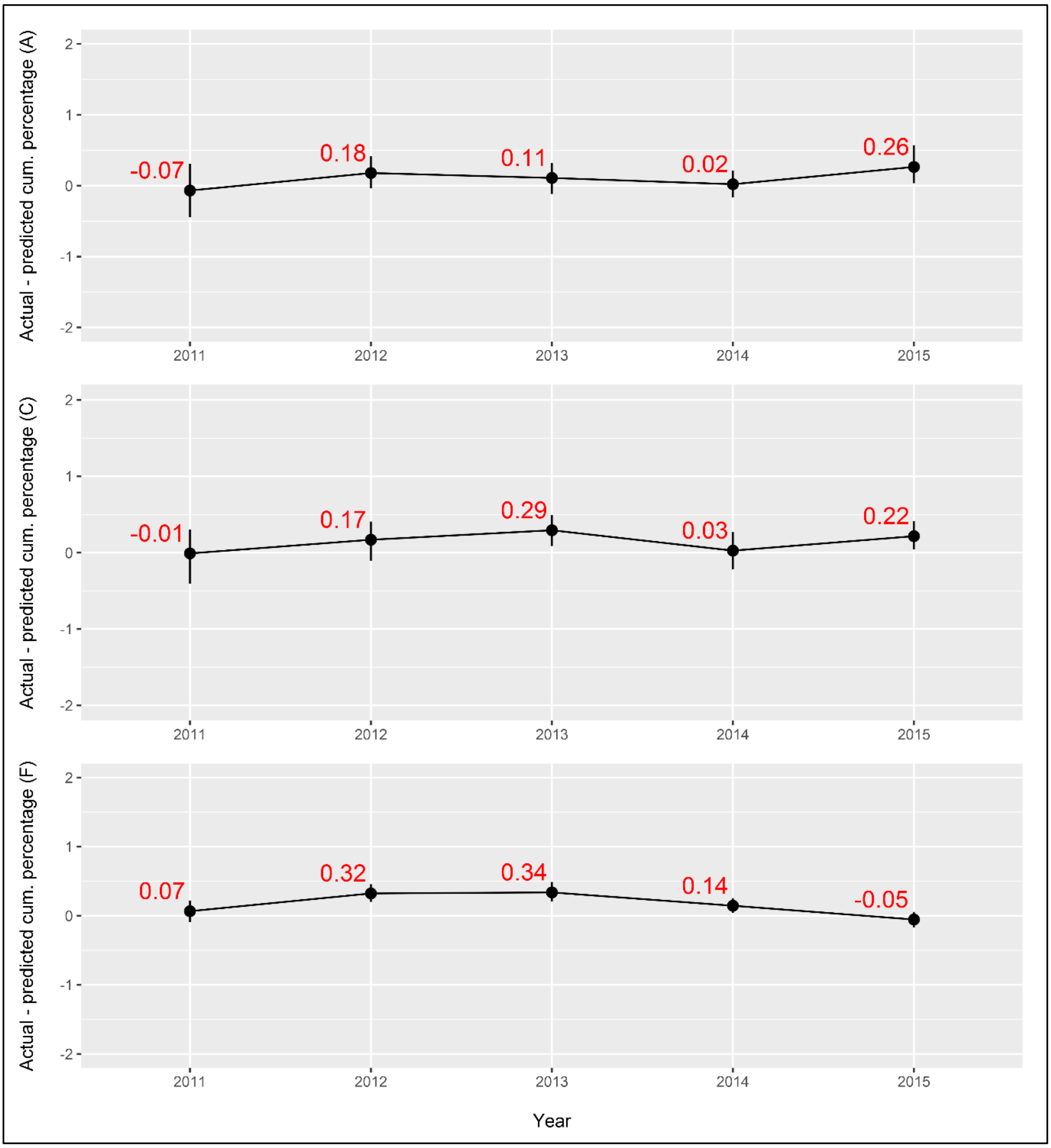

Figure 11. *Deviation of actual GCSE outcomes from predicted outcomes over time*

Note. Number of subjects = 94. Two outlier subjects have been removed*.*

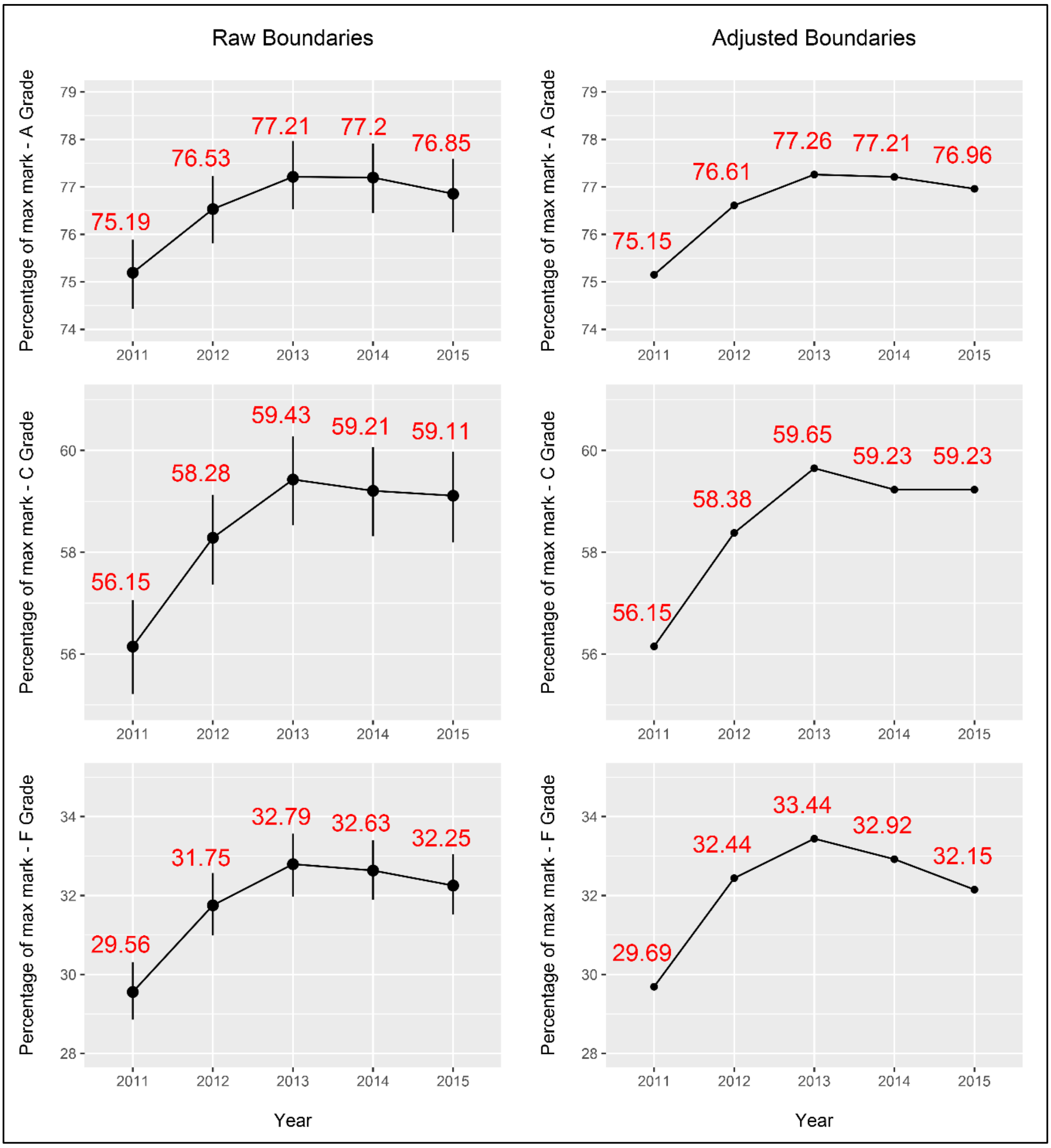

Figure 12*. Grade boundaries adjusted for deviation from predicted outcomes.*

As with the AS / A level analyses, the graphs were replotted, removing any subjects from the analyses which contained any controlled assessments. As can be seen in Figure 13, roughly the same pattern is produced as in Figure 9, although it appears that grade boundaries for these examined units exhibited a visible decrease between 2013 and 2015.

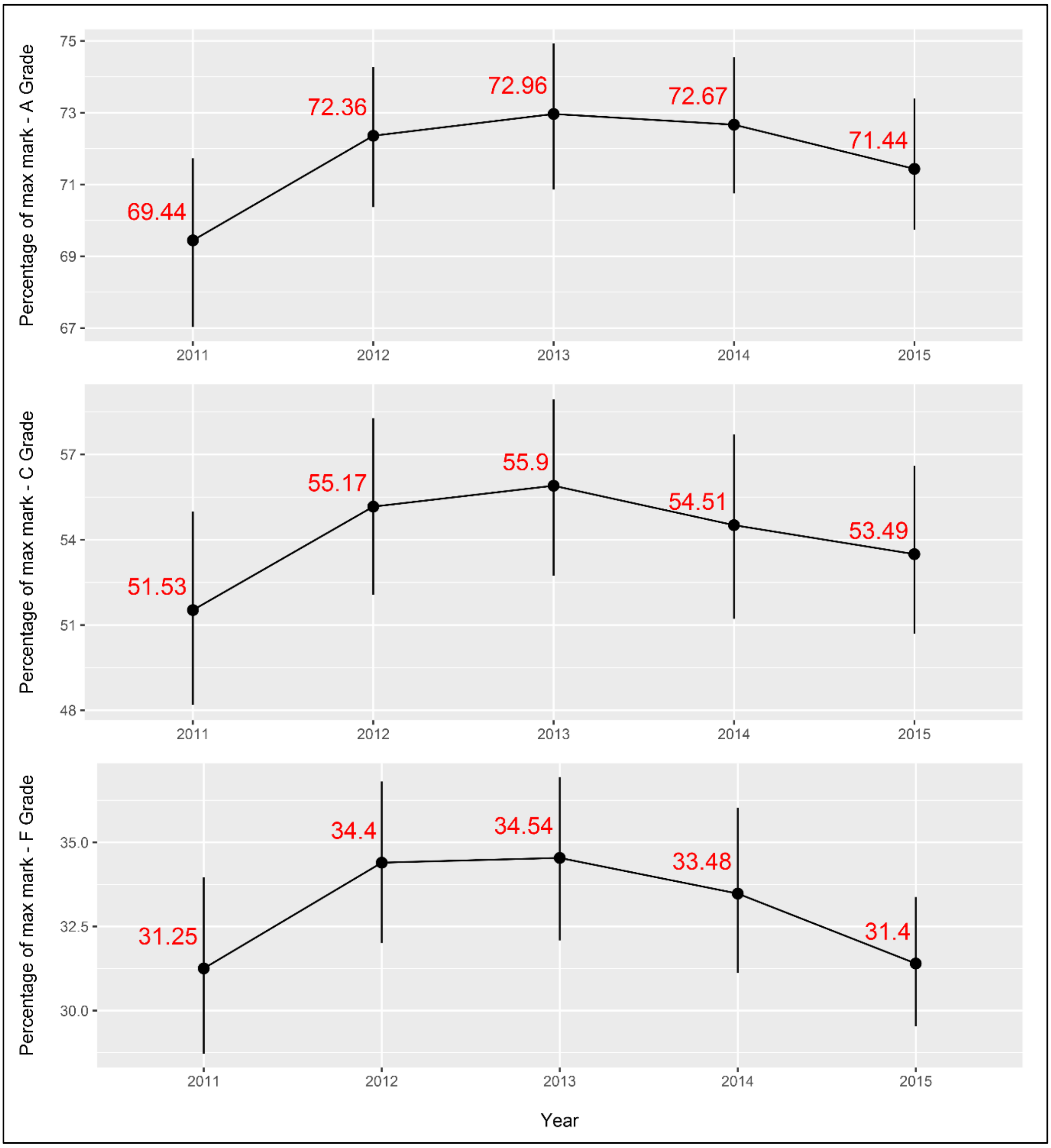

Figure 13. *GCSE boundaries. Subjects that were assessed entirely by exam*

Note. n = 53, 79, and 52 for A-, C-, & F- grades respectively.

Boundaries were again plotted for subjects that were not assessed entirely by exam (Figure 14). The results were similar to Figure 9 (all units), with relatively rapid changes occurring over the first 3 years, but no significant changes occurred between 2013 and 2015.

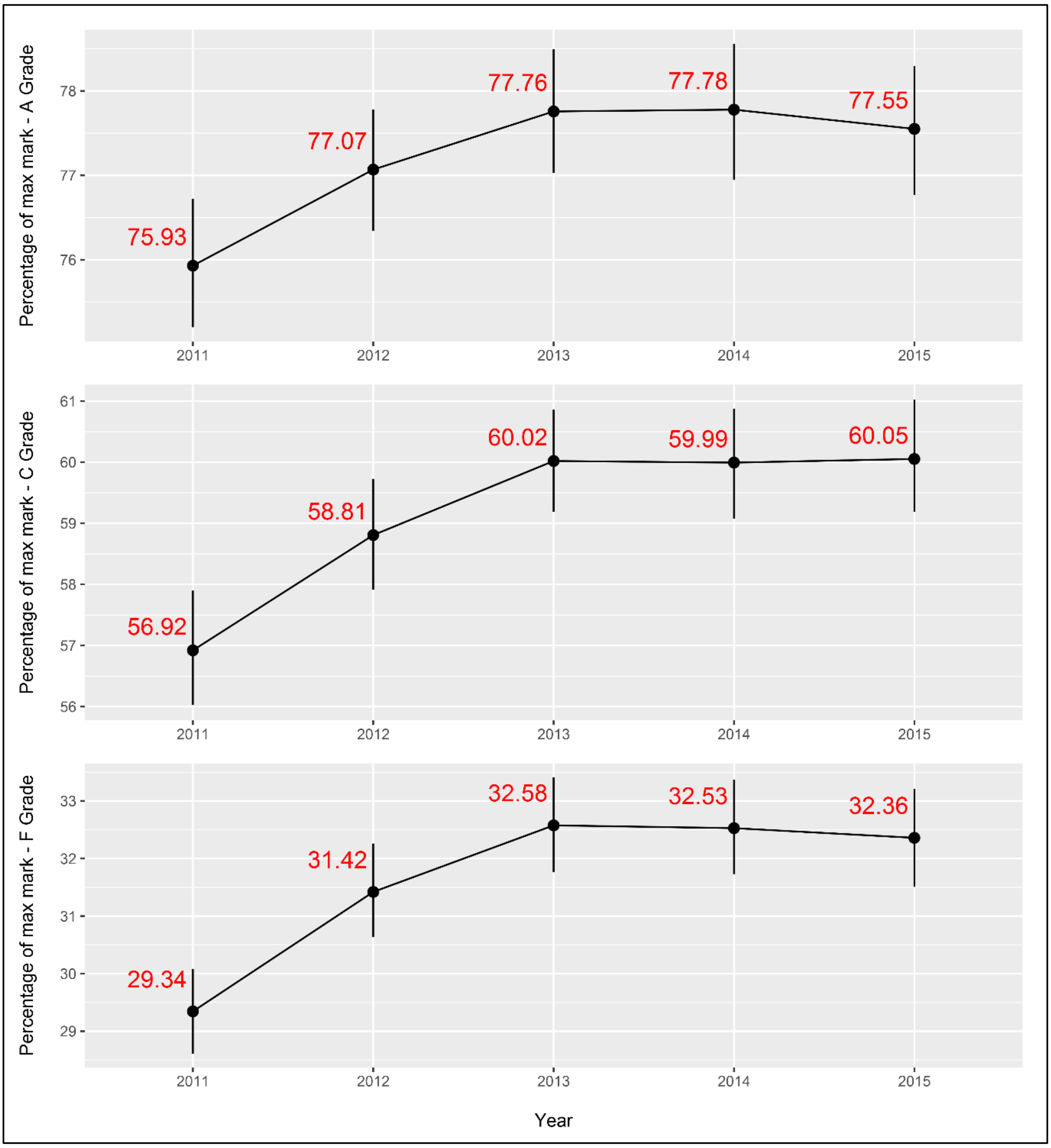

Figure 14. *GCSE boundaries. Subjects that were* ***not*** *assessed entirely by exam*

Note. n = 413, 471, and 414 for A-, C-, & F- grades respectively.

It was interesting to note that there was no downward trend observed for these non-examined-only units between 2013 and 2015, as was observed for units from examined only subjects (ie Figure 13). This was investigated further, by splitting non-examined-only subjects into examined and controlled assessment units. These results (Figure 15) again suggest that controlled assessment units plateaued somewhat after 2013, but boundaries for examined units decreased after 2013.

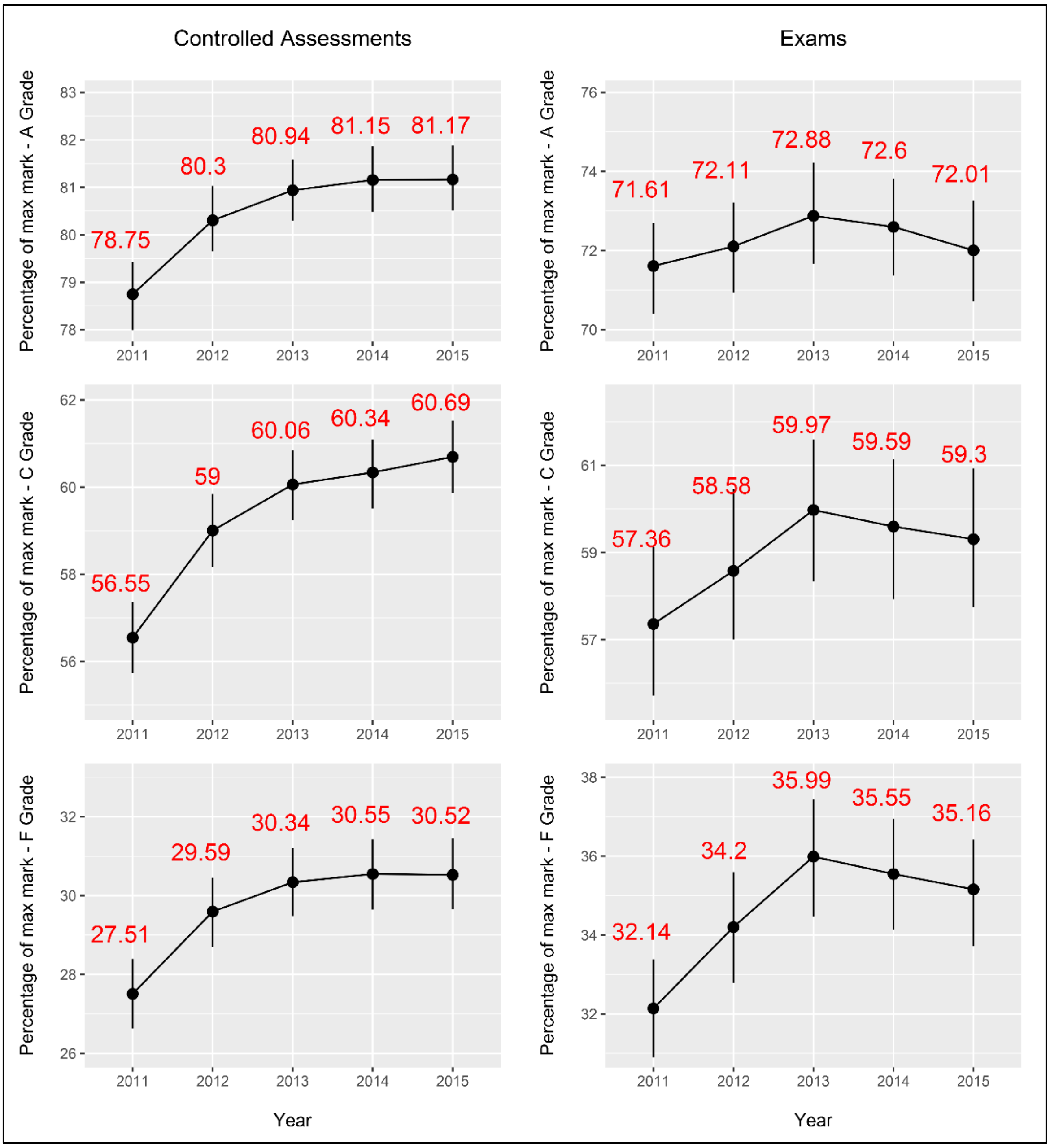

Figure 15. *GCSE boundaries. Units from mixed subjects split by assessment type.*

Note. Controlled Assessment Units: *n* = 250, 254, and 250 for A-, C-, & F- grades respectively. Examined Units: *n* = 163, 217, and 164 for A-, C-, & F- grades respectively.

