Fifty years of A-level mathematics: Have standards changed?

Ian Jones*, Chris Wheadon**, Sara Humphries* and Matthew Inglis*

*Mathematics Education Centre, Loughborough University

**No More Marking Ltd.

Correspondence concerning this article should be addressed to Ian Jones,

Mathematics Education Centre, Loughborough University,

Loughborough, LE11 3TU, UK.

Email: I.Jones@lboro.ac.uk

Abstract

Advanced-level (A-level) mathematics is a high-profile qualification taken by many school leavers in England, Wales, Northern Ireland and around the world as preparation for university study. Concern has been expressed in these countries that standards in A-level mathematics have declined over time, and that school leavers enter university or the workplace lacking the required mathematical knowledge and skills. The situation in England, Wales and Northern Ireland reflects more general international concerns about decreasing educational standards. However, evidence to support this concern has been of limited scope, rarely subjected to peer-review, and of questionable validity. Our study overcame the limitations of previous research into standards over time by applying a comparative judgement technique that enabled the direct comparison of mathematical performance across different examinations. Furthermore, unlike previous research, all examination questions were re-typeset and candidate responses rewritten to reduce bias arising from surface cues. Using this technique, mathematics experts judged A-level scripts from the 1960s, 1990s and the current decade. We report that the experts believed current A-level mathematics standards to have declined since the 1960s, although there was no evidence that they believed standards have declined since the 1990s. We contrast our findings with those from previous comparison studies and consider implications for future research into standards over time.

Keywords: A level mathematics, standards, assessment, comparative judgement

Background

Numerous articles and reports have been published over recent years decrying the mathematical knowledge of school leavers in England and Wales (e.g. Walport et al., 2010; ACME, 2011). This includes those who have achieved high grades in Advanced-level (A-level) mathematics (Croft, Harrison & Robinson, 2009; Hawkes & Savage, 2000), a course usually associated with achieving university entrance to science, engineering and mathematics courses in England and Wales. High-profile and on-going media coverage (e.g. Willis & Paton, 2009) suggests that standards were higher some time in the past, but have declined since. In this article we investigate whether this is in fact the case.

Concerns about declining standards perhaps go back as far as accredited education itself, but of particular relevance to the current debate in England and Wales is the influential Dearing report (National Committee of Inquiry into Higher Education, 1997). The report expressed strong concern about falling standards, and set in motion a process by which all subjects at A level would undergo an in-depth review to ensure that standards are maintained over time. There ensued a vast national archiving process that continues to this day (Robinson, 2007), alongside reports into the maintenance of standards over periods varying from one to five years (QCA, 2006a). Independent studies also investigated standards over time using alternative methods (Bramley, Bell & Pollitt, 1998; Christie & Forest, 1980; Coe, 2007; Lawson, 2003; Quinlin, 1995). Taken together these reports and studies suggest a decline since the late 1980s, but differ on its extent or when it may have occurred.

In this article we present a study that spans a wider time period than previous studies into A-level mathematics standards over time. In addition, the study reported

here was specifically designed to minimise the limitations that threaten the validity of previous findings.

First, we summarise the methods used in standards comparability studies, and attempt to synthesise findings, for the case of A-level mathematics. We then present a case for using the method adopted here, called comparative judgement, and describe how we sought to minimise the common threats to validity that plague comparisons of examination standards. Following this we present the main study and discuss the findings in light of previous results.

Measuring standards over time

Four distinct approaches have previously been applied to the comparison of standards in A-level mathematics over time, which we call here *cross-moderation*, *common test*, *expert review* and *comparative judgement*. Statistical methods are also used to compare standards across different syllabuses (Stringer, 2012; Wheadon, 2013), but these have not commonly been applied to the case of A-level mathematics and are not discussed further here.

Cross-moderation.

A cross-moderation exercise involves examiners scrutinising candidates' examination scripts (Adams, 2007). The scripts are drawn from different assessments, such as examination papers for a given course from different years, and conclusions about relative standards are drawn. Cross-moderation exercises have a long history in the monitoring and adjusting of assessment materials and arrangements.

An early version of a cross-moderation exercise to compare standards in A-level mathematics was reported by Christie and Forest (1980). In essence, the grade awarding processes of the time were repeated using scripts from 1963 and 1973, and examiners were asked to mark and grade 30 scripts from the two time points. Christie and Forest reported that "there is some evidence for a downward shift in standards" (p. 41) but the findings were equivocal and "any shift is negligible" (p. 41). Moreover, the results were undermined by the authors' conviction that the examiners' task was unwieldy due to the necessity to mark and grade against the explicit criteria of 1973. This resulted in student achievement from 1963 being ignored as it did not fit the criteria of 1973 following changes in the syllabus in the intervening years.

A second cross-moderation exercise to investigate changes in standards in A-level mathematics was reported by Quinlin (1995). Experts judged whether boundary scripts (at grades A/B, B/C and E/N, where N is a fail) from syllabuses in 1989 and 1994 were better than, worse than, or typical of a given boundary region. Quinlin reported an overall decline in standards over this period: "at all three boundaries the 1989 syllabus was seen as severely graded compared to the 1994 syllabus" (p. 25). However, similar limitations to those reported by Christie and Forest (1980) threatened the validity of the study.

Common test.

Common test approaches involve administering a test, which remains constant over time, to subsequent cohorts and comparing achievement (Murphy, 2007). If the grades awarded to candidates who achieved the same score on a test increase over

time, then standards may be interpreted to be falling. This is the basis of the approach adopted by international surveys such as PISA and TIMSS (NCES, 2006).

Two common tests have provided evidence about standards over time in A-level mathematics. Coe (2007) reported the relationship between the Test of Developed Abilities (TDA) and A-level grades between 1988 and 2006. The results suggested that A-level mathematics standards fell by over three grades during this period, substantially more than other A-level subjects, which fell by about two grades. However, Coe reported threats to the validity of this finding arising from substantial changes in mathematics syllabuses, as well as slight modifications to the TDA.

Lawson (2003) reported the relationship between a diagnostic test, administered to students embarking on Higher Education courses at an English university in 1991 and 2001, and A-level mathematics grades. He concluded that a grade N in 1991 (a fail) was equivalent to a grade B in 2001, an apparently steep decline. However, Lawson cautioned that this finding was "not evidence that A-level standards have fallen" (p. 174). The diagnostic test did not change in the intervening time, but in 1991 it was administered only to students "at risk" in terms of previous mathematical achievement, whereas in 2001 it was administered to all students studying subjects involving substantial mathematics.

Expert review.

Expert review approaches combine cross-moderation exercises with examiner judgement of syllabuses, examination papers and other documents. For example, Quinlin's (1995) study involved the cross-moderation exercise summarised above and a review of syllabus documents and examination questions. The syllabus review

reflected the decline in standards perceived in the cross moderation exercise. However, there were no perceived changes in the demand of examination questions.

In 1996, the United Kingdom government sponsored five-year reviews of standards in A levels following the Dearing review (National Committee of Inquiry into Higher Education, 1997). The reviews sought to establish whether the demand of syllabuses and examination papers, and the grading of candidate work, had changed over time. The reviews produced outcomes relating to A-level mathematics covering three periods. For 1995 to 1998 (QCA, 2001), experts reviewed syllabus documents, examination papers, mark schemes, examiner reports and sample candidate scripts. The review reported a slight decline over the period, which was particularly notable for one examination publisher. For 1998 to 2004 (QCA, 2006b), a time of substantial change to A-level mathematics including a shift from synoptic to modular assessments, the review concluded that standards of performance had been maintained. For 2004 to 2007 (Ofqual, 2009), also a time of substantial change to A-level mathematics, the review again concluded that standards had been maintained. Together, the reviews offer no clear evidence for a decline in standards from 1998 to 2007.