One possible explanation for decreasing examination boundaries was a change in policy relating to 'early entries'. Before 2013 it was relatively common for students to be entered for a GCSE a year early (ie in Year 10), with the idea that students can 'bank' good grades early on, and resit if they fail to achieved the desired grade in the first sitting (Department for Education, 2015a). Research has shown both benefits and disadvantages associated with early entries (eg Department for Education, 2011; Gill, 2014; Rushton, 2013; Vidal Rodeiro & Nádas, 2010, 2012), meaning that students' performance might have been affected in 2014 and 2015 examinations, after the introduction of a new early entry policy in September 2013 (after the June 2013 examination series) (see Department for Education, 2015a). While it is unclear exactly how this might have caused the observed pattern of result, this may offer a potential explanation. However, given that we did not have unit level entry data available to us, we were not able to investigate this further.

3.3.3 Data simulation procedure to approximate changes in performance

The results reported thus far broadly support the hypotheses made in Section 3. Specifically, grade boundaries have increased following the last set of reforms (on average), consistent with the idea that test-specific performance has increased due to increasing familiarity and preparedness with the new tests. The observed trends are also generally consistent with the proposal that changes should be fairly rapid in the first few years, with the degree of changes becoming more gradual after the limits of familiarity have been reached. The current results for GCSEs in particular (somewhat supported by changes in AS / A level A grades) broadly suggest that these rapid increases occurred over approximately 3 years.

It would be useful at this stage to extend the analyses to provide estimates of changes in performance. This would allow us to glean information about the plausible improvements that teachers / students can make in the years following assessment reform. We were able to use the same simulated data employed previously to estimate the changes in percentage outcomes that would have occurred, had the grade boundaries not been set in the same place year-on-year, therefore reflecting underlying changes in test-specific performance. The reader is again referred to Appendix B.1 for details on how the frequency distributions were produced. Appendix B.3 describes how this data was used to estimate underlying changes in performance (via outcomes) over time. As before, it was necessary to assume that average changes at a unit level correlated with average changes at a subject level (because we were comparing unit grade boundaries to subject level outcomes). In reality, there will be a margin of error with this assumption, and so the analyses to follow should not be considered precise. However, this method does allow for an **estimate** of the **approximate** sizes of changes in performance over time.

Table 2 shows the estimated percentages of simulated students that would / would not have received each grade had grade boundaries remained the same between adjacent years (following the method described in Appendix B.3). These results again generally show that, on average, the largest changes occurred during the first 3 years of the new assessments, with changes being more gradual after this time. As mentioned previously, the differences in AS / A level boundaries between 2010 and 2011, and 2013 and 2014 might be considered smaller than would otherwise be expected (which are reflected in the estimated outcomes). Although the rate of change in AS / A level E-grade boundaries was the largest across the sample, the changes in estimated outcomes were amongst the smallest. This is because the mark distributions were negatively skewed, meaning that fewer students were found at the lower end of the mark range. Thus, a relatively large change in boundaries at the lower end (ie E-grade boundaries) affected relatively few students, compared to changes at the upper end of the mark range (see Figure 37 in Appendix B.1).

Table 2. *Estimated changes in outcomes (the frequency of students achieving each grade) between years associated with changes in average grade boundaries.*

	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015
AS / A level					
A-grade	1.31%	2.87%	1.45%	0.40%	1.17%
E-grade	0.36%	0.65%	0.38%	0.38%	0.40%
GCSE					
A-grade	-	2.88%	1.48%	-0.02%	-0.77%
C-grade	-	4.23%	2.31%	-0.44%	-0.18%
F-grade	-	0.97%	0.47%	-0.08%	-0.17%

Note. Values indicate the estimated percentage of simulated students that would have achieved each grade in the latter year, had grade boundaries not been set in the latter year to ensure comparable outcomes with the former year.

The reader is reminded that these values are representative of **average** changes in performance, and that the size of these changes did vary across individual subjects / units. This is demonstrated in Figure 16. Indeed, grade boundaries for a number of subjects also decreased over time, suggesting that performance may have also decreased in those subjects (although it is also possible that grade boundaries for these units / subjects may have decreased due to changes in question paper demands), therefore not exhibiting the expected sawtooth pattern.

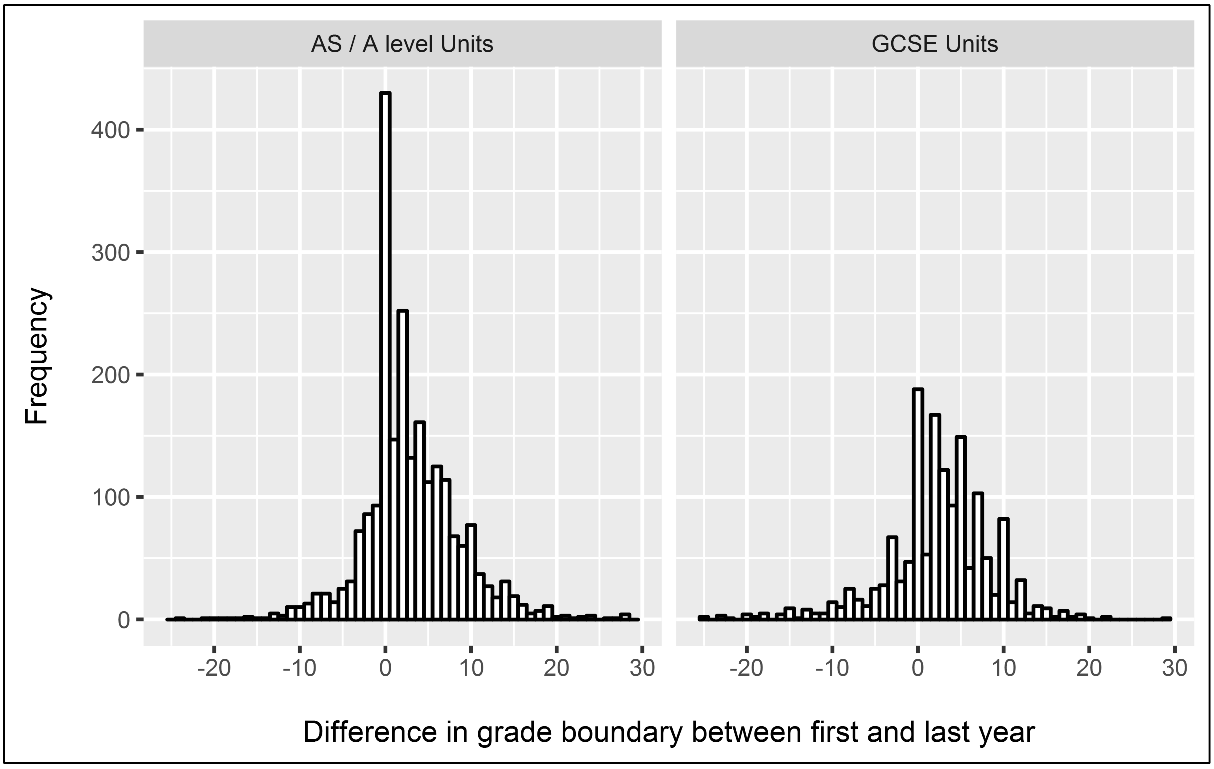

Figure 16. *Histogram to demonstrate that post-reform performance gains do not appear to occur for all assessments, or at the same rate*

Note. Bars have a bin-width of 1%

3.4 Discussion

Overall, the results discussed thus far seem to be consistent with a pattern of post-reform performance gains following the introduction of the last set of GCSE and AS / A level specifications, providing support for the existence of the Sawtooth Effect in the UK education system. Although it was not possible to observe the characteristic 'dip' in performance, improvements in performance (measured via grade boundary changes) in the early years of the new specifications are consistent with the expected pattern (ie that theorised in Figure 1, to the right of the red line).

The fact that most boundaries exhibited relatively rapid increases which seemed to lessen after the third assessment year (with AS / A level E grade boundaries being the exception to this trend) is also consistent with the proposition that post-reform improvements in performance are test-specific and related to student / teacher familiarity (rather than reflecting more general improvements in ability[14]). This is because one would expect improvements to subside once the limits of test familiarity and preparedness are gradually reached. As well as supporting past research into the Sawtooth Effect, the current research expands our understanding of this effect, demonstrating that it applies across a range of subjects at GCSE and A level, and not just English and maths, which past research has exclusively (to our knowledge) focused upon.

Although our finding that the period of relatively rapid improvements lasted approximately 3 years is broadly consistent with previous research (see Koretz et al., 1991, and Cresswell, 2003, whose results suggested post-reform periods of around 2 and 4 years respectively), it should be noted that during these first 3 years there were other testing series available to students (eg January series). It is therefore possible that changes in performance were somewhat accelerated during this time, because there were more opportunities to become familiar with the tests. This may mean that the Sawtooth Effect could occur at a slower rate during future reforms (relative to the changes reported here), when fewer assessment series are available (ie when assessment opportunities are only available once a year).

AS / A level boundaries did not increase as much as was expected between 2010 and 2011, which perhaps warrants further discussion. We propose 2 possibilities that may account for this. Firstly, it is possible that the relative demands of AS / A level assessments meant that it took longer for students and teachers to adapt to the new specifications. By the third assessment year they might have had sufficient time to become more prepared, and exhibited the rapid gains that this allowed. Secondly, it might be that the introduction of the new GCSEs in 2011 took precedence in terms of teacher focus over AS / A levels, which meant that AS / A level performance did not increase as much as it could have in that year. This is made more likely when one considers that schools are held particularly accountable for their GCSE outcomes, which is less true for AS / A levels (the GCSE accountability policy for the period of interest is described by the Department for Education, 2010, paragraph 6.26). However, it is difficult to know which explanation is correct at this time.

It was also interesting to note that E-grade AS / A level boundaries changed more dramatically over time than any of the other boundaries investigated here. Although the rate of change in E-grades did seem to lessen somewhat after 2012, the rate of change in later years was still relatively large. It is possible that lower achieving students may be affected more by reforms (causing a larger 'dip', which might also be more severe at AS / A level due to the added demands) and so have the potential to make larger gains across cohorts in order to return to previous levels of

[14] Therefore supporting the use of the comparable outcomes approach to account for test-specific gains during this time.

performance. It is also possible that lower achieving students could be more reliant upon test familiarity, again meaning that larger performance gains can be made as this familiarity is gained, compared to higher achieving students who may rely less on test familiarity. We have not tested this, and further research is needed to test these possibilities.

Although these analyses of grade boundaries have allowed for an exploration of the Sawtooth Effect, and the results do seem to be generally consistent with its existence, the conclusions drawn are admittedly limited by a reliance on the assumption that grade boundary changes accurately reflect underlying changes in average cohort performance over time. However, there are alternative explanations that might account for these patterns of grade boundary change. For example, it is possible that the demands of question papers lessened over time, which would have meant that grade boundaries rose in the manner observed to account for the rising outcomes caused by these changes (according to the comparable outcomes approach to awarding). Should this be the case, grade boundaries may have risen without there necessarily being any changes in average cohort performance. A second study was therefore conducted to address this limitation of Study 1, in which we assessed the validity of the assumed relationship between patterns of grade boundary change and changes in cohort performance over time.

4 Study 2: Comparative judgements of post-reform performance change

4.1 Design

A comparative judgement methodology was employed to support (or refute) the proposal that performance changed in the manner suggested by the changes in grade boundaries. The specifics of comparative judgment methods have been dealt with in greater detail elsewhere (eg Bramley, 2007; Pollitt, 2012) but, in brief, they make use of expert judgements to produce a rank order of a number of items according to an attribute of interest. For the purposes of the current study, subject experts can be asked to rank order a number of scripts according to the quality of work demonstrated by the candidates. This process can be repeated for different sets of scripts, and once several judgements have been made, Rasch modelling (a statistical technique) can be used to produce estimates of the 'quality' of each individual script, relative to the other scripts in the dataset. This method allows one to deduce the rank order of a large number of scripts, without requiring the judges to review and place into rank order all the materials in one go (this would create significant cognitive load). The specific approach taken in the current study was a 'Thurstone Triples' design (as previously used by Black, 2008), whereby judges are asked to rank order sets of 3 scripts from best to worst. The advantage of this design for the current study was that it allowed us to directly compare performance across 3 testing series (before, after, and at the point at which test familiarity appears to have been reached for each subject), as well as allowing us to gain an appropriate balance between cognitive load, quality of judgment, and the number of comparisons per script. By modelling the subject experts' judgements, we can assess whether performance (as perceived by these experts) has changed in the manner suggested by grade boundaries during this time.

One advantage of using relative judgment over absolute judgment (eg on a 1 to 100 scale) is that the impact on the results of any individual differences between judge leniency or severity is removed. For example, should one judge consistently makes more severe judgements than another, each judge should nevertheless agree on a rank order between scripts, even if they might disagree over their absolute level of quality. As a result, these individual differences are ultimately removed from the analyses (Andrich, 1978; Bramley, 2007). A further benefit of comparative judgement methods is that experts are able to make more reliable and accurate judgements when comparing one item against another (eg "which script is better?") compared to when making absolute judgments (eg "how good is this script?") (Gill & Bramley, 2008; Thurstone, 1927). This benefit is especially advantageous for the current design; given that we were expecting the differences in quality between series to be quite small.

Although we were expecting small degrees of change, and therefore expected that this task would be quite difficult for our judges, previous research has suggested that experts in comparative judgement exercises are able to detect relatively small changes in quality (eg Benton & Elliott, 2016). As the grade boundaries for the subjects included in the current study differed by 5.6% – 11.3% of total raw marks (presented in the following section), we hoped that our judges would have been able to detect such changes, should the grade boundary changes indeed reflect an increase in student performance.

4.2 Method

4.2.1 Materials

As it would be unfeasible to examine the performance exhibited in each of the many units examined in Study 1, it was first necessary to identify a small sample of units that would be appropriate for this exercise. Five criteria drove this search: Firstly, to narrow our scope, we decided to focus only upon GCSE assessments in this study. Secondly, controlled assessment units were not included, as this would have increased the burden on our judges (due to the length of these projects) and would have made cross-unit comparisons more difficult (due to the large differences in the style of the assessment). Thirdly, we wished to include all examined units from within a subject, rather than a mixture of unrelated units, so that we could examine performance change at a subject level. To identify patterns of interest at this subject level, (examined) unit boundaries were averaged to produce an approximate subject-level grade boundary for each assessment series[15]. Patterns were sought at this averaged subject level, rather than for individual units. Fourthly, although we did not include units from subjects that were not present throughout the whole period of interest in Study 1 (to avoid the biasing influence of late starting or early ending units on averaged patterns), the same exclusion criterion was not necessary for Study 2 because we were purely interested in the relationship between grade boundaries and judgements of performance, rather than establishing patterns of performance change. We therefore decided to focus upon popular, large entry subjects in the current study, which were not necessarily included in Study 1. Finally, we deliberately chose subjects that exhibited relatively large changes over time, to allow our judges a greater possibility of being able to detect any changes in cohort performance that may have occurred.

A small number of subjects that either broadly matched or did not match the sawtooth pattern were identified, and can be seen in Figure 17. To increase coverage, we decided to focus upon performance at both A- and C-grade boundaries for included units, although 2 of the patterns identified (Exam Board [EB] 1 maths and EB2 geography) were foundation tier entry routes and so A-grade boundaries were not available[16]. The years (ie assessment series) investigated for each subject depended upon the specification lifespan. The final specifications selected (and the years investigated) were EB1 history (2011, 2013, 2014), EB1 maths (2012, 2013, 2015), EB3 maths (2012, 2013, 2015), EB2 religious studies (2011, 2013, 2015), and EB2 geography (2011, 2013, 2015).

[15] In actual awarding, the process of determining subject-level boundaries is more complex (eg the conversion to Uniform Mark Scales [UMS]). However, calculating simple averages was sufficient for the purposes of sampling for this study.

[16] For foundation tier assessments, the maximum grade available is a C-grade, giving a grade range of G-C.

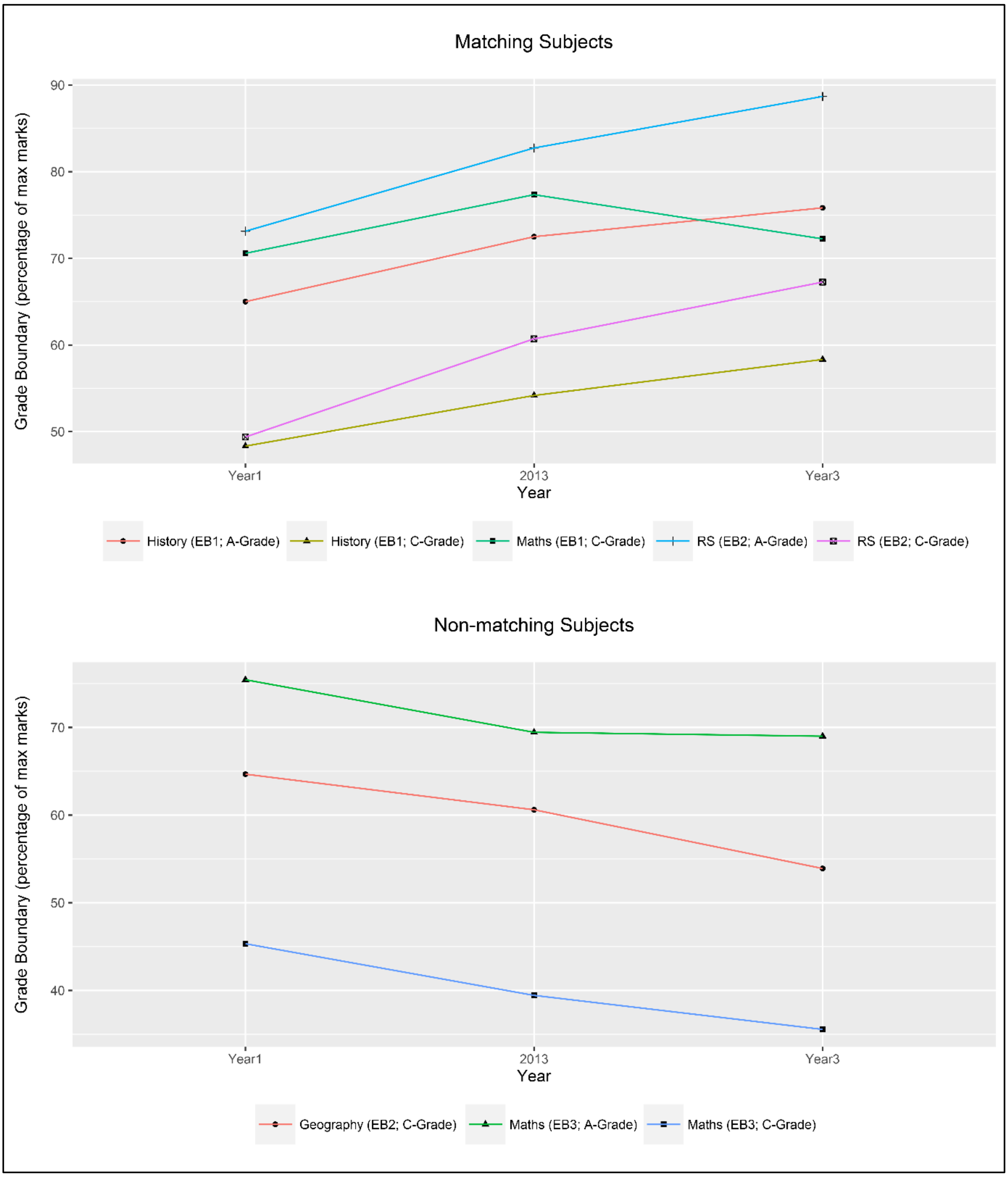

Figure 17. *Subjects identified that matched or did not match the sawtooth pattern*

Notes. Data points indicate the mean grade boundary position, as an average of the boundaries for all externally assessed units of that subject. "Year 1", and "Year 3" were dependent upon the specification lifespan of each subject. We examined religious studies and geography boundaries for 2011, 2013, and 2015. The new maths specifications were introduced a year late, so 2012 boundaries were investigated instead of 2011. History specifications changed after the 2014 testing series and so 2014 boundaries were investigated instead of 2015.

Once these subjects had been chosen, the relevant archive scripts held by the exam boards were obtained. Five scripts per grade per assessment series were provided by the relevant exam boards (the typical total number of scripts that they hold in their archives). Helpfully, scripts selected for archiving are those that are deemed to be representative of the grade boundary mark for that assessment series, allowing us to examine the performance exhibited at each of the data points plotted in Figure 17. To avoid influencing the participants' judgments, each script was 'cleaned' of all markers' annotations, as well as any information that identified the year, the exam board, or the candidate. So that we could identify each script, and so that the judges could relay their judgements back to us, a cover page was added with a clearly printed ID number[17].

After the scripts had been prepared they were collated into packs of 3 for the examiners to judge. In the interest of brevity, the pack design shall only be described briefly here; full details can be found in Appendix C. Each script was included within 15 packs in total, to give enough comparisons per script for Rasch analyses. The number of packs that each examiner judged depended on their subject area, according to the number of scripts under investigation. As we wished to produce estimates of performance change at an aggregated subject level, rather than for each unit in isolation, it was necessary to combine scripts from different units of the same subject within some of the packs. We did this in order to ensure that Rasch analyses would allow us to compare estimates of quality between different units. To make the task easier for our judges, packs were presented to them in order, according to the type of comparison that we were asking them to make. Easier comparisons were presented first (eg scripts from 3 years of the same unit), and more cognitively demanding comparisons (eg scripts from 3 different years and 3 different units) were presented once the judges had become more familiar with the task. Scripts were presented in a random order within each pack.

4.2.2 Participants

Each of the 3 exam boards that provided scripts to us also nominated 2 suitably experienced judges (chief and / or principal examiners) for each of the subjects identified in Figure 17. A total of 24 judges from 4 subject areas (6 per subject) were recruited and each examiner completed judgements from all of the grade boundaries in their subject area. For example, geography examiners only judged 1 set of scripts

[17] This ID number was generated to include the first letter of the exam board (substituted for numbers in the current report), the first letter of the subject, the grade boundary ('A' or 'C'), the unit number (1-3), the year (eg '1' for 2011), and the candidate number (1-5). For example, 1MA312, would be EB1, maths, A-grade, unit 3, 2011, candidate 2 of 5. Judges were not made aware of this coding scheme.

(C-grade scripts from EB2), whilst the maths examiners judged 3 sets of scripts (C-grade scripts from EB1, and C- and A-grade scripts EB3). Each judge was paid a fee in exchange for making their judgements, according to the amount of work completed.

4.2.3 Procedure

Each judge was posted paper packs to conduct their judgements, which were ordered in the manner outlined previously. Judges were not provided with mark schemes, as we did not want them to re-mark the scripts. Included within their materials was an instruction sheet, which outlined what they were expected to do. In brief, they were asked to work through each pack in the order given to them and to decide upon a rank order of the scripts within each pack of 3 scripts, from best to worst in terms of the quality of work, taking into account how difficult each paper is. This latter instruction was included to control for any subtle differences in the demands of different question papers, meaning that judges were expected to compensate for a worse / better performance when one question paper was harder / easier than another. In order to help them with this requirement, they were first provided with a pack containing blank question papers (the same papers as the scripts) and were asked to rank order these in terms of their difficulty on a cover sheet provided. This exercise was used simply to familiarise the judges with the materials, so that they could form an impression of question paper difficulty before having to compensate for this during the actual script judgments.

After completing the short familiarisation task, judges were asked to judge their packs of scripts. In essence, they were simply asked to decide upon which candidate was the best geographer / historian / mathematician / religious studies student. Judges were encouraged to base their decisions upon holistic judgements of quality, and were discouraged from simply re-marking the scripts. Once they had decided which script was the best, middle, and worst in terms of quality (taking paper difficulty into account), they were asked to note their responses on a cover sheet provided, by writing the script ID number next to each level of response (ie "best", "middle", and "worst"). Ties or equal ratings were not permitted, and judges were encouraged to make a decision, even when they thought 2 scripts might be of equal quality.

Once each judge had completed all their script judgements, and had returned all the materials, they were asked to complete a short questionnaire to provide some feedback on the task, especially regarding how difficult they found it.

4.3 Results

Once all the judgements had been made, each triples comparison was converted into 3 paired comparisons (1 vs. 2; 1 vs. 3; 2 vs. 3 – see Figure 18), required by the

Rasch analysis. The 'SIRT' package (Robitzsch, 2016) for R was used for all Rasch modelling. However, the equations provided by Pollitt (2012) were used to calculate infit statistics, which are not supported by this package.

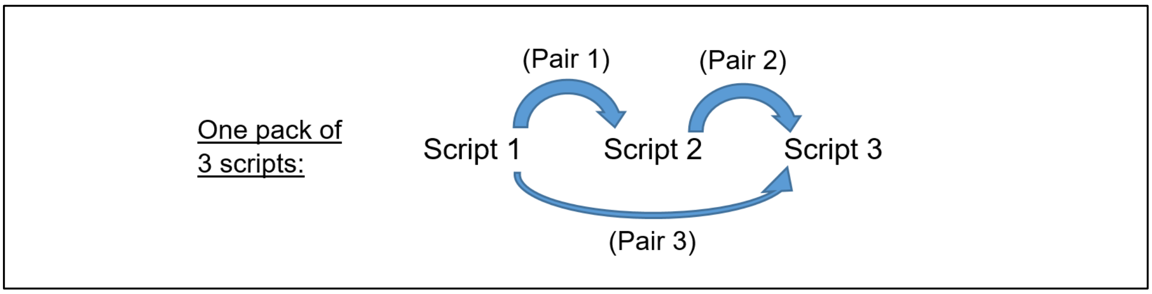

Figure 18. *Converting one triples comparison into 3 paired comparisons*

4.3.1 Religious studies

A-grades

A separation reliability[18] of .84 was achieved for the Rasch analysis, indicating that the model was successfully able to discriminate between scripts of different quality (as perceived by the judges). Infit statistics were also investigated, and items / judges were classified as being misfitting when individual infit scores were greater than the mean plus 2 standard deviations (calculated from all infit scores). This criteria is standard convention (Pollitt, 2012). One item ('2RA111') marginally exceeded this threshold (1.41 vs. 1.33), indicating that judges found this item more difficult to judge reliably than the others. All judge infit scores were within tolerance.

The scatter plot displayed in Figure 19 shows the estimates of script quality, known as 'theta' scores. As discussed by Bramley (2007), it would be unwise to use these results to conduct statistical tests for differences between years, because the measurement error for each of these points would be ignored. As such, Bramley suggested combining the measurement errors of each item within a group (in our case, each script within a year) to produce a measurement error value for the mean of each group (ie the mean of each year). One then can assess any differences between groups by seeing whether the means differ outside of the 95% limits of their measurement error. This method was taken to produce the bottom 2 panels of Figure 20. Using the same method, theta scores were also aggregated across units to produce overall estimates of subject-level performance in each year, which is

[18] This is an equivalent measure to Cronbach's alpha (Wright and Masters, 1982, cited in Pollitt, 2012). Standard convention states that alphas above .7 are usually considered sufficient (Nunnally, 1978).

displayed in the second panel of Figure 20. To allow for quick visual comparisons to be made, the expected pattern (ie the changes in grade boundaries from Figure 17) is displayed in the top panel of Figure 20.

Upon inspection of the error bars (ie whether or not they overlap), the results suggest judges deemed that students, on average, demonstrated a significantly better performance in 2013 and 2015 compared to those in 2011, for Unit 8 and for the subject overall (as expected). Although there were overlapping error bars between 2011 and 2013 cohorts for Unit 1, judges did still rate the performance of the 2015 candidates as being better than 2011 candidates, as was expected.

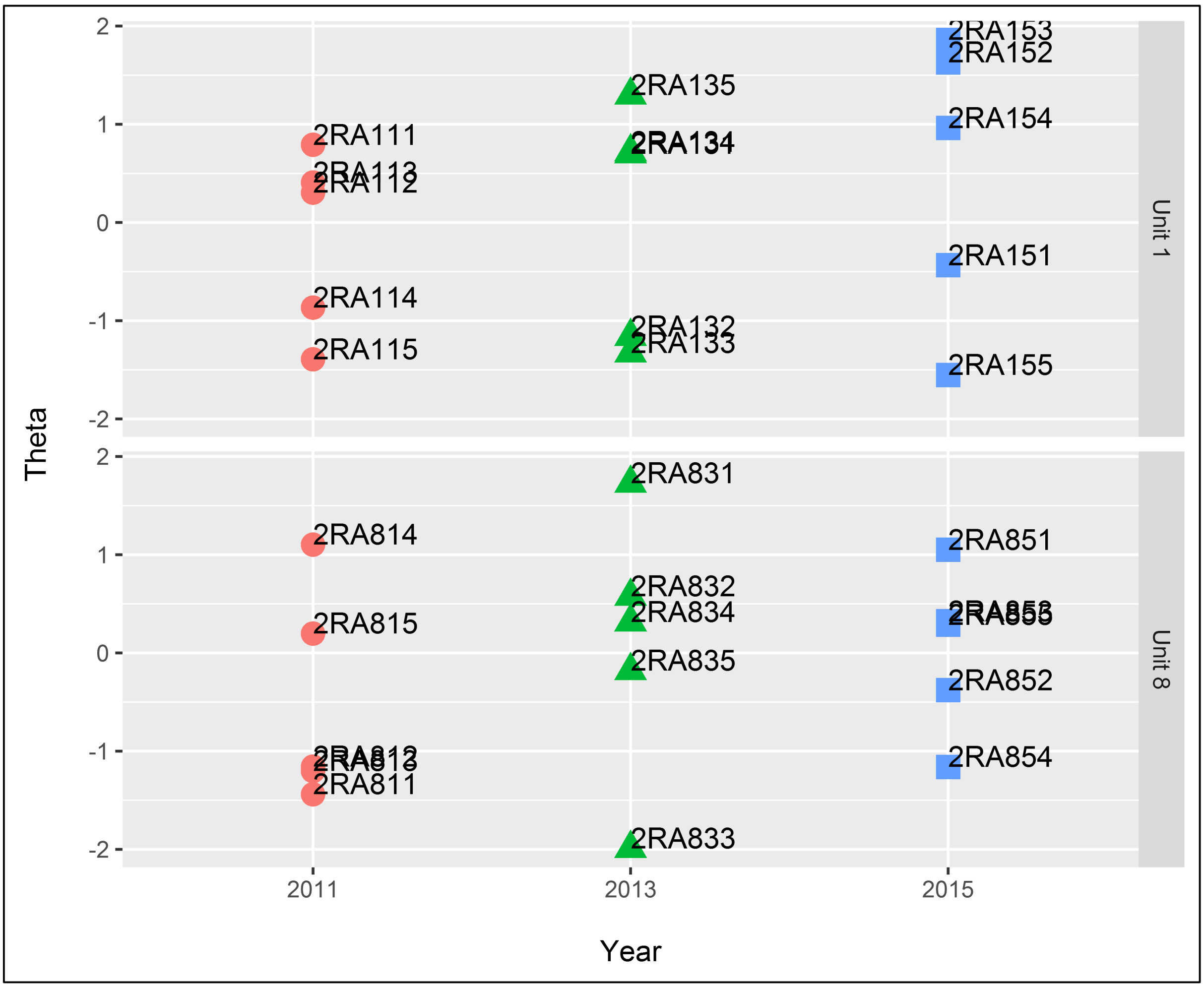

Figure 19. *Scatter plot for religious studies A-grade scripts across 3 series*

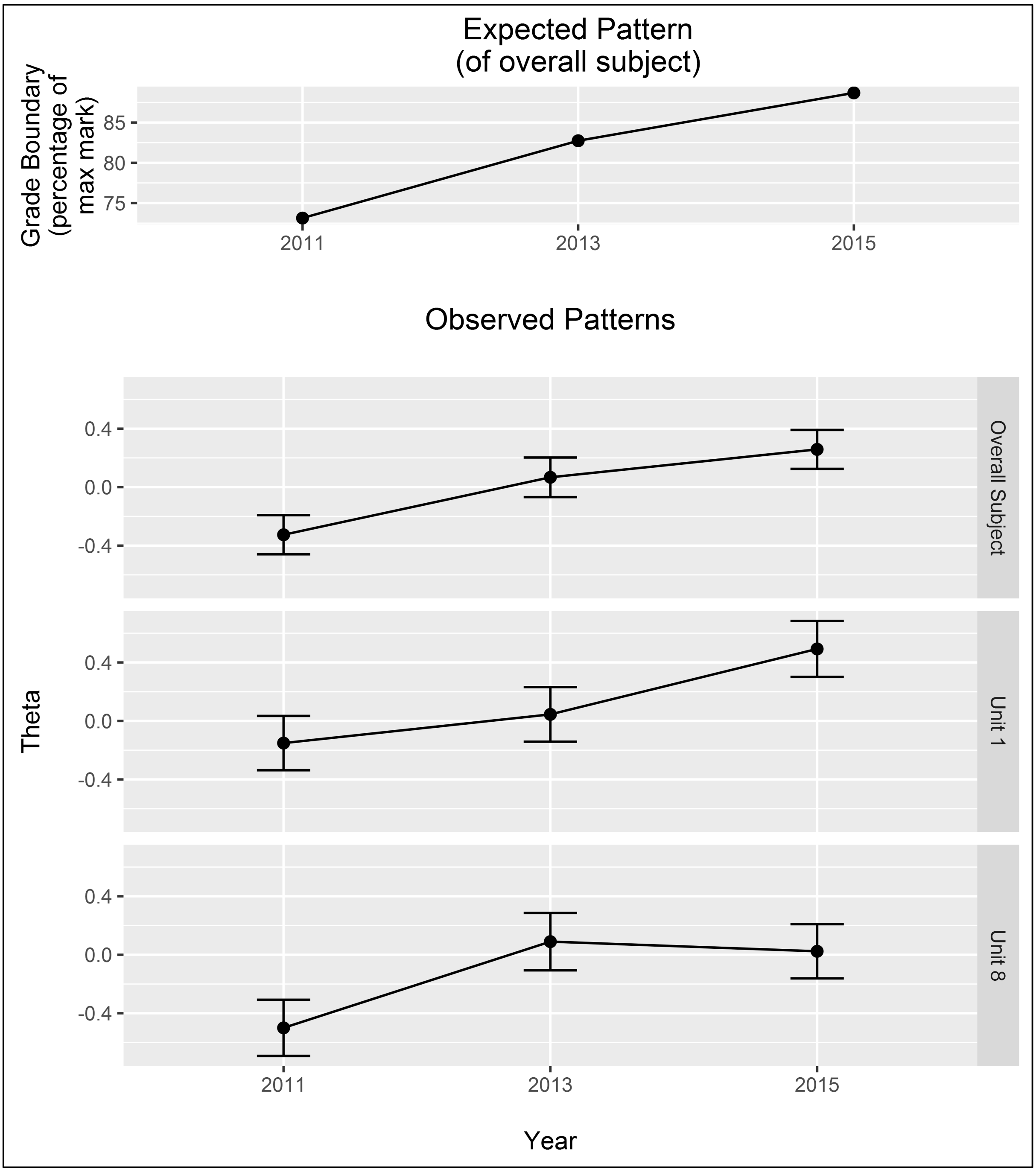

Figure 20. *Expected and observed patterns for religious studies A-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

C-grades

A separation reliability of .78 was achieved for C-grade boundary scripts, which was again within acceptable limits. As with the A-grade scripts, just one script ('2RC813') had an infit value marginally above the tolerance threshold (1.28 vs. 1.24), indicating that judges found it slightly more difficult to judge this script reliably, compared to the others. As before, none of the judges had infit values that exceeded the tolerance threshold.

A scatter graph (Figure 21), and a graph of means / combined errors (Figure 22) were produced. The same conclusions can be drawn from these graphs as from the comparisons of A-grade scripts, that judges perceived a difference in subject-level performance between 2011 and 2013, but no significant change between 2013 and 2015.

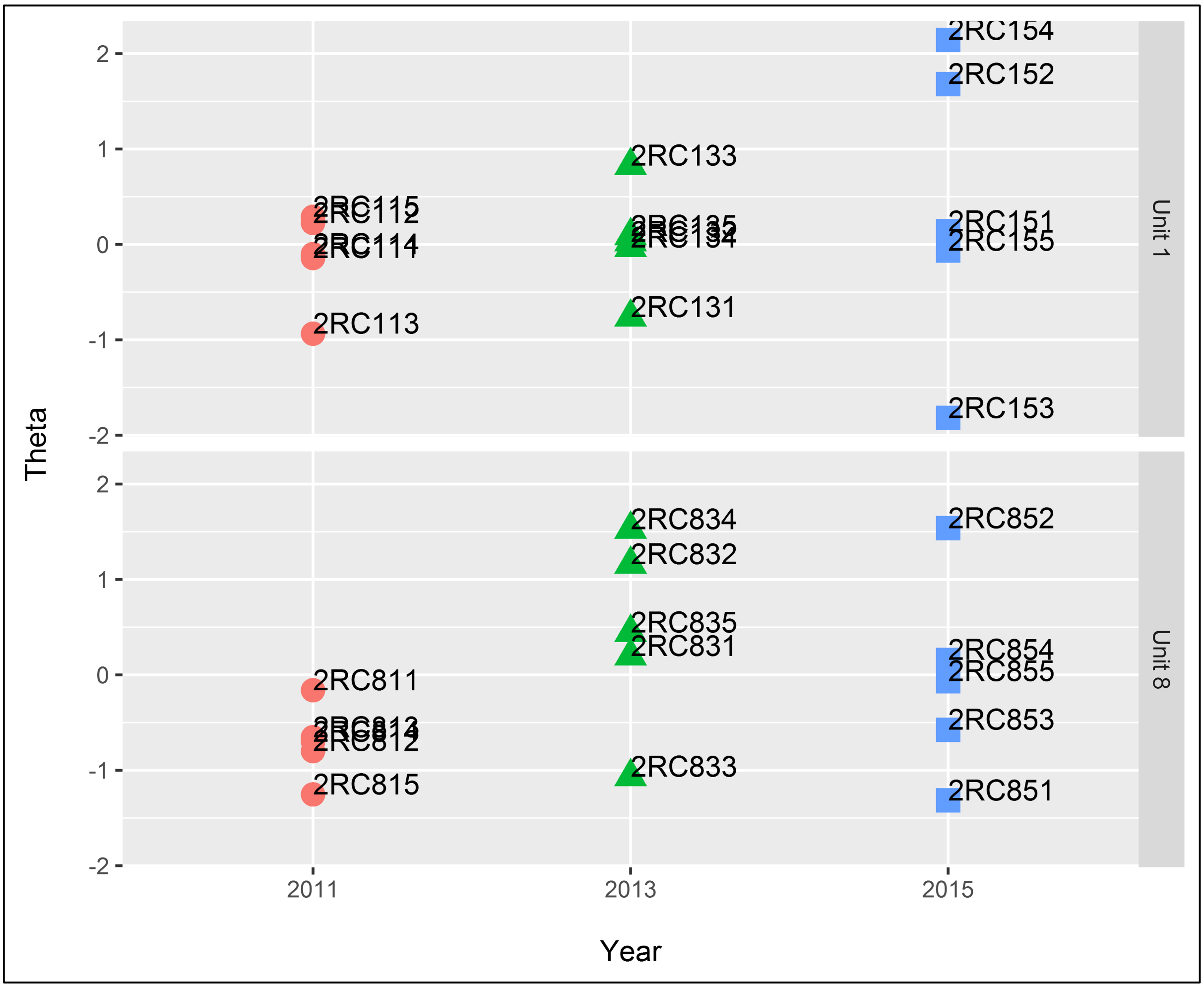

Figure 21. *Scatter plot for religious studies C-grade scripts across 3 series*

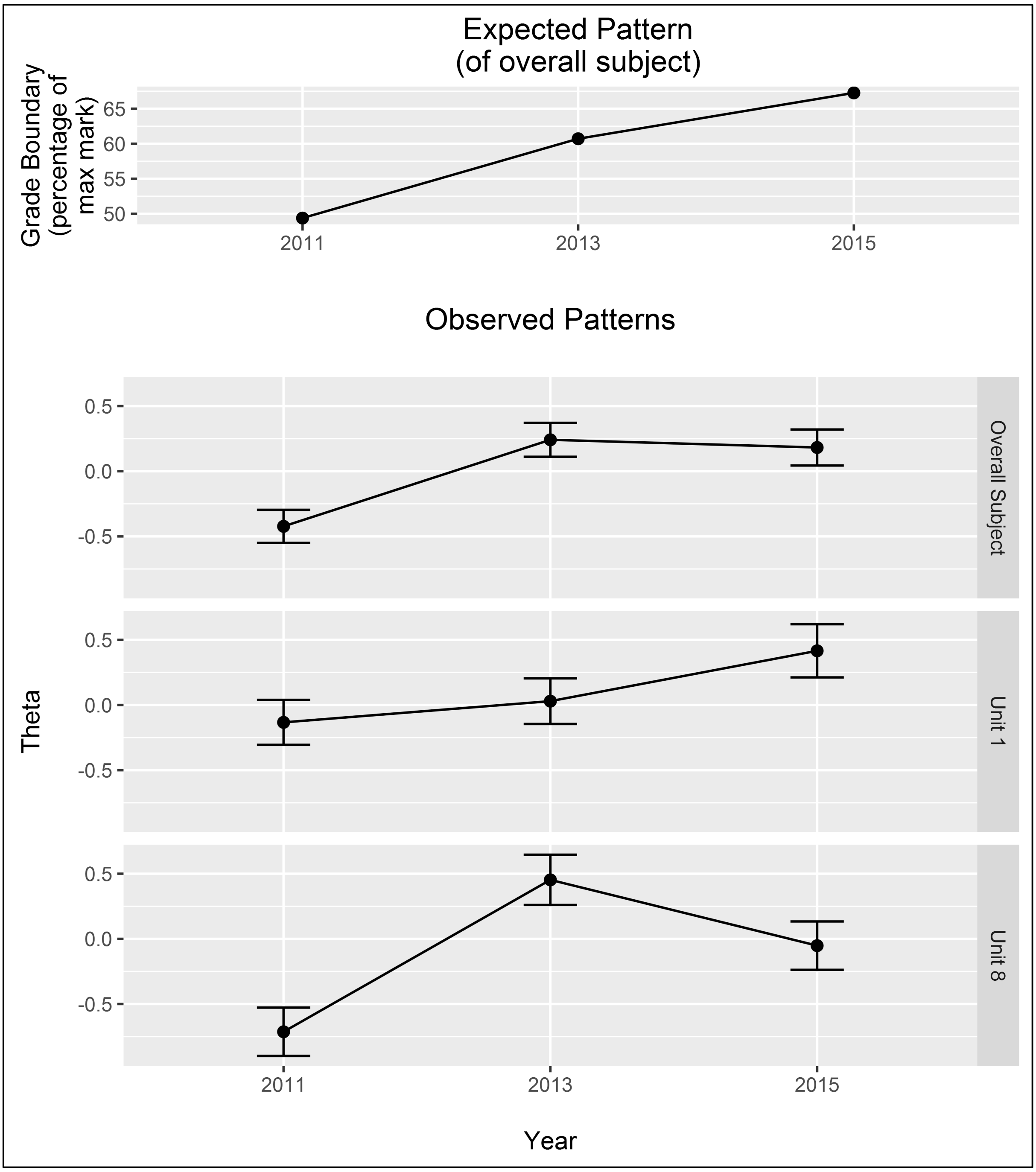

Figure 22. *Expected and observed patterns for religious studies C-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

4.3.2 History

A-grades

A separation reliability of .88 was achieved, suggesting that the model was reliable. An analysis of infit values suggested that judges were equally reliable in their judgements, but one item was judged less reliably than the others, as the infit value for script '1HA211' was above tolerance (1.87 vs. 1.48).