However, the reviews suffered limitations and warned that the experts involved lacked confidence in their judgements. Particular concerns related to the numerous and substantial changes to content and syllabuses over the period, the limited quantity and range of sample candidate scripts, and the difficulty of making complex decisions against grade descriptors (QCA, 2006a). The reviews have now been discontinued.

Comparative judgement.

Expert judgement plays a role in some of the approaches to described above. Here we use the term comparative judgement to refer specifically to experts judging directly the merit of a candidate's work relative to other candidates' work (Pollitt, 2012). It is the approach adopted in the present study.

In a comparative judgement exercise, subject experts are presented with pairs of scripts and asked to decide, based on the evidence before them, which script is over a higher standard. The decision is based on a global criterion such as "the better mathematician". The outcomes of many such decisions from several experts are then statistically modelled to produce a relative parameter estimate of the "quality" of each script. The parameter estimates can then be used to construct a scaled rank order of scripts from "highest" to "lowest" quality. Scripts at specified grades from different cohorts are included, and their relative positions in the final rank order used to draw inferences about changes in standards.

A key advantage of the comparative judgement approach is that candidate responses to different questions are compared directly. Human beings have been shown to be more reliable at comparing one object relative to another than they are at evaluating an object in isolation (Thurstone, 1927). That is, experts are more consistent when judging one script relative to another than when judging a lone script against grade descriptors. In addition, the use of collective expertise to construct a single rank order means the relative severity or leniency of individual experts is cancelled out, and the statistical modelling allows the precision of the estimates to be quantified (Pollitt, 2012). Comparative judgement approaches prevail today as the methodology underpinning comparability studies in England and Wales (Bramley, 2007; Bramley & Gill, 2010).

Comparative judgement was applied to the comparison of syllabuses from 1986 and 1995 in A-level mathematics (Bramley, Bell & Pollitt, 1998). For each syllabus, 10 scripts were selected from each of the A, B and E boundaries in 1986 along with 7 scripts around each of those boundaries in 1995. Two panels of experts, who were familiar with the examinations, undertook the judging. The authors reported evidence of a small decline in standards at Grade A in one syllabus. However, the study suffered from limitations. For example, although the judges were not informed of the purpose of the study, they correctly inferred it. We return to limitations with this and other comparability studies below.

Summary of standards over time in A-level mathematics.

In the remainder of the article we present a study that used a comparative judgement method, which was adapted to minimise the limitations associated with previous work. First we synthesise the findings reported above to provide an overview of evidence for changes in A-level mathematics standards.

An incomplete and not entirely coherent picture can be presented from 1963 to 2007. From the 1960s into the 1970s there appears to have been a slight decline in standards (Christie & Forest, 1980). No data are available for the period from the early 1970s to the late 1980s. Between the late 1980s and the mid-2000s there may have been a sharp decline, perhaps equivalent to over three grades (Coe, 2007; Lawson 2003). If so, it is difficult to pin down how and when this decline took place. There appears to have been some decline between the late 1980s and mid-1990s (Bramley, Bell & Pollitt, 1998; Quinlin, 1995), and slight further decline into the late 1990s (QCA, 2001), but results are nuanced and equivocal. Recent studies provide

little evidence for a decline from the late 1990s until the mid-2000s (QCA, 2006b; Ofqual, 2009), despite substantial changes to syllabuses and assessments. We are not aware of published studies into changes in the standards of A-level mathematics since the mid-2000s.

The study

The present study used a comparative judgement approach to investigate standards over time. We sought to minimise limitations to previous comparability studies that used one or more of the four approaches described above, while paying particular attention to limitations of comparative judgement studies as summarised by Bramley (2007). We list relevant limitations here and describe how they were minimised, before going on to present the study and its findings.

First, many comparability studies span a relatively short time period, as was the case for most of the research summarised above. While it is possible to attempt to synthesise findings over a longer duration the outcome is not necessarily entirely coherent, as we reported above. This is likely to arise in part from a cumulative effect of the threats to validity across studies using widely different assumptions and approaches. To help minimise this problem, we obtained an historic archive of candidates' examination responses from the mid-1960s to 2012, a broader span than has previously been used. Nevertheless, the archive available was incomplete as detailed below, with no scripts available for the 1970s and 1980s. Consequently, a detailed picture of changes over time was not possible.