The results, presented in Figures 23 and 24, are somewhat consistent with the expected pattern when aggregated across the whole subject. Judges rated the performance of 2013 and 2014 candidates as being significantly better than 2011 candidates. Performance in 2014 was rated lower than 2013, but this was within the limits of measurement error. The low average rating for 2014 scripts in Unit 2 might be due to a potential outlier (see Figure 23 – script '1HA242'), which might also explain the slight unexpected dip in the overall subject ratings for 2014.

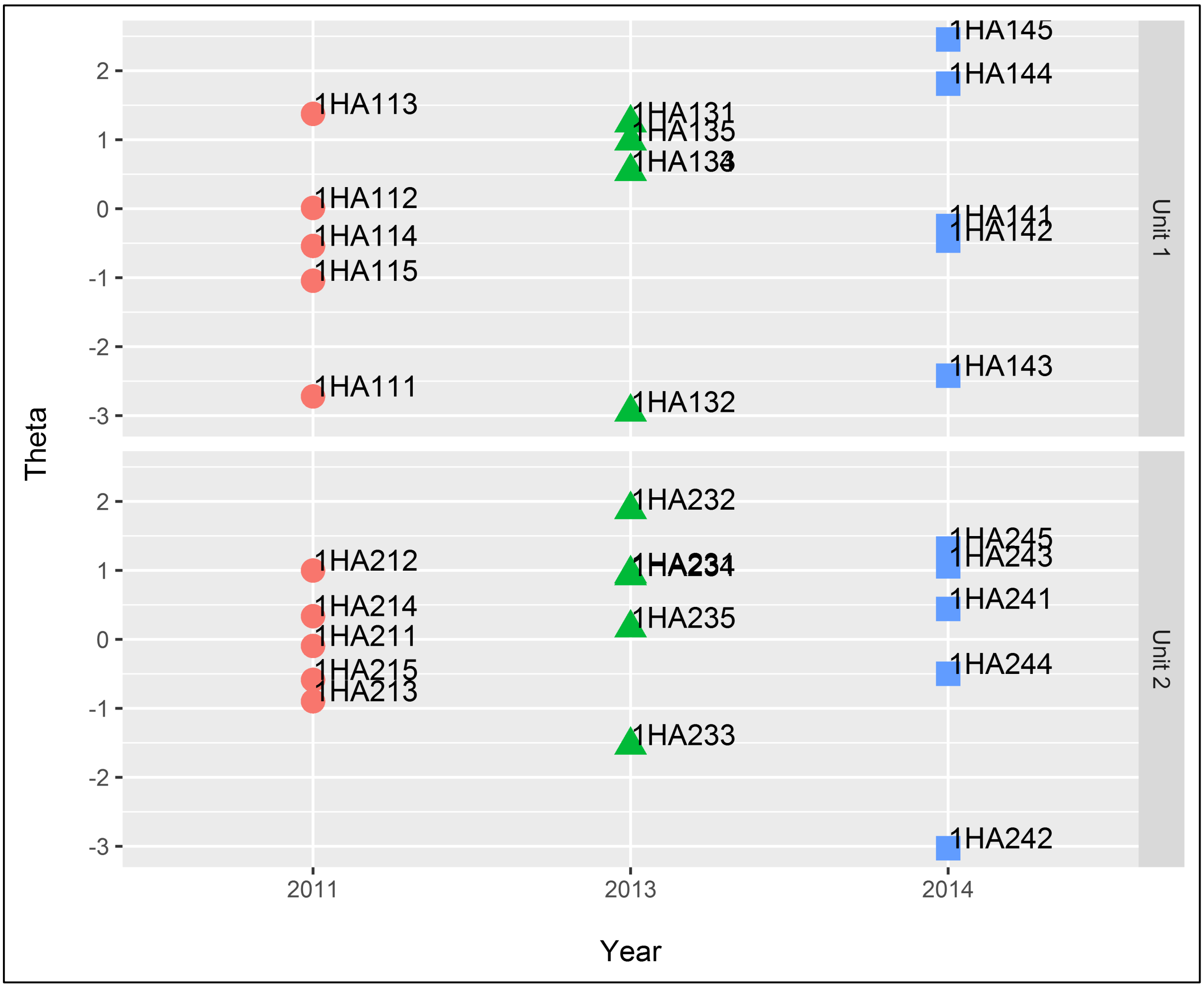

Figure 23. *Scatter plot for history A-grade scripts across 3 series*

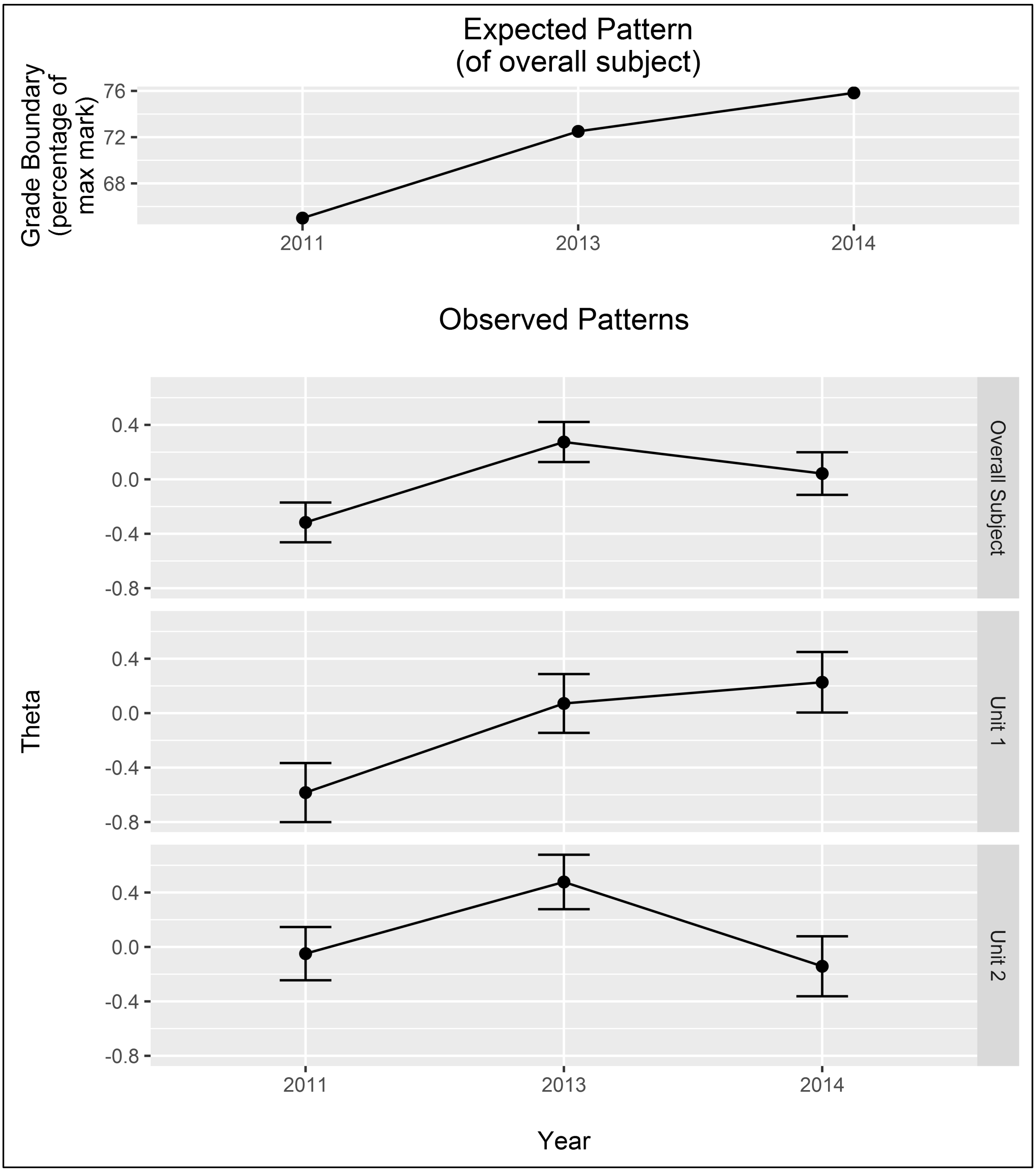

Figure 24. *Expected and observed patterns for history A-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

C-grades

A separation reliability of .88 was again achieved, suggesting that model was reliable. Infit values again supported the reliability of judges, but suggested that one script ('1HC145') was judged less reliably, as its infit score was marginally above tolerance (1.37 vs. 1.35).

Figures 25 and 26 show that, when aggregated, judges mean ratings were significantly higher in each year. Although there appeared to be a decrease in ratings between 2013 and 2014 for Unit 2, this was within the limits of measurement error. The overall subject-level pattern closely matched the expected pattern (ie changes in grade boundaries).

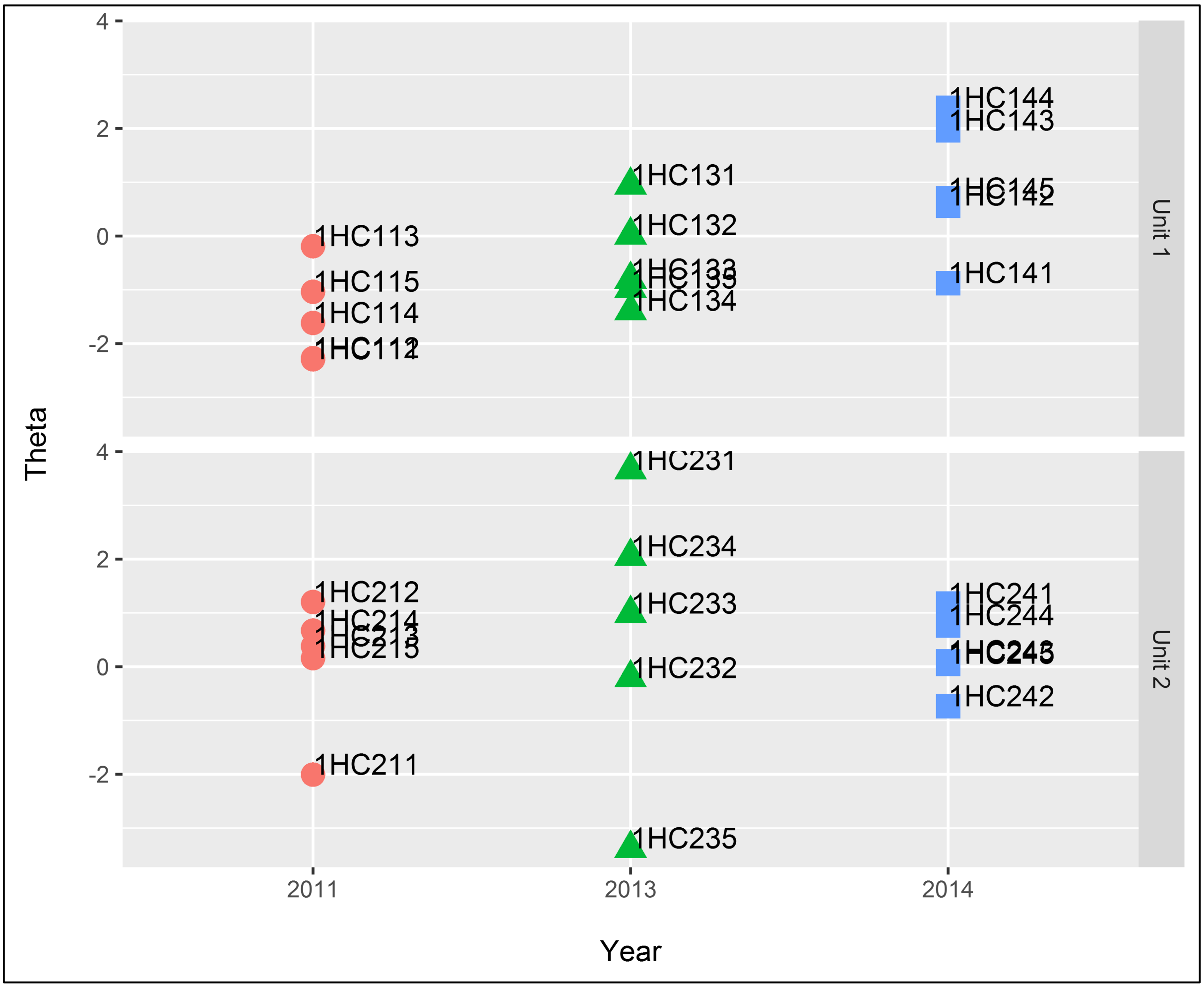

Figure 25. *Expected and observed patterns for history C-grade scripts across 3 series*

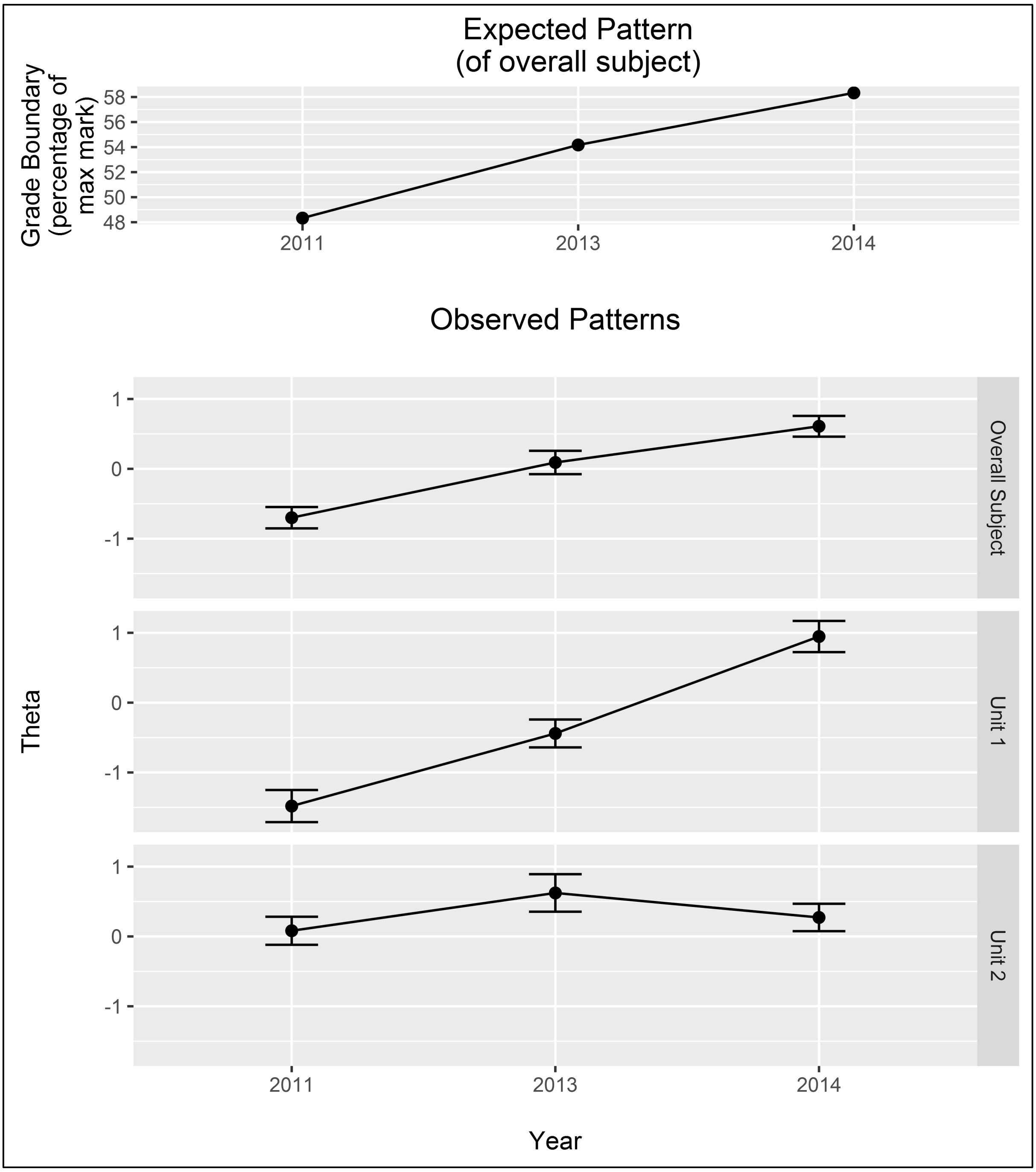

Figure 26. *Expected and observed patterns for history C-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

4.3.3 Maths (EB1)

C-grades

A separation reliability of .82 was achieved, and no infit values (either for items or judges) exceeded the tolerance threshold. Results are plotted in Figures 27 and 28. When aggregated across the whole subject, there were no statistically significant differences between the observed mean performance in 2012 and 2013, but judges mean ratings were significantly higher on average in 2015 compared to 2012. Although judges ratings of performance did not match the large changes in grade boundaries in 2013, there was an overall increase between 2012 and 2015, as expected.

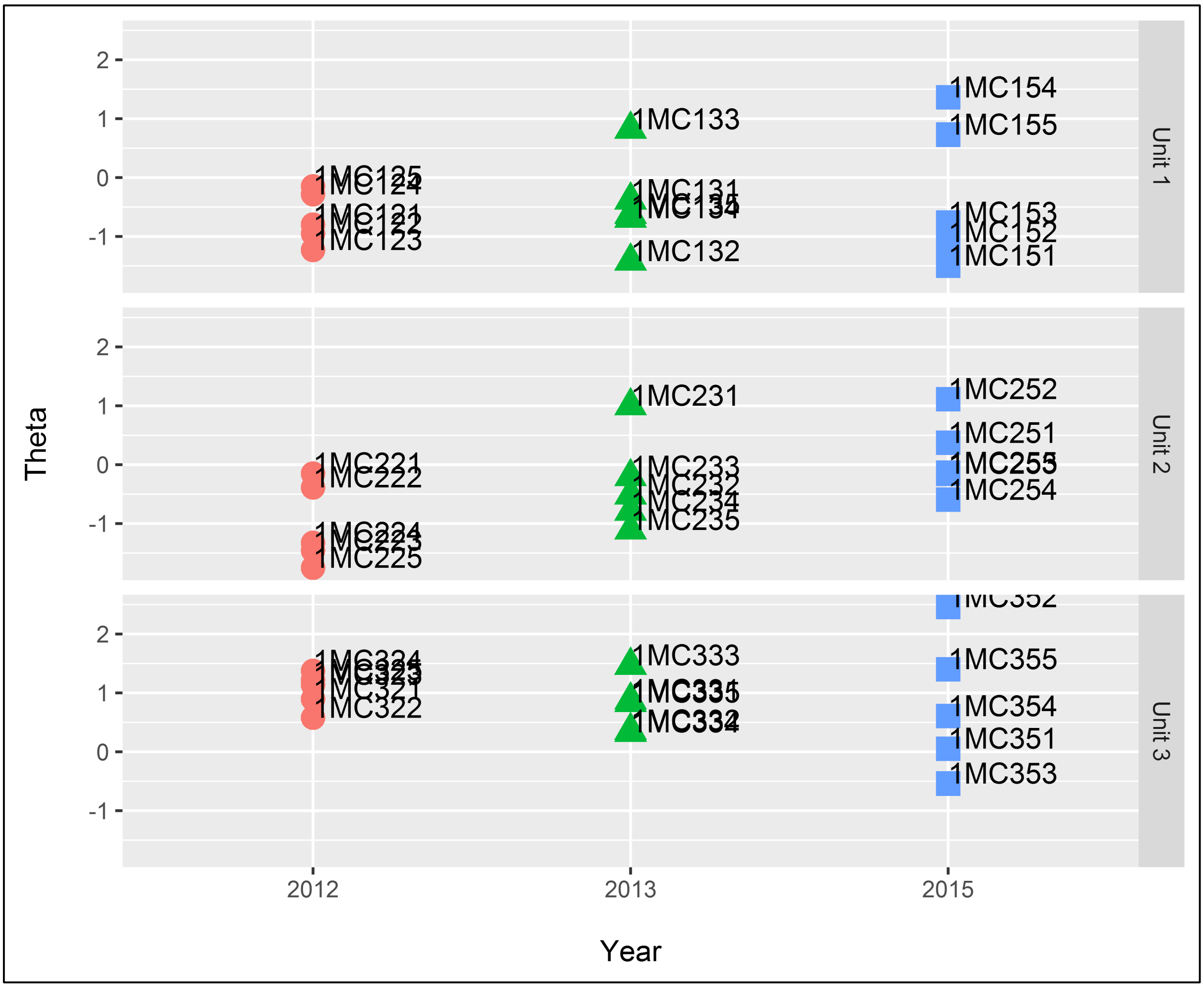

Figure 27. *Expected and observed patterns for maths (EB1) C-grade scripts across 3 series*

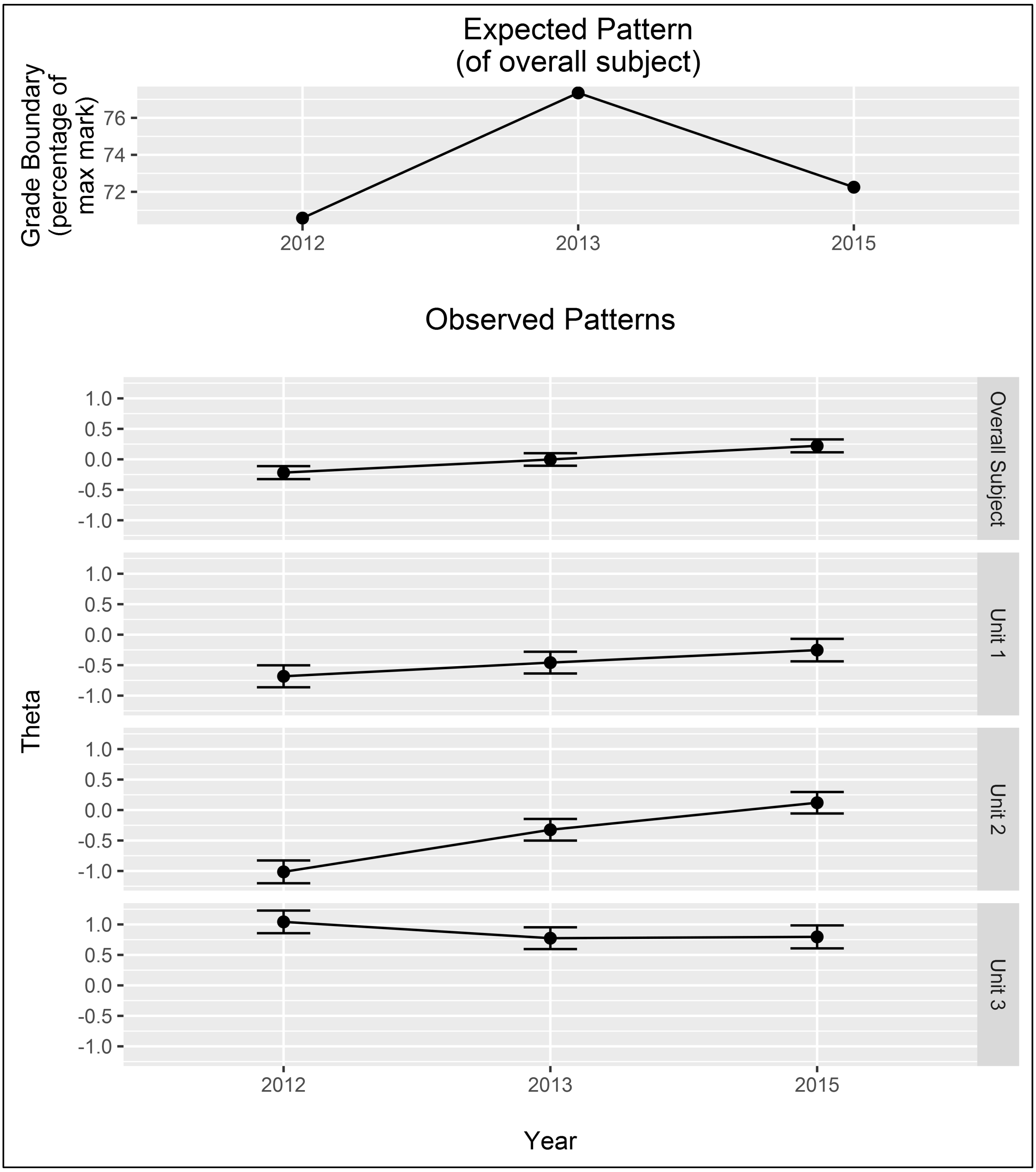

Figure 28. *Expected and observed patterns for maths (EB1) C-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

4.3.4 Maths (EB3)

A-grades

A separation reliability of .91 was achieved, and no infit values (either for items or judges) exceeded the tolerance threshold. Results are plotted in Figures 29 and 30. When aggregated across the whole subject, the difference between the observed mean performance in 2012 and 2013 was within the limits of measurement error, but judges' ratings were higher on average in 2012 compared to 2015, as expected. Although ratings of performance were somewhat higher in 2013 than might have been expected, there was an overall decrease in judges' ratings between the first and last year, which were broadly consistent in direction (though not necessarily in magnitude) with the changes in grade boundaries.

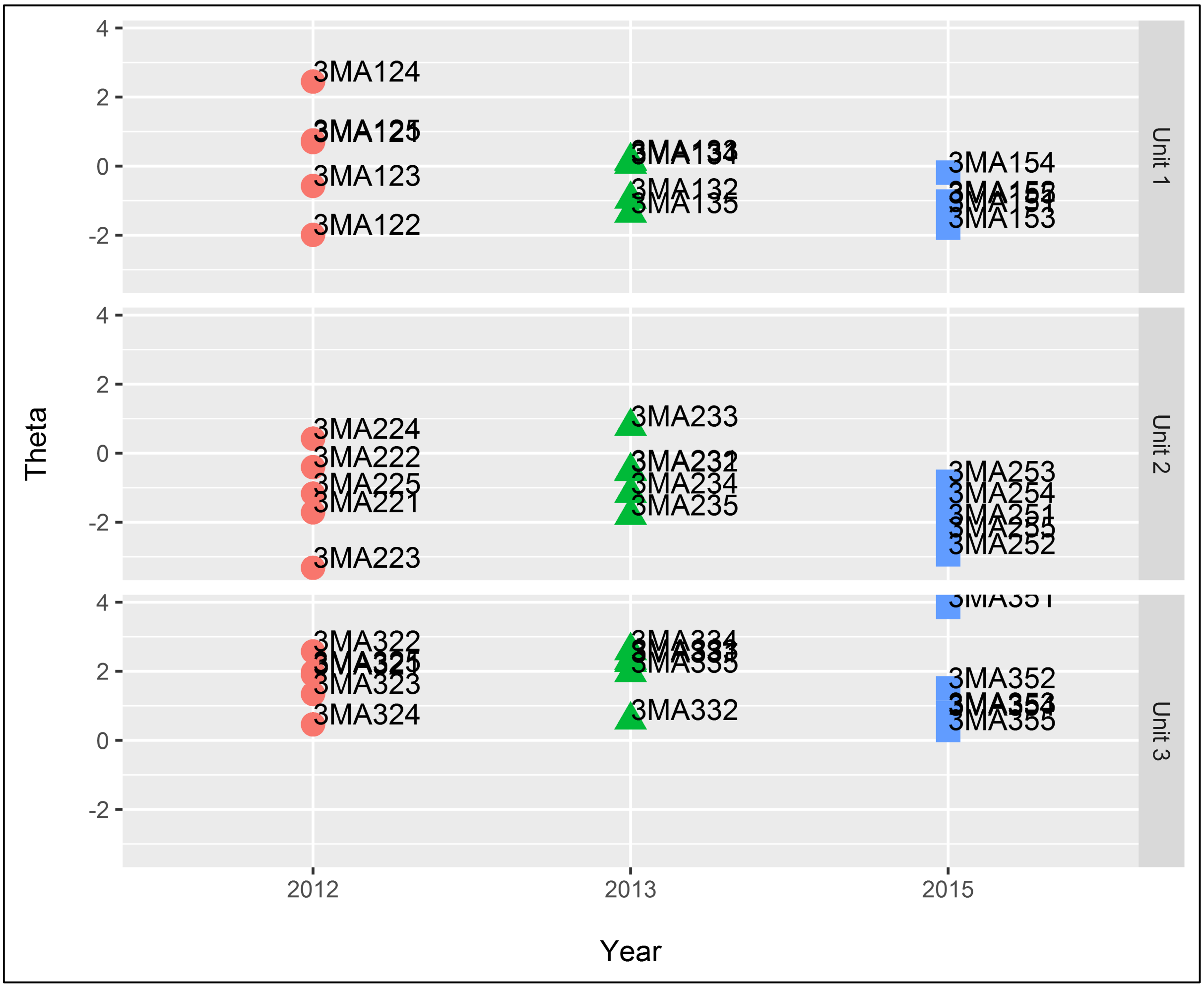

Figure 29. *Expected and observed patterns for maths (EB3) A-grade scripts across 3 series*

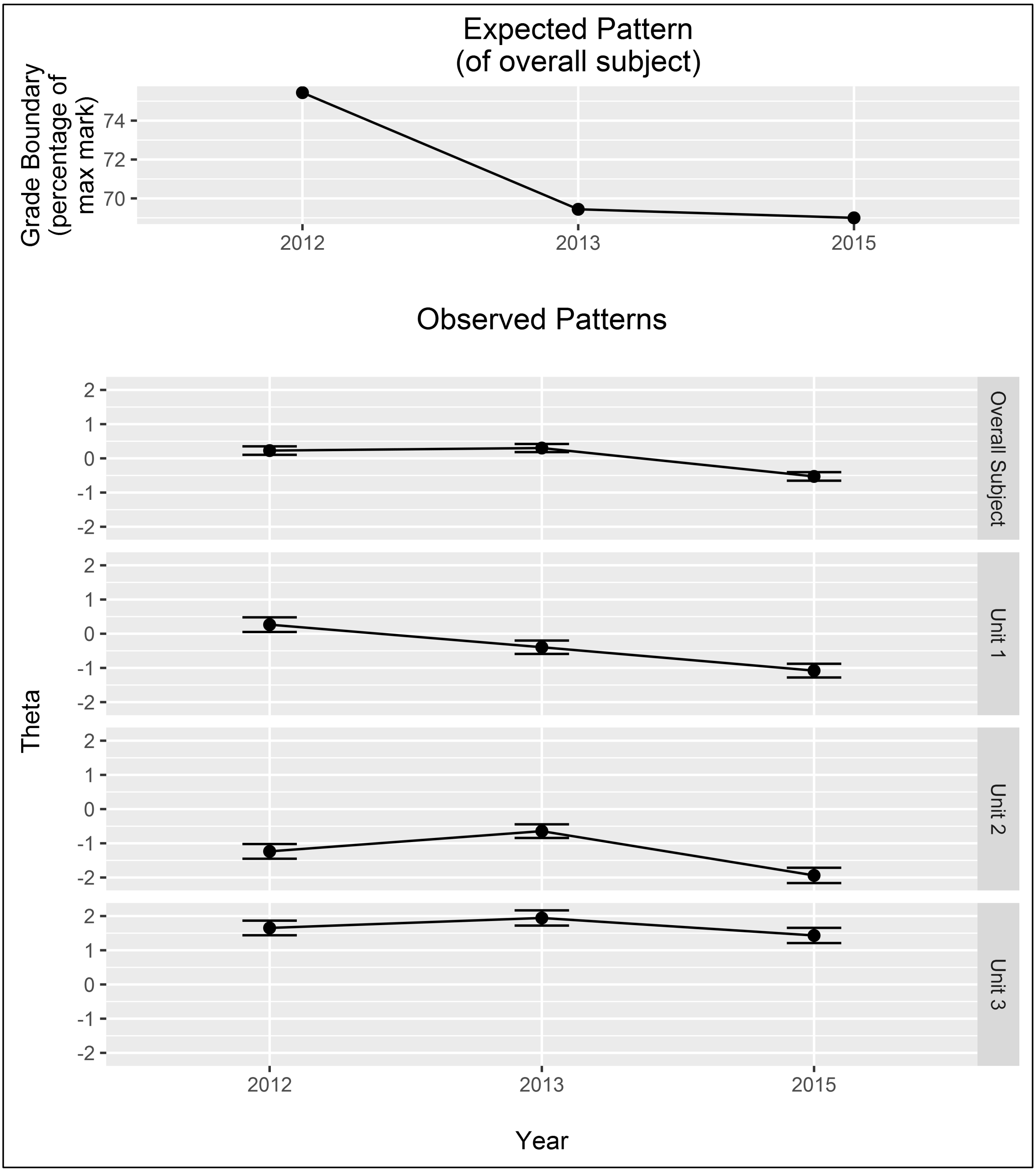

Figure 30. *Expected and observed patterns for maths (EB3) A-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

C-grades

A separation reliability of .91 was again achieved. None of the judges’ infit values exceeded the tolerance threshold. One item (‘3MC135’) did marginally exceed this threshold (1.53 vs. 1.50), suggesting that it was slightly more difficult to judge reliably than the others. As expected, judges’ mean ratings decreased each year, when the theta values were aggregated across the subject (Figures 31 and 32). This seemed to be mostly due to changes occurring to Unit 3, with the other 2 units being comparably more stable over time (as perceived by our judges).

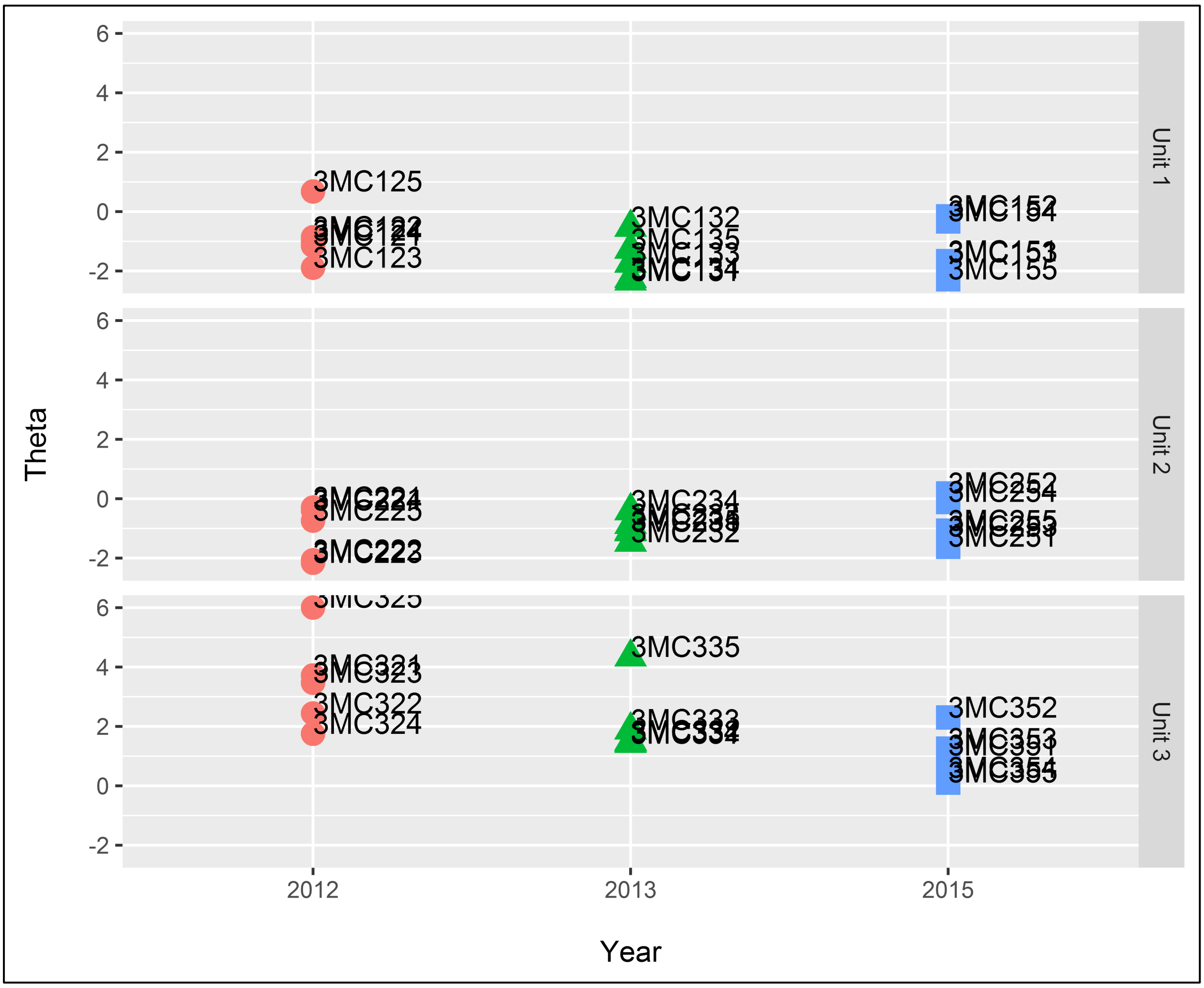

Figure 31. *Expected and observed patterns for maths (EB3) C-grade scripts across 3 series*

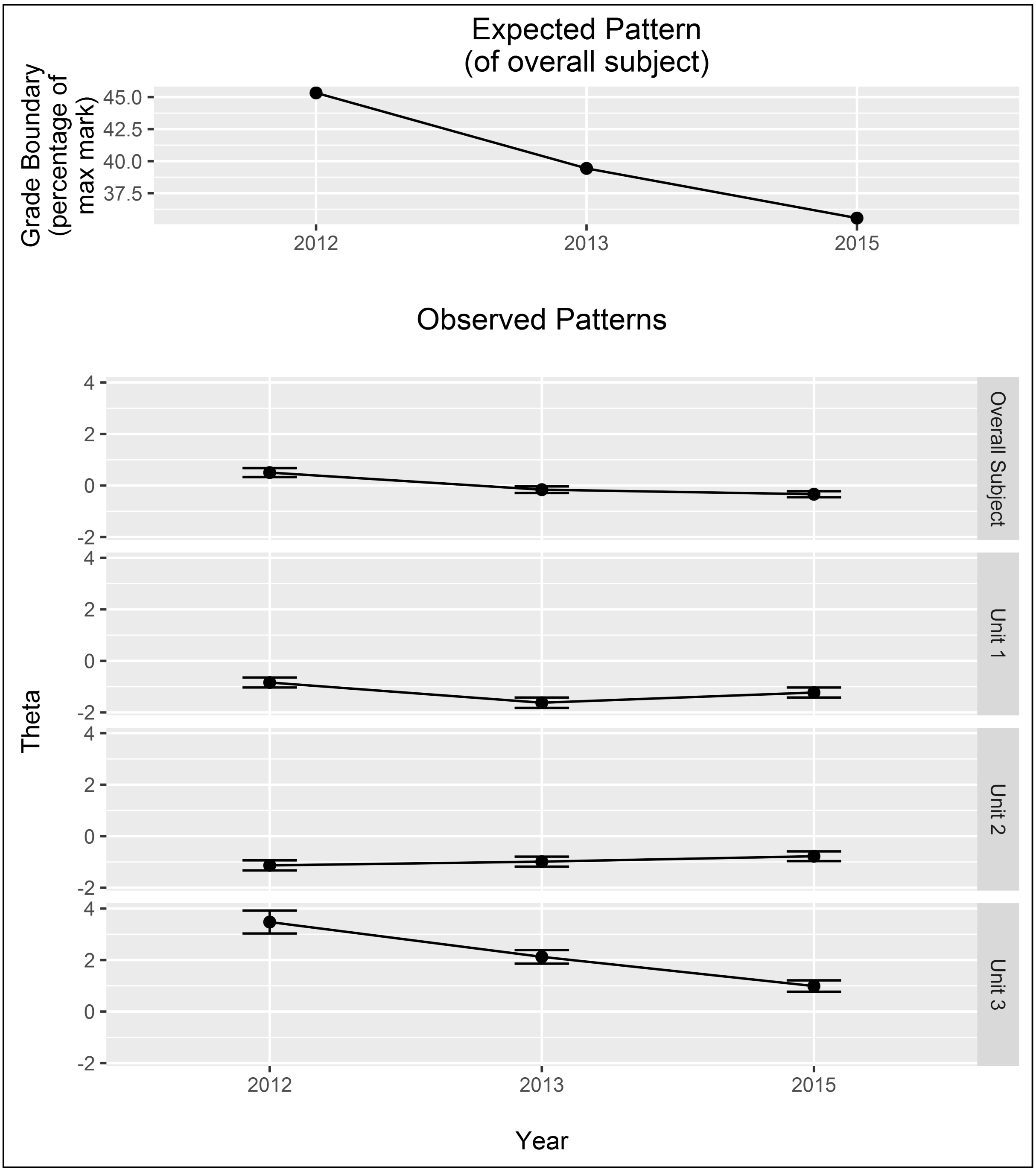

Figure 32. *Expected and observed patterns for maths (EB3) C-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

4.3.5 Geography

C-grades

A separation reliability of .82 was achieved, and no infit values (either for items or judges) exceeded the tolerance threshold. The results (Figures 33 and 34) were unexpected in light of the changes in grade boundaries for this subject (ie the top panel of Figure 34). Although we expected performance to drop over this time, judges perceived the quality of scripts to increase overall. This finding reflects a degree of disconnect between grade boundaries and underlying performance change for this subject.