Second, examiners are usually familiar with the examination materials being judged, and are likely to infer the purpose of a study, as was the case for the research

reported by Bramley, Bell and Pollitt (1998). This can introduce bias, and in particular the belief that recognisably older materials are of higher standard might influence participants' judgements. To minimise this we recruited research mathematicians rather than professional examiners. They were from various countries and so did not have a shared knowledge or expectation of A-level mathematics from their own schooling experience. As such our judges were content experts largely unfamiliar with A-level examination papers. In addition, the mathematicians were kept naive to the purpose of the study, and a post-questionnaire demonstrated that none inferred the purpose. We are therefore confident that bias based on preconceptions about the perceived age of examination papers was eliminated.

Third, bias might be introduced by differences in superficial appearance such as the typeset of examination questions and candidates' handwriting. In contrast to previous studies, all questions were re-typeset and all responses rewritten by a researcher to eliminate bias based on superficial appearance.

Fourth, comparability studies have traditionally focused on changes in grade boundaries, and accordingly used candidate scripts on or near grade boundaries. However, this reduces confidence that a candidate awarded, say, a grade A was genuinely representative of a grade A candidate for that examination. To maximise confidence in our findings we selected only those candidates who had been awarded "secure" grades.

Fifth, a well-known threat to validity is that an expert's judgement of a candidate's response can be biased by the particular test question (Good & Cresswell, 1988). Consider an extreme case in which two candidates of equal ability have attempted different questions. Candidate A attempted a demanding question and wrote very little; Candidate B attempted less demanding question and answered in

full. We might expect an expert judge to deem Candidate B the "better mathematician". If, however, both candidates answered the questions correctly and in full we might expect an expert judge to deem Candidate A, who attempted the more demanding question, to be "better" than Candidate B. This scenario would provide a more accurate reflection of Candidate A achieving a higher standard of mathematics than Candidate B. In order to investigate judge bias towards less demanding questions, we undertook a small follow-up study in which experts judged the responses of a fictitious "perfect" candidate who scored full marks on every question in the main study. If the results from the main study reflect a genuine change in standards at specific grades, then we should expect judgements of a single "perfect" candidate to produce the same broad pattern of results. That is, the relative standard of each year of examination from the main study (real candidates) and the follow up study (fictitious candidate) should be the same. If, however, those candidates judged better mathematicians in the main study were in fact those who tackled the easier test questions, consistent with the Good and Cresswell hypothesis, then we would expect symmetrical results from the judgements of a perfect candidate. That is, the relative standard of each year of examination from the main study (real candidates) and the follow up study (fictitious candidate) should be reversed.

The above research design decisions were taken to minimise, and in some cases eliminate, specific threats to the validity of previous comparability studies. However, it is not possible to avoid all problems and some scholars have argued that the comparison of standards over time is a flawed enterprise because of changes in curricula (Coe, 2010), changes in what is valued by the community of experts (Cresswell, 1996), and even because changes in standards are uninteresting

(Goldstein, 1979). We consider limitations of the present study, and indeed attempts to compare standards over time per se, later in the article.

Method

Materials. We obtained a historic archive of graded candidate responses, referred to as *scripts* here, to A-level mathematics examinations. The archive, obtained from Ofqual, the regulator for A-level qualifications in England, was somewhat sparse and contained just 66 pure mathematics examination scripts appropriate for inclusion in the study. These scripts were at grades A, B and E for the examinations published in 1964, 1968, 1996 and 2012, as shown in Table 1. The grade scale has remained largely unchanged over this period, with A being the highest grade and E being the lowest passing grade, although a new highest grade, A*, was introduced in 2010. An example pairing of questions and responses is shown in Figure 1.