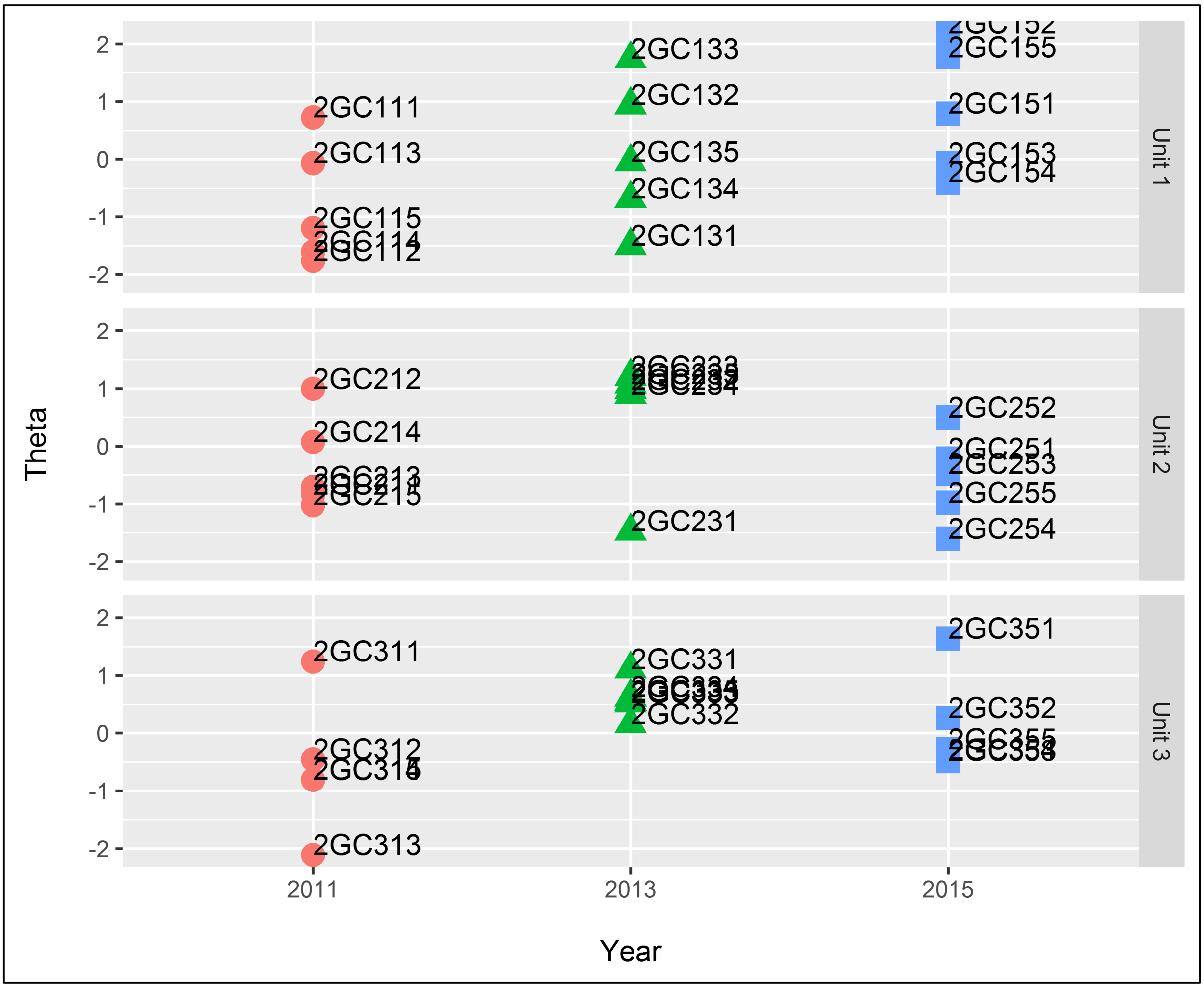

Figure 33. *Expected and observed patterns for geography C-grade scripts across 3 series*

]

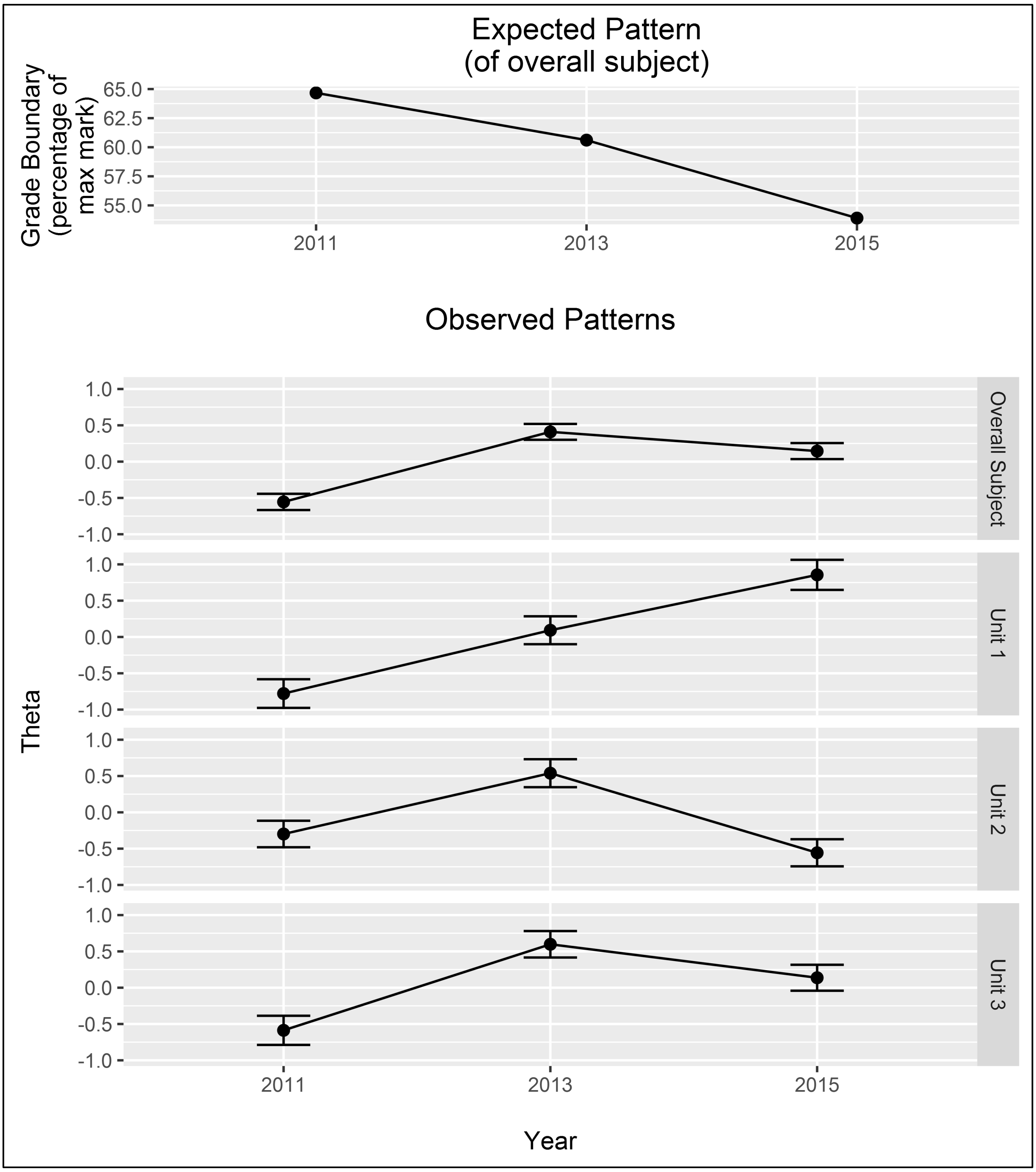

Figure 34. *Expected and observed patterns for geography C-grade scripts*

Note. The scales are different and therefore non-comparable between the expected and observed patterns. Theta scores are also placed on a logit scale, and therefore cannot be used additively. Readers should therefore avoid drawing conclusions about the size of any differences, but should rather focus upon the overall patterns of changes.

4.3.6 Survey findings

Judges were also given the opportunity to provide some feedback on the task via a short online survey. Twenty-two of twenty-four judges completed this survey. Overall, judges felt reasonably confident in their judgements, which was also supported by the good separation reliabilities and judge infit statistics already presented.

Encouragingly, although 64% of judges rated the task overall as being "difficult", none stated that they found it "very difficult". When asked what they found difficult about the task, the most common response was that it was difficult to judge scripts from different units. This perhaps suggests that the intra-unit comparisons (ie the changes in performance over time; the main patterns of interest) were less difficult to make. Judges were also able to make their judgements within a reasonable time (modal time = 15 minutes). This short time also supports the ability of judges to make holistic judgements of quality, rather than feeling the need to fully re-mark the scripts, which would have taken longer to do. This was also confirmed by the judges themselves, as more than half said that they made no attempt to re-mark, and the rest either informally or mentally marked some questions, but not the whole paper. Less than 25% of judges felt that the mark scheme would have helped them make more accurate judgments.

Judges said that they found it more challenging to take the difficulty of question papers into account when judging the candidates' performance. 41% of judges found this "difficult" and 14% found this "very difficult". As before, judges stated that this was made more difficult when comparing question paper demands across different units (possibly suggesting that their intra-unit judgements were less affected). Although several judges stated that there were no great differences in difficulty between the papers, it should be noted that past research has shown that such differences can bias judges' perceptions (eg Good & Cresswell, 1988 – discussed further in the following section).

We also sought to gain some insight into how judges attempted to make their decisions. Some judges stated that they focused upon certain questions. This depended upon the subject, but responses suggested that they were able to identify certain types of questions that gave more information about candidate performance. For example, one geography judge noted that (s)he focused mostly upon "longer, more developed responses" and focused less on multiple choice questions. Other judges relied more on certain features of the candidates' responses, such as the quality of writing, depth of understanding, and / or depth of analysis.

4.4 Discussion

The general agreement between the judges' ratings of script quality and the changes in grade boundaries from Study 1 supports the proposal that changes in grade boundaries over time did reflect changes in cohort performance during the same period. With one major exception (geography), and some minor inconsistencies, this was true both for subjects showing patterns consistent with the Sawtooth Effect, and those that did not. Although some of the changes in grade boundaries might reflect changes in marking leniency or question paper demands over time (because of the minor inconsistencies between Study 1 and Study 2), the fact that in most subjects the experts' judgments did not depart substantively from the expected trends suggests that these alternate possibilities do not explain the results of Study 1. This is because the judges in the current study judged scripts in a random order (addressing the alternate explanation of changes in marker leniency over time) and were asked to compensate for question paper difficulty when making their judgements (addressing the alternate explanation of changes in question paper difficulty over time).

However, although judges' responses to the survey suggested that they felt confident in their ability to take question paper difficulty into account when making their judgements, past research has demonstrated that differences in question paper difficulty can bias judgements of script quality. For example, Good and Cresswell (1988) found that perceptions of script quality will often be lower when a paper is more difficult, and higher when a paper is easier. As the demands of papers do fluctuate over time (ie between assessment series), perceptions of script quality may be unknowingly affected. This therefore represents a potential limitation of the current study, if judges were indeed unable to appropriately control for the difficulty of different question papers.

Geography was an interesting case. In particular, there was an apparent disconnect between the grade boundaries, which reduced overall, and mean perceptions of cohort performance, which increased overall. It is possible that this was caused by some of the minor reforms to geography specifications that occurred during the period of interest. As noted in Figure 2, the new spelling and grammar rule came into effect for geography assessments in 2013, and demands were raised in 2015 assessments. Each of these changes may have affected the positions of the grade boundaries in those years, whilst not affecting judges' ratings of performance. In other words, it is possible that student performance in these assessments did increase as expected under the Sawtooth Effect, but the grade boundaries did not reflect this change. The same might be also true for those other subjects that were affected by the new spelling and grammar rule in 2013 (ie history and religious studies). In 3 of the 4 grade boundaries investigated for these subjects, the difference in mean judges' ratings between 2013 and 2015 could be attributed to measurement error, while a difference was perhaps expected according to the grade boundaries. As before, it is possible that grade boundaries in 2013 were lower in these subjects than in 2015 because they were affected by the spelling and grammar rule in 2013, whilst underlying performance might have actually changed little between these years. Such claims need further exploration, but offer a potential explanation for these findings.

5 General discussion

Each of the 2 studies reported within this paper have provided general support for the existence of the Sawtooth Effect in UK secondary school assessments. In Study 1 we demonstrated that average unit-level performance (measured by proxy via changes in judgemental grade boundaries over time) increased during successive years immediately following the last set of reforms to a range of GCSE and AS / A level assessments, with the rate of this increase seemingly lessening after around the third assessment year. In Study 2, we tested the proposal that these changes in grade boundaries did indeed reflect changes in underlying performance, using a small sample of GCSE assessments. We concluded that, with some exceptions, this did appear to be the case.

Several novel contributions have been made by this research, enhancing our understanding of the nature of the Sawtooth Effect. Firstly, we have demonstrated that this effect appears to have an influence across a range of high-stakes assessments at secondary school, and not just for English and maths subjects, which previous research has exclusively (to our knowledge) focused on. Secondly, we have contributed to the literature regarding estimates of magnitude and duration, suggesting that post-reform performance gains occur to quite a small degree on average over a period of approximately 3 assessment years (around 2% change in estimated outcomes each year), followed by a more stable change in each year thereafter (around 0.5% change in estimated outcomes each year). However, it is worth emphasising again that these are average values, and some subjects did change to a lesser or greater degree. It is also worth noting that we have not been able to disentangle increases in performance due to test familiarity from those due to more genuine improvements in ability (although one would expect that improvements due to the former can be made more quickly than due to the latter – Koretz & Barron, 1998 – meaning that improvements due to test familiarity might explain a greater portion of the more rapid changes occuring in the first few years). Thirdly, we have demonstrated that not all subjects appear to be affected by the Sawtooth Effect in the same way, as patterns of grade boundary change for some subjects did not follow the typical sawtooth pattern (eg see Figure 17).

The fact that grade boundaries (and by proxy, performance) rose over time on average and for most assessments means that those involved in the awarding of assessments (not necessarily restricted to those investigated here) should be aware of the possibility of test-specific gains following assessment reform, especially for any awarding processes that are not statistically controlled (eg by the comparable outcomes method). In particular, relevant parties should be mindful of making comparisons across cohorts in the early years of a new assessment, to avoid drawing unfair conclusions about a cohort’s performance simply because they were the first students to be entered for the new assessment (which is the underlying ethical imperative of the comparable outcomes approach to awarding). Our results suggest that it may take about 3 years on average for the majority of these adjustments to happen. This is consistent with the results of Koretz et al. (1991) and Cresswell (2003), whose results suggested post-reform periods of around 2 and 4 years respectively. However, this finding needs confirming in different contexts, particularly due to the presence of multiple testing series during the first 3 years of the period of interest, which may have led to an accelerated rate of change due a greater opportunity to gain test familiarity. One should also be cautious against drawing conclusions for individual students / schools, as there will be variation in the size of this effect at these lower levels due to differences in way assessment reform is handled. Similarly, it may be easier to adapt to changes made in some subjects compared to others (ie students and teachers may gain familiarity at different rates for different subjects, leading to a greater rate of change in cohort performance in some subjects compared to others).

Finally, although outside of the main focus of this report, the current results seem to provide support for the validity of general assessments in the UK, and can be used to help allay concerns of a systematic ‘dumbing down’ of assessments. While it is not impossible that the small increases in grade boundaries observed are due to assessments becoming easier, this is unlikely, and changes in test familiarity is a more plausible explanation. Given that at least a proportion of the largest increases (ie those in the first 3 years) are most likely attributable to increases in test familiarity, and that the more recent increases are very small, any changes in demand, if present at all, have not occurred to any great degree. The fact that boundaries began to plateau after around 3 years is also perhaps more likely explained by the limits of familiarity being approached, rather than the ease of assessments coincidentally following a similar pattern. Of course, it is possible that there have been instances of reductions in demand for individual subjects, but this does not appear to have occurred at an aggregate level.

5.1 Limitations and suggestions for future research

Although the current research has allowed us to explore the nature of the Sawtooth Effect in UK awarding, there are certain caveats that may limit one’s ability to make predictions about future effects based on these findings.

Firstly, the extent to which we can generalise from these results is limited, because the size and duration of any future Sawtooth Effects will likely be dependent upon the size of the reforms taking place. For example, should future reforms prove to be more substantial and wide-ranging than the last set of reforms (eg in terms of how much teaching needs to change) then the degree and duration of post-reform performance gains may be larger / longer than was observed here. As already mentioned, the rate of change between 2010 and 2013 might also be somewhat quicker than what might occur in the future, because the presence of January testing series (which were discontinued in 2014, after most of the gains to through test familiarity had been achieved) might have meant that teachers and students were able to gain familiarity more quickly. The opposite might also be true. Continued monitoring of grade boundary change would be needed to determine the point at which differences between years begin to plateau. Although we might be able to make some qualitative judgements regarding the likely difference in severity between specification periods, this difference is difficult to quantify. Nevertheless, this might only affect our interpretation of the post-reform performance gains period, which will depend on the size of the reforms. Once these rapid improvements have waned, one might reasonably expect the size of changes in performance to be comparable across ‘stable’ periods of each specification lifespan.

Secondly, further work may need to be done to assess the accuracy of our estimates of the magnitude of changes occurring. The results of Study 2 demonstrated that although grade boundaries did appear to be a fairly good indication of underlying performance, there were some disagreements between the patterns of grade boundaries and perceptions of performance change. Geography was a good example of this, whereby the grade boundaries suggested decreasing performance over time, whilst the judges perceived the opposite. In addition, although we ruled out the possibility that subtle changes in question paper demands explained the entire pattern of results, this explanation might still have a minor effect on our estimates of outcome change. Confirmatory work is therefore suggested to establish whether our estimates are a fair reflection of the changes in performance occurring during the post-reform period.

Thirdly, the change from modular to linear might affect our interpretation of the plateau in overall GCSE trends. The apparent restabilisation of grade boundaries (and by proxy, performance) in the 2014 series may in part be explained by the discontinuation of modular assessments. As discussed by Vidal Rodeiro and Nádas (2012), the move from modular to linear routes increases the amount of content needed to learn for the test (ie a whole year’s information versus one unit’s), can place greater stress on students (as they have to take more exams during one series), and provides less formative feedback for students during the course (ie students can use results from January exams to identify learning needs, as well as having the opportunity to make corrections for the June exams). Modular deliveries can also offer much more flexibility for schools (eg Vidal Rodeiro & Nádas, 2012, reported over 5000 different unit/series combinations for GCSE maths in 2009), which means they can adapt their teaching to match their own strengths. These factors might therefore lead one to expect a dip, or at least no rise, in performance following linearisation, compensated by a lowering of grade boundaries to maintain outcomes. However, although this might explain the lack of change between 2013 and 2014, the fact that grade boundaries did not once again rise in 2015 (when schools/teachers have had time to adapt to this change) suggests that the plateau in performance is not entirely explained by linearisation, as one would expect boundaries to rise once more if that were the case. In addition, the fact that the rate of change in AS / A level boundaries appeared to reduce somewhat after 3 years (one year before linearisation) also gives credence to the assumption that GCSE boundaries have also plateaued after 3 years partly due to familiarity, rather than linearisation.

Fourthly, some caution is needed because of the difficulties in interpreting the current results imposed by the aggregation of unit outcomes to arrive at overall subject grades. We have already discussed how choice of boundaries for examined units might be limited by the need to compensate for high gains in controlled assessments. The discontinuation of January exams may again have had some effect on the way units are aggregated. Bramley (2013a) discussed how under a modular route the outcomes of units assessed in January cannot be changed once awarded, therefore boundaries in June may reflect some compensation for awards made in earlier series in order to meet predicted subject-level outcomes. Such adjustments will have no longer occurred in 2014 and 2015 (because January exams had been discontinued), and so this may have affected the trends observed in the current findings. Again, however, although this might lead one to expect a change in trends between 2013 and 2014, it is difficult to see why this would carry over to produce no change between 2014 and 2015.

Finally, although fulfilling its purpose of maintaining standards, there are some limitations with the comparable outcomes approach that might possibly affect our explanation of *why* these trends have occurred. For example, Bramley (2013b) found (using simulated data) that prediction matrices can underestimate true outcomes when students and teachers have a choice over which exams students are entered for. To put it another way, students may do better than what was predicted when they enter into exams that they are expected do better in (without necessarily having any increases in underlying ability). Such choices might, for example, be based upon personal preferences, course content, and/or methods of assessment (Bramley, 2013b). It is possible, therefore, that the rising trends in performance reported here reflect student entry patterns (eg tactical entry into exams where students are expected to do their best, with no change in actual ability over time), rather than a change in ability caused by measurement driven instruction (which would suggest a change in ability on tested materials over time). Tactical entry patterns might be expected to cause similar changes as measurement driven instruction because both practices depend upon familiarity with assessments. Therefore, although this potentially limits our understanding of why this pattern has occurred, this would not change our interpretation of the results as reflecting underlying test-specific performance gains following assessment reform due to increases in assessment familiarity. As such practices would likely continue into the new specification lifespan, a similar pattern of results would still be expected.

In addition to conducting research to address the aforementioned limitations, the current study has also raised a number of questions that might also draw focus in future research activities. Firstly, although students' performance appears to have improved over time, it is unclear which party is most 'responsible' for these improvements, and it would be interesting to determine whether these trends mainly reflect changes in teachers' or students' familiarity and preparedness. Secondly, as we have emphasised throughout, not all assessments / units / subjects follow the average trends reported here. Further research is necessary to identify the factors that determine whether the Sawtooth Effect applies for an individual assessment, and what factors determine the duration and size of this effect. A further avenue for future research would be to see whether the 3 year adjustment period remains consistent for other qualification types, such as vocational qualifications or Key Stage 2 assessments, as there may be differences in institutions' abilities to adapt to assessment change. Thirdly, researchers might investigate further why performance in GCSE examinations appears to have declined in 2014 and 2015. Finally, in Study 2, we did not assess whether changes in grade boundaries appropriately reflect changes in underlying performance on controlled assessments or for AS / A levels. This therefore perhaps also requires further exploration.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 451–462. http://doi.org/10.1177/014662167800200319

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*, 258–267. http://doi.org/10.3102/0013189X07306523

Baird, J.-A. (2007). Alternative conceptions of comparability. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 124–156). London, UK: Qualifications and Curriculum Authority.

Baird, J.-A., Ahmed, A., Hopfenbeck, T., Brown, C., & Elliott, V. (2013). *Research evidence relating to proposals for reform of the GCSE*. Oxford, UK: University of Oxford. Retrieved from http://oucea.education.ox.ac.uk/about-us/research-evidence-relating-to-proposals-for-reform-of-the-gcse/

Baird, J.-A., Chamberlain, S., Meadows, M., Royal-Dawson, L., & Taylor, R. C. (2009). *Students' views of stretch and challenge in A-level examinations*. Manchester, UK: Centre for Education Research and Policy. Retrieved from https://cerp.aqa.org.uk/research-library/students-views-stretch-and-challenge-level-examinations

Baird, J.-A., Daly, A. L., Tremain, K., & Meadows, M. (2009). *Stretch and challenge in A-level examinations: Teachers' views of the new assessments*. Manchester, UK: Centre for Education Research and Policy. Retrieved from https://cerp.aqa.org.uk/research-library/stretch-and-challenge-level-examinations-teachers'-views-new-assessments

Benton, T., & Bramley, T. (2015). *The use of evidence in setting and maintaining standards in GCSEs and A levels: Discussion paper*. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/204310-maintaining-standards-discussion-paper-tom-benton-and-tom-bramley.pdf

Benton, T., & Elliott, G. (2016). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, *31*, 352–376. http://doi.org/10.1080/02671522.2015.1027723

Benton, T., & Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE*. Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/investigating-the-relationship-between-a-level-results-and-gcses

Benton, T., & Sutch, T. (2014). *Analysis of use of Key Stage 2 data in GCSE predictions*. Cambridge, UK: Cambridge Assessment. Retrieved from https://www.gov.uk/government/publications/analysis-and-use-of-key-stage-2-data-in-gcse-predictions

Black, B. (2008, August). *Using an adapted rank-ordering method to investiage January versus June awarding standards*. Paper presented at The Fourth Biennial EARLI/Northumbria Assessment Conference. Berlin, Germany.

Bramley, T. (2007). Paired comparison methods. In P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards2* (pp. 246–294). London, UK: Qualifications and Curriculum Authority.

Bramley, T. (2013, August). *Maintaining standards in public examinations: Why it is impossible to please everyone*. Paper presented at The 15th biennial conference of the European Association for Research in Learning and Instruction (EARLI). Munich, Germany.

Bramley, T. (2013b). *Prediction matrices, choice and grade inflation*. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Bramley, T., Dawson, A., & Newton, P. E. (2014, April). *On the limits of linking: Experiences from England*. Paper presented at The 76th annual meeting of the National Council on Measurement in Education (NCME). Philadelphia, PA.

Bramley, T., & Dhawan, V. (2012). Estimates of reliability of qualifications. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp. 217–320). Coventry, UK: Office of Qualifications and Examinations Regulation.

Bramley, T., & Vidal Rodeiro, C. L. (2014). *Using statistical equating for standard maintaining in GCSEs and A levels*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/182461-using-statistical-equating-for-standard-maintaining-in-gcses-and-a-levels.pdf

Cheng, L. (2000). *Washback or backwash: A review of the impact of testing on teaching and learning*. Washington, D.C.: Educational Resources Information Center (ERIC). Retrieved from http://eric.ed.gov/?id=ED442280

Cresswell, M. (2003). *Heaps, prototypes and ethics: The consequences of using judgements of student performance to set examination standards in a time of change*. London: Institute of Education, University of London.

Daly, A. L., Baird, J.-A., Chamberlain, S., & Meadows, M. (2012). Assessment reform: students' and teachers' responses to the introduction of stretch and challenge at A-level. *The Curriculum Journal*, *23*, 139–155. http://doi.org/10.1080/09585176.2012.678683

Department for Education. (2010). *The Importance of Teaching: The Schools White Paper 2010*. London: Department for Education. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/175429/CM-7980.pdf

Department for Education. (2011). *Early entry to GCSE examinations report*. (DfE Research Report DFE-RR208). London: Department for Education. Retrieved from https://www.gov.uk/government/publications/early-entry-to-gcse-examinations

Department for Education. (2015a). *Early Entry – Guidance for schools*. London: Department for Education. Retrieved from http://www.education.gov.uk/schools/performance/secondary_14/GCSE_Early_Entry_Guidance.pdf

Department for Education. (2015b). *Progress 8 measure in 2016: Guide for* *maintained secondary schools, academies and free schools*. London: Department for Education. Retrieved from https://www.gov.uk/government/publications/progress-8-school-performance-measure

Elwood, J., Hopfenbeck, T., & Baird, J.-A. (2015). Predictability in high-stakes examinations: students’ perspectives on a perennial assessment dilemma. *Research Papers in Education*, 1–17. http://doi.org/10.1080/02671522.2015.1086015

Gill, T. (2014). An investigation of the effect of early entry on overall GCSE performance, using a propensity score matching method. *Research Matters: A Cambridge Assessment Publication*, *18*, 28–38. Retrieved from http://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Gill, T., & Bramley, T. (2008, September). *How accurate are examiners’ judgments of script quality?*. Paper presented at the British Educational Research Association Annual Conference. Edinburgh, UK. Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/175604_TG_TB_Zones_of_Uncertainty_BERA08.pdf

Good, F. J., & Cresswell, M. (1988). *Grading the GCSE*. London, UK: Secondary Examinations Council.

House of Commons. (2008). *Testing and Assessment: Government and Ofsted Responses to the Committee’s Third Report of Session 2007-08*. London: House of Commons. Retrieved from http://www.publications.parliament.uk/pa/cm200708/cmselect/cmchilsch/1003/1003.pdf

Jacob, B. A. (2002). *Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools*. (NBER working paper no. 8968). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w8968

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND. Retrieved from http://www.rand.org/pubs/issue_papers/IP202.html

Koretz, D. M. (2005). *Allignment, high stakes, and the inflation of test scores*. (CSE Report 655). Los Angeles, CA.: National Center for Research on Evaluation. Retrieved from http://eric.ed.gov/?id=ED488711

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Washington, D.C.: RAND. Retrieved from http://www.rand.org/pubs/monograph_reports/MR1014.html

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at The Annual Meetings of the American Educational Research Association. Chicago, IL. Retrieved from http://eric.ed.gov/?id=ED340730

Linn, R. L. (1998). *Assessments and Accountability*. (CSE Technical Report 490). Los Angeles, CA.: National Center for Research on Evaluation. Retrieved from https://eric.ed.gov/?q=Assessments+and+Accountability&ff1=autLinn,+Robert+L

.&id=ED443865

Linn, R. L. (2000). Assessments and Accountability. *Educational Researcher*, *29*, 4–16. http://doi.org/10.3102/0013189X029002004

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). *Comparing state and district test results to national norms: Interpretations of scoring “above the national average.”* (CSE Technical Report 308). Los Angeles, CA: UCLA Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://eric.ed.gov/?id=ED338676

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Ofqual. (2011). *GCSE, GCE, principal learning and project code of practice*. Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/gcse-gce-principal-learning-and-project-code-of-practice

Ofqual. (2013). *Review of controlled assessment in GCSEs*. (Ofqual Report 13/5291). Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/review-of-controlled-assessment-in-gcses

Ofqual. (2015a). *Developing New GCSEs , AS and A Levels for First Teaching in 2017 – Part 1*. Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/consultations/development-of-new-gcses-and-a-levels-for-teaching-from-2017

Ofqual. (2015b). *Setting GCSE , AS and A level grade standards in summer 2014 and 2015*. Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/setting-gcse-and-a-level-grade-standards-in-summer-2014-and-2015

Ofqual. (2015c). *Summer 2015 Data Exchange Procedures: GCE , GCSE and Level 1/2 Certificates*. Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/data-exchange-procedures-for-a-level-gcse-level-1-and-2-certificates

Ofsted. (2008). *Mathematics: Understanding the score*. London, UK: The Office for Standards in Education. Retrieved from http://webarchive.nationalarchives.gov.uk/20141124154759/http://www.ofsted.gov.uk/resources/mathematics-understanding-score

Pollitt, A. (1998, September). *Maintaining Standards in Changing Times*. Paper presented at International Association for Educational Assessment. Barbados.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*, 157–170. http://doi.org/10.1007/s10798-011-9189-x

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappa*, *68*, 679–682. Retrieved from http://www.jstor.org/stable/20403467

Robitzsch, A. (2016). SIRT: Supplementary Item Response Theory Models. Retrieved from https://cran.r-project.org/web/packages/sirt/index.html

Rushton, N. (2013). Changing times, changing qualifications. *Research Matters: A Cambridge Assessment Publication*, *16*, 2–9. Retrieved from http://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Shepard, L. A. (1998, April). *Should instruction be measurement-driven?: A debate*. Paper presented at The Annual Meeting of the American Educational Research Association. New Orleans.

Shepard, L. A., & Dougherty, K. C. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at The Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education. Chicago.

Stecher, B. M. (2002). Consequences of large-scale, high stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making Sense of Test-Based Accountability in Education* (pp. 79–100). Santa Monica, CA: RAND.

Stecher, B. M., Chun, T., & Barron, S. I. (2004). The effects of assessment-driven reform on the teaching of writing in Washington State. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 53–72). New Jersey: Lawrence Erlbaum Associates.

Stringer, N. S. (2012). Setting and maintaining GCSE and GCE grading standards: The case for contextualised cohort-referencing. *Research Papers in Education*, *27*, 535–554. http://doi.org/10.1080/02671522.2011.580364

Sturman, L. (2003). Teaching to the test: science or intuition? *Educational Research*, *45*, 261–273. http://doi.org/10.1080/0013188032000137256

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286. http://doi.org/10.1037/h0070288

Vidal Rodeiro, C. L., & Nádas, R. (2010). *Effects of modularisation*. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/images/109794-effects-of-modularisation.pdf

Vidal Rodeiro, C. L., & Nádas, R. (2012). Effects of modularity, certification session and re-sits on examination performance. *Assessment in Education: Principles, Policy & Practice*, *19*, 411–430. http://doi.org/10.1080/0969594X.2011.614218

Wuertz, D., & Chalabi, Y. (2013). fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. Retrieved from http://cran.r-project.org/package=fGarch

Appendices

Appendix A – Graphs with all units included

AS / A level

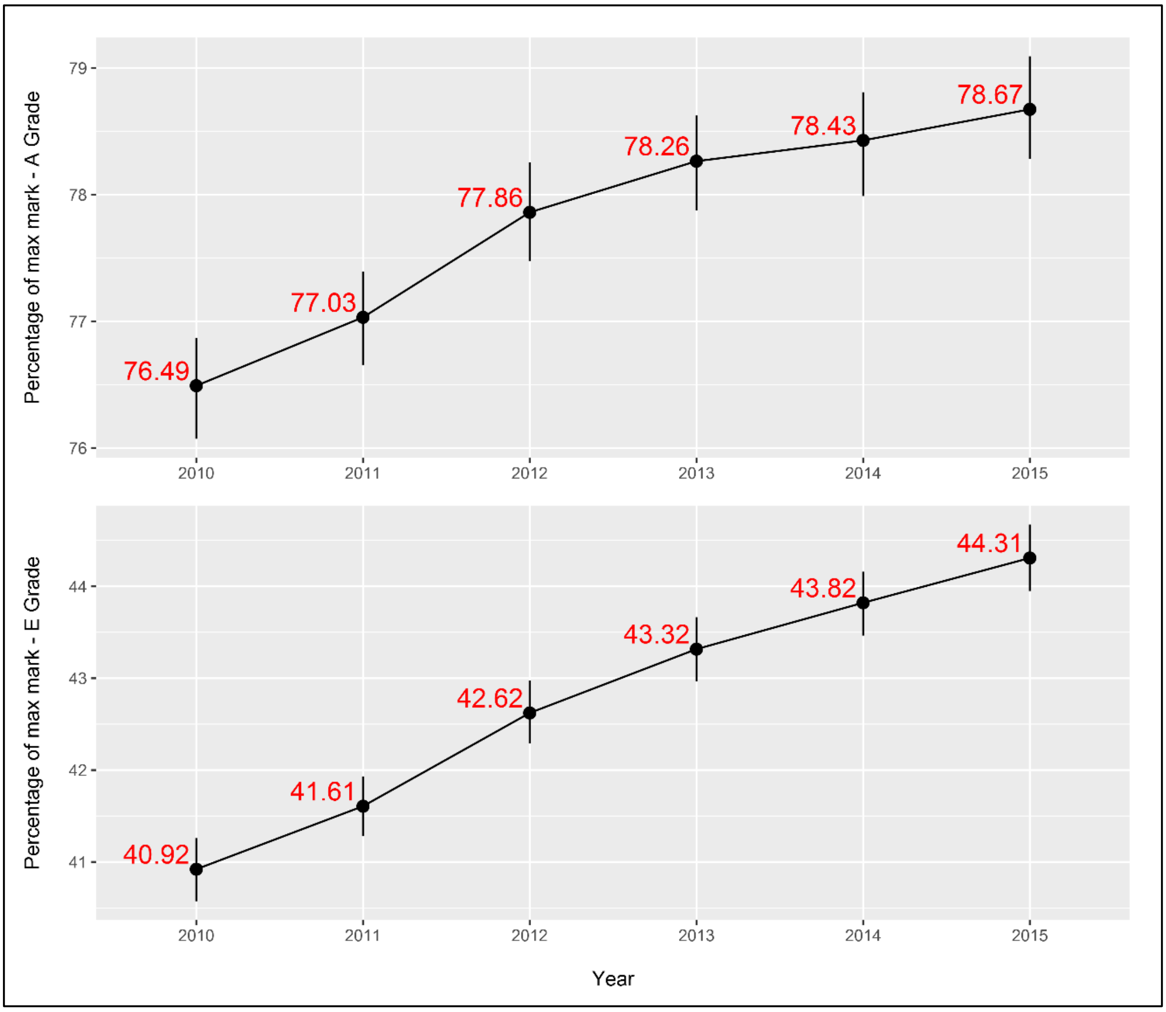

Figure 35. *All AS / A level units (ie with no exclusion criteria)*

Note. n = 1496 for both grades.

GCSE

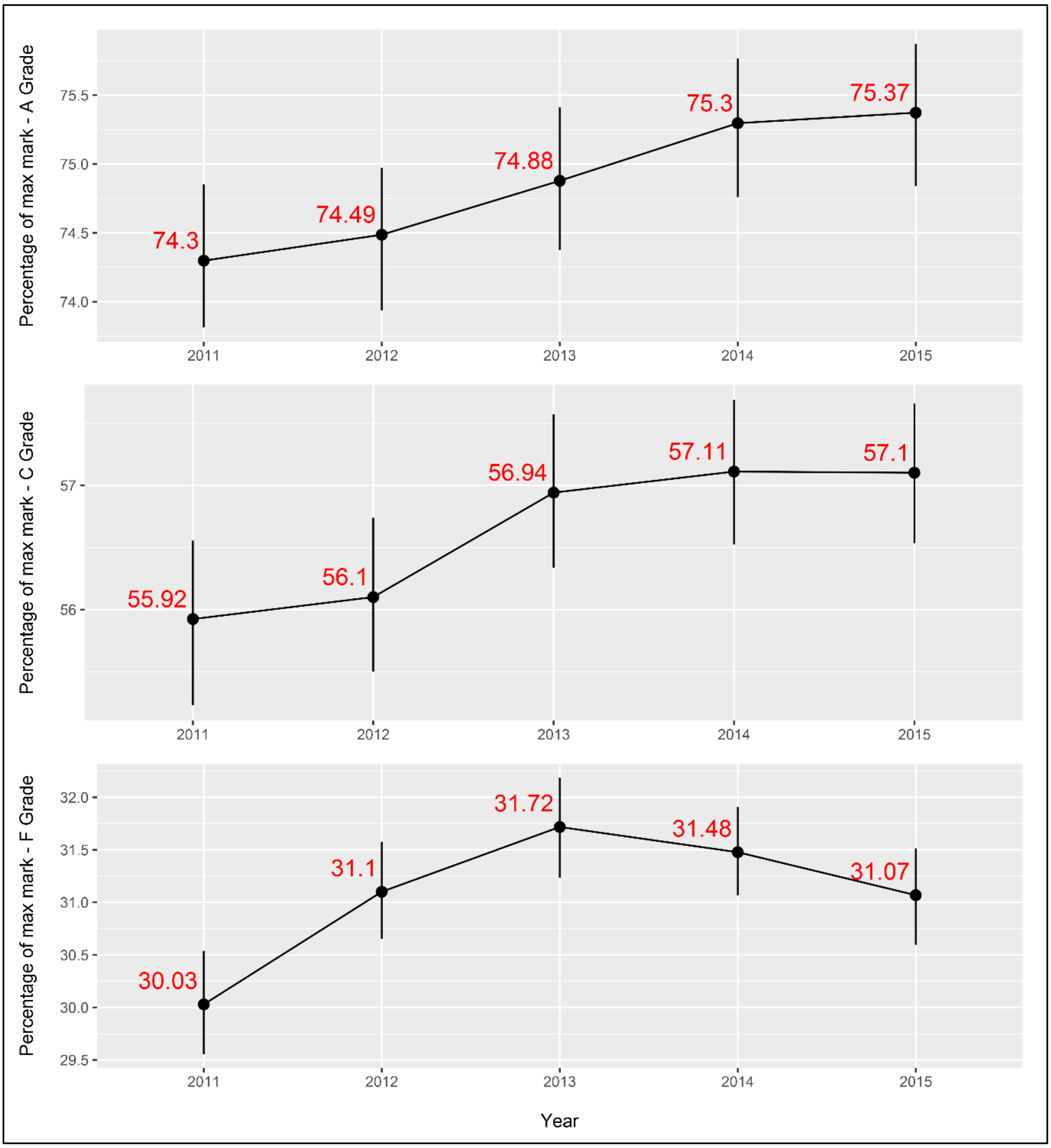

Figure 36. *All GCSE units (ie with no exclusion criteria)*

Note. n = 1531, 1669, and 1526 for A-, C-, and F-grades respectively.

Appendix B – Data simulations

B.1 – Procedure for generating the distributions

The distributions of students' marks were not available for all subjects, and so data was randomly simulated (within certain constraints), using the 'fGarch' (Wuertz & Chalabi, 2013) package for R to produce frequency histograms representative of the data for each year at AS / A level and GCSE level. Percentage marks were plotted on the x-axes, with frequencies on the y-axes. Various parameters were set in place to produce distributions that closely matched the 'average' distribution of subject outcomes, which can be seen in Tables 3 and 4. Although providing adequately representative distributions, this procedure does rely on the assumption that outcomes follow a near perfect Gaussian distribution. Real-life distributions may deviate somewhat from this assumption, representing a limitation of this method.

The grand mean and standard deviation of the curve was calculated from subject level outcomes (ie the average cumulative frequency of students achieving each grade; using the same subjects as in Figures 5 and 11). Kurtosis was not considered an issue, as this will have been determined by the standard deviation. In order to generate the correct amount of skew, constraints were also placed based on the mean cumulative frequency of students exceeding each mean grade boundary. Because it was not possible to achieve these values exactly, the simulation was repeated until a match could be found within the smallest tolerance limits (eg for GCSEs in 2011, the cumulative frequency at each grade was considered a match when the simulation was within 1.9% of stipulated values). As each of the above parameters changed over time, a different simulated dataset was produced for each year. Figures 37 and 38 show the simulated distributions for AS / A level and GCSE. The increasing skew over time for AS / A level is a reflection of the fact that E-grade boundaries increased more rapidly than A-grade boundaries, squashing the distribution towards the upper end. This might also be why a greater tolerance was needed for each of the AS / A level simulations, when compared to GCSEs.

Table 3. *Descriptive statistics for simulated distributions - AS / A levels*.

	2010	2011	2012	2013	2014	2015
Mean						
Input	65.73	66.28	66.89	67.32	67.42	67.76
Achieved	65.78	66.34	66.92	67.42	67.50	67.77
SD						
Input	13.70	13.46	13.08	12.77	12.61	12.45
Achieved	13.71	13.48	13.10	12.74	12.60	12.45
Tolerance for cumulative percentages	3.4%	3.4%	3.7%	3.7%	3.9%	4.1%
A cumulative percentage						
Input	25.13	25.23	24.33	23.99	23.40	23.12
Achieved	21.78	21.99	20.72	20.32	19.55	19.12
E cumulative percentage						
Input	98.25	98.45	98.48	98.66	98.55	98.57
Achieved	95.20	95.07	94.93	94.98	94.91	94.75
Skew	-0.42	-0.57	-0.69	-0.77	-0.71	-0.78
Kurtosis	0.11	0.23	0.33	0.41	0.37	0.48

Table 4. *Descriptive statistics for simulated distributions - GCSEs*.

	2010	2011	2012	2013	2014	2015
Mean						
Input	-	61.14	63.03	63.77	63.18	63.40
Achieved	-	61.09	63.07	63.73	63.13	63.37
SD						
Input	-	17.60	17.12	16.81	17.02	17.00
Achieved	-	17.57	17.08	16.84	17.03	16.99
Tolerance for cumulative percentages	-	1.9%	2.4%	2.5%	2.2%	2.8%
A cumulative percentage						
Input	-	19.83	20.05	19.60	19.08	19.88
Achieved	-	21.61	22.39	21.92	21.20	22.44
C cumulative percentage						
Input	-	65.30	65.85	65.73	64.58	65.85
Achieved	-	63.42	63.61	63.24	62.48	63.09
F cumulative percentage						
Input	-	96.57	96.86	96.93	96.56	96.53
Achieved	-	95.32	95.48	95.44	94.95	95.23
Skew	-	-0.36	-0.42	-0.47	-0.52	-0.50
Kurtosis	-	0.08	0.14	0.20	0.21	0.19

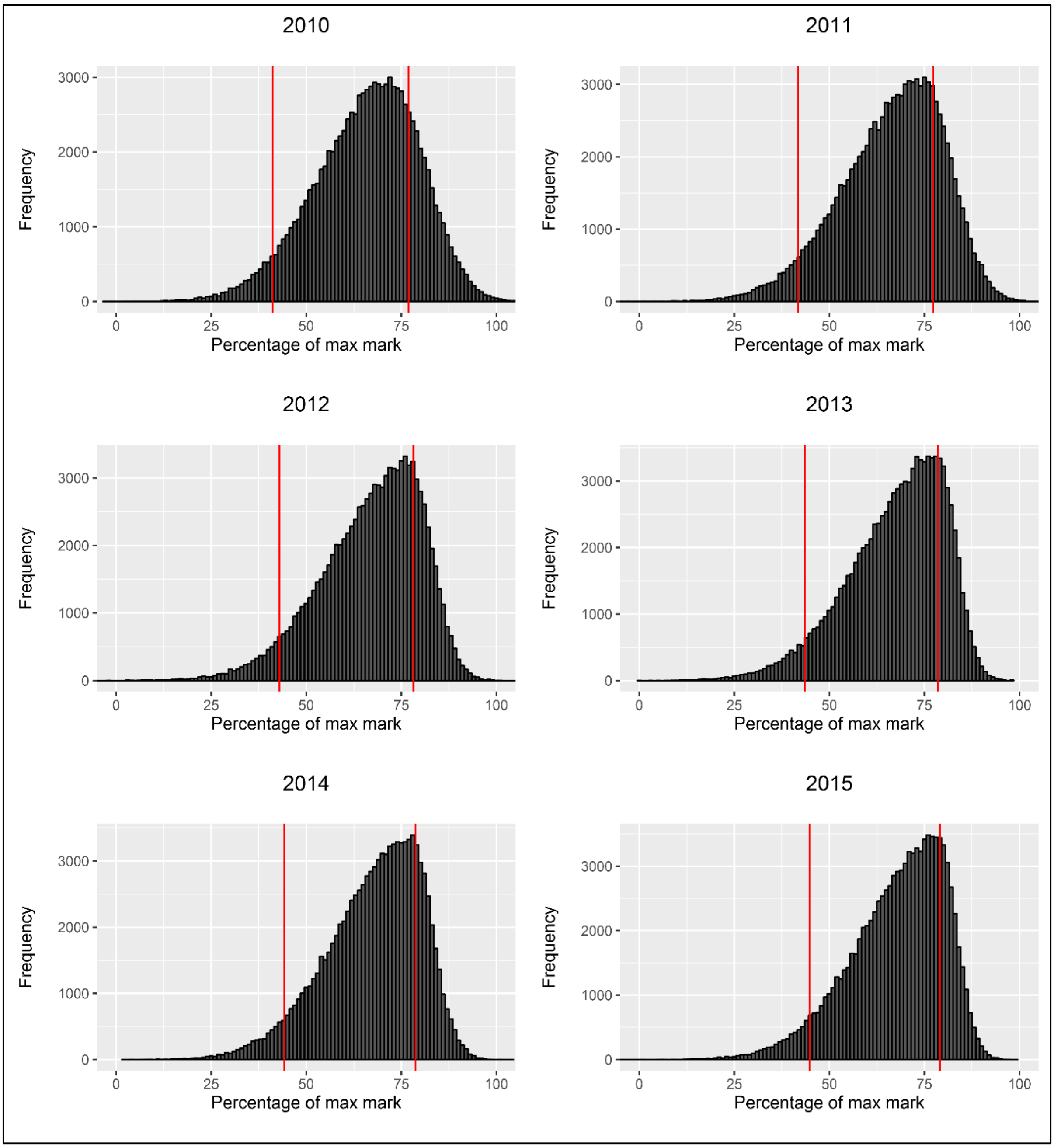

Figure 37. *Simulated distributions for AS / A levels.*

Note. The red vertical lines indicate the position of each judgemental grade boundary (taken from Figure 3).

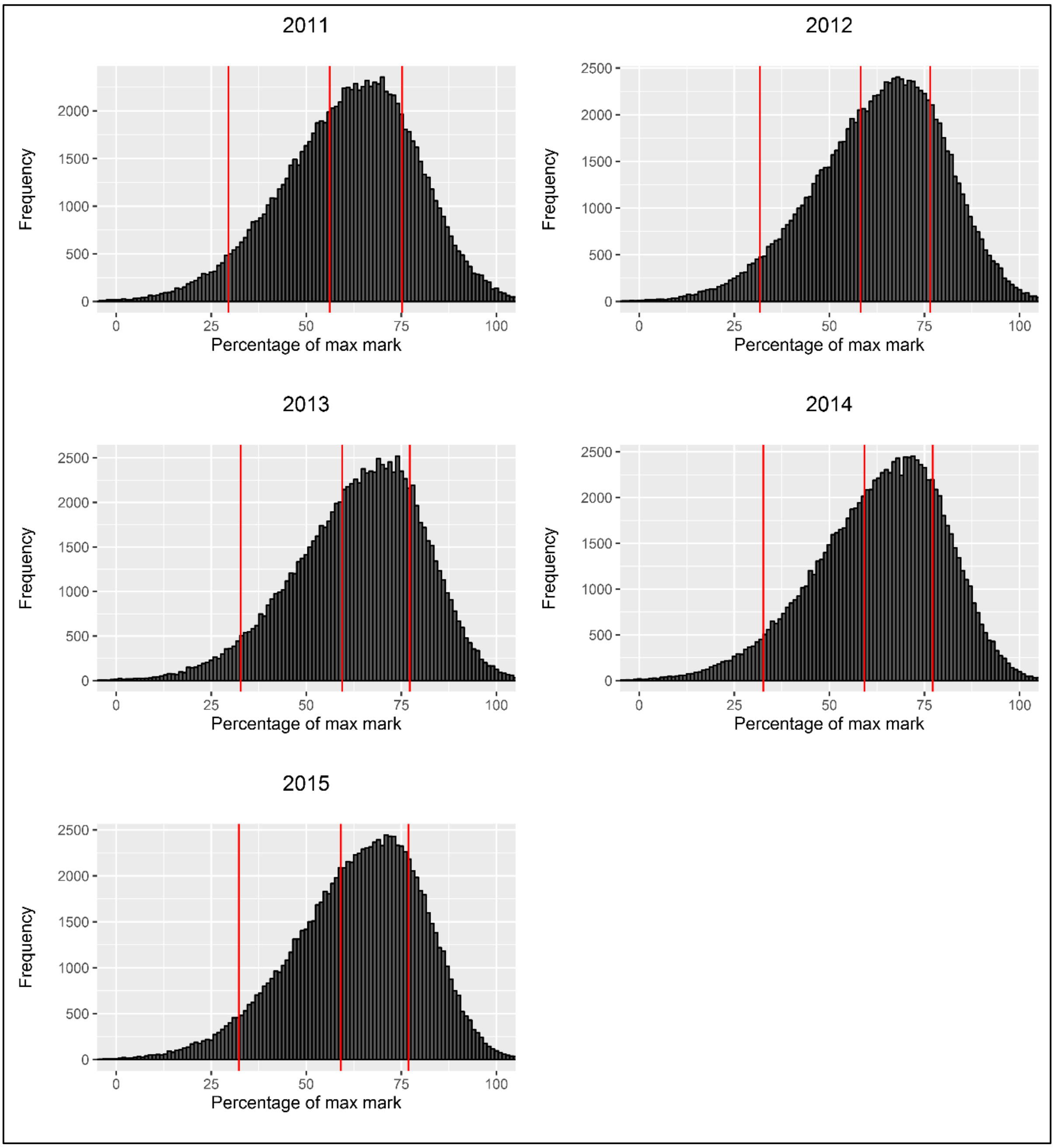

Figure 38. *Simulated distributions for GCSEs.*

Note. The vertical lines indicate the position of each judgemental grade boundary (taken from Figure 9).

B.2 – Adjusting boundaries to return to predicted outcomes

By using the simulated distributions from Appendix B.1, one can adjust each grade boundary to take into account the fact that actual outcomes did not exactly match what was predicted. This was an important consideration, as the trends in grade boundary change may have simply been explained by a drift from predicted outcomes, rather than any change in underlying performance.

To give an example, in the awarding of 2010 AS / A levels it seems as though, on average, 0.23% (of the total entry) more students achieved an A-grade than were predicted. To adjust for this, the A-grade boundary can be increased, so as to remove 0.23% of the simulated sample from those exceeding the A-grade boundary. In this example, moving the mean boundary from 76.85 to 76.94 achieved this. It was not always possible to move the boundaries in a way that affected exactly the number of simulated students that we desired. In such cases, the boundary was set to the closest possible value. These adjustments were made for each boundary, in each year, at both AS / A level and GCSE level (Table 5). Nevertheless, making these adjustments did not change the trends observed in Section 3.3, as adjusted values rarely departed far from the raw grade boundaries.

Table 5. *Raw and adjusted grade boundaries.*

Boundary	2010	2011	2012	2013	2014	2015
AS / A level						
A - Raw	76.85	77.30	78.18	78.60	78.72	79.07
A - Adj.	76.94	77.45	78.36	78.76	78.81	79.18
E - Raw	41.17	41.76	42.88	43.57	44.21	44.83
E - Adj.	41.32	42.08	43.21	44.23	44.89	45.45
GCSE						
A - Raw	-	75.19	76.53	77.21	77.20	76.85
A - Adj.	-	75.15	76.61	77.26	77.21	76.96
C - Raw	-	56.15	58.28	59.43	59.21	59.11
C - Adj.	-	56.08	58.38	59.56	59.23	59.23
F - Raw	-	29.56	31.75	32.79	32.63	32.25
F - Adj.	-	29.69	32.44	33.44	32.92	32.15

B.3 – Approximating outcome changes from boundary changes

As a reverse of the process described in Appendix B.2, the simulated distributions can also be used to estimate how many students will have been affected by changes in grade boundaries over time. For example, it can allow us to determine how many students would have achieved a C-grade at GCSE in 2012, had the comparable outcomes approach not been used to set grade boundaries to account for test-specific performance gains between the 2011 and 2012 cohorts. This can allow us to provide an estimation of test-specific performance change between these years.

To give an example, Figure 39 shows the GCSE distribution for 2012. The average C-grade boundary in 2012 (58.28%; the green line) had increased by 2.13% from 2011 (56.15%; the red line). If this boundary is brought back down by 2.13%, then the number of students that would have received a C-grade based on the gains in their test-specific performance, had the boundaries not been set with comparable outcomes in mind, can be identified. In effect, this will be the number of students that fall between the red and green lines. In this example, 4,228 simulated students fell between the 2 lines, meaning that 4.23% of the sample (of the 100,000 that were simulated) would have improved their grade (ie from a D to a C) between years, had the grade boundaries remained the same. These calculations were repeated for each boundary, in each year, for AS / A levels and GCSEs (see Table 2).

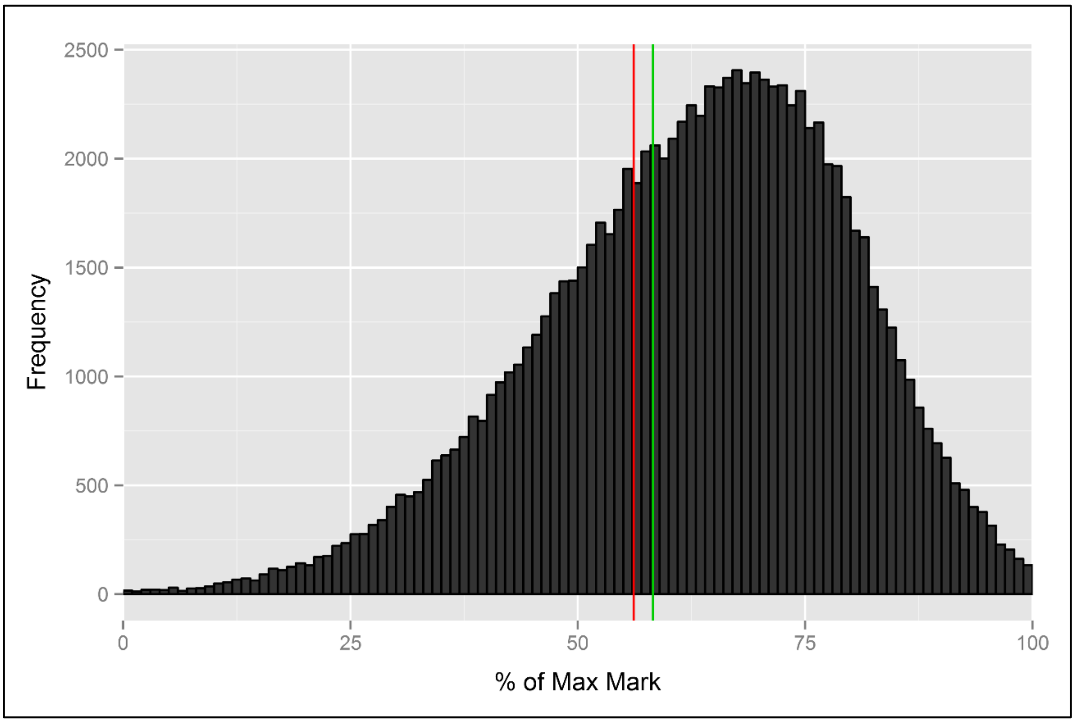

Figure 39. *Simulated distribution for 2012 GCSEs.*

Note. The vertical lines indicate the position of mean C-grade boundaries in 2011 (red) and 2012 (green).

Appendix C – Pack design for Study 2

Once the materials had been 'cleaned', it was necessary to arrange them into 'packs' of 3 scripts for the examiners to judge. To give enough total comparisons per script, each script was included within 15 packs, and because each script was compared twice within each pack (ie to each of the other 2 scripts), this gave a total of 30 paired comparisons per script. To accommodate these numbers, a total of 150 packs per grade boundary were collated for each 2-unit subject (history and religious studies)[19], and 225 packs per grade boundary were collated for each of the 3-unit subjects (maths, history, and geography)[20].

One aspect of the Rasch analysis that needed to be taken into account within the pack design was the fact that estimates of quality (known as 'theta' scores) are placed on an arbitrary logit scale, meaning that this scale cannot be linked across analyses using different materials. Therefore, if separate analyses were conducted for each unit within a subject, we would have been unable to aggregate unit outcomes to deduce subject level changes in performance (as perceived by judges). We therefore needed to include all scripts from within a subject in the eventual analyses, and therefore combined them within the packs. This presented an additional complication, however, because comparing quality across different content areas (eg human geography versus physical geography) might have been difficult for the judges. To help make the task somewhat easier, packs were presented in a deliberate order, according to the different types of comparison that could be made.

As we believed that the easiest comparison to make would be between 3 scripts from different years of the same unit (ie within the same content area), this type was presented to judges first. Packs from different units were presented in a random order, rather than having the examiners judge all packs from one unit followed by all packs from another. The next set of packs again contained scripts from 3 different years, but had 2 scripts from the same unit and one script from a different unit. The penultimate set of packs contained 3 scripts from 3 different years and 3 different units (this type was excluded for 2-unit subjects). The final set of packs contained 3 scripts from different units of the same year (for 2-unit subjects, this type had 2 scripts from different units of the same year; the third script was from a different year). As the first type of comparison was the easiest, and because we were most interested in this type (ie changes in performance over time; the other types were included to link the units together), judges received more of these packs than from the other types. Scripts from the same unit and year were never included within the

[19] (30 scripts [5 per unit per year] x 15 instances) / 3 scripts per pack = 150 packs

[20] (45 scripts [5 per unit per year] x 15 instances) / 3 scripts per pack = 225 packs

same pack, because they were assumed to be of very similar quality, due to the way in which they had been selected for archiving.

For a triples comparison design and 15 scripts per unit, there were 85,140 possible packs to choose from for the 3-unit subjects, and 24,360 possible packs to choose from for the 2-unit subjects. Using R for Windows, packs were randomly sampled so that the correct number of packs were chosen for each type of comparison, whilst making sure that each individual script featured in exactly 15 packs overall. Although this meant that each script was not necessarily compared with all of the other scripts in the sample, paired comparison designs are able to handle missing data well[21]. Once the packs had been chosen, they were equally divided amongst the judges within each subject area. The resulting pack design is summarised in Table 6.

Table 6. *Summary of the pack design*

	Maths	History	RS	Geography
Years	2012 2013 2015	2011 2013 2014	2011 2013 2015	2011 2013 2015
Number of scripts per unit (5 per year)	15	15	15	15
Number of scripts per grade boundary (15 per unit)	45	30	30	45
Number of packs per grade boundary	225	150	150	225
Total number of packs (across all grade boundaries)	675	300	300	225
Number of packs per judge	3 judges with 112 3 judges with 113	50	50	3 judges with 37 3 judges with 38

[21] Although one needs to ensure that each script is involved in enough comparisons overall, non-random missing data does not affect the results because the separation between 2 scripts on the scale of quality produced by the Rasch analysis does not depend on which other scripts they had been compared to (Bramley, 2007).

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346