The criteria for candidate selection were that grades were available and the qualification gained was in pure mathematics. We required the same grades at each year to enable a direct comparison, although no B grade papers for 1996 were available. All candidates sat two papers and one paper per candidate was selected, specifically that considered "standard" A-level mathematics: Paper 1 in Pure Mathematics for 1964, 1968 and 1996, and Unit Pure Core 4 for 2012. Only papers graded unambiguously (not near the boundaries) were included. The questions were re-typeset and the responses rewritten by a single researcher for uniformity. All marks, examiner comments and candidate details were omitted. The 66 scripts were spliced into 546 individual question responses. Perhaps surprisingly given the

timespan covered, no questions needed to be excluded on the grounds of archaic phraseology or contexts.

Participants. Twenty judges were recruited from the cohort of mathematics PhD students at Loughborough University, and were paid for their time. They were required to complete a two-hour examination made up from questions sampled from the examination papers used in the studies, and only those achieving >70% were allowed to undertake judging (all passed).

Procedure. The judging was conducted using an online comparative judgement system (www.nomoremarking.com). Judges received a unique url to access 250 pairwise comparisons of question responses via the comparative judgement website. 5000 pairwise judgement decisions were collected in total. A one-page user guide was sent to the judges, instructing them to decide, for each pairing, "which student you think is the better mathematician". Judges completed the judging online within a two-week window.

Judges were blind to the research aims. Upon completion of the judging they responded to an online survey that included the following two questions designed to check whether any had guessed or inferred the purpose of the study:

1. The questions you judged came from three different A-Level equivalent syllabuses. Although you did not know which questions came from which examinations, did you think that there were differences in difficulty between the syllabuses? If so, and if you can, please speculate on what you think might be behind these differences.

2. Is there anything else you would like to say about your experience of the judging process? (Optional)

All judges responded and none suggested that the study might have been about standards over time.

Analysis and results

The 5000 judgement decisions were modelled using the Bradley-Terry 2 package in the statistical software *R* (Firth, 2005), which assigned a parameter estimate and standard error to each question response (see Appendix for technical details on the modeling procedure). Each parameter estimate represented the relative quality of the question response, ranging between 0 (low ability) to 4.5 (high ability).

Preliminary analysis was conducted to ensure the coherence of the data. First, parameter estimates were used to construct a scaled rank order of questions from "best" to "worst". The internal reliability of the scaled rank order of responses was checked by calculating the Scale Separation Reliability (SSR, a measure of the "separatedness" of parameter estimates), judges' misfit figures and question-response misfit figures (measures of the consistency of the judging) (Pollitt, 2012). The SSR was .80, one of the 20 judges was a marginal misfit, 17 of the 546 question responses were marginal misfits and one question response was a large misfit. Overall these measures suggest the internal consistency of the outcome of the modelling procedure was acceptably high.

The mean of the parameter estimates of all the questions completed by a given candidate was then calculated to produce an overall "ability" parameter for that candidate. To address the question of whether the grading of A-level mathematics

examinations was perceived to have varied over time, a mean parameter estimate was calculated for the six scripts for each grade at each year, as shown in Table 1 and Figure 2. As can be seen, the mathematical performance of candidates, as collectively perceived by the experts, appears to have declined overall since the 1960s. We conducted a multiple regression analysis on the scripts' mean parameter estimates, using grades (E = 1, B = 4, A = 5) and year of examination (1964, 1968, 1996, 2012) as predictors. The full regression analysis is shown in Table 2. The analysis revealed that grade and year explained 74.8% of the variance in the scripts' mean parameter estimates, $F(2,63) = 93.56$, $p < .001$. Grade was a significant predictor of parameter estimate, $\beta = .69$, $p < .001$, and, crucially, so was year of examination, $\beta = -.51$, $p < .001$.

The multiple regression analysis confirmed a perceived overall decrease in mathematical performance, but did not establish when this decrease occurred. An inspection of Figure 2 indicates that the perceived decrease appears to have happened between 1968 and 1996, and there appears to have been little perceived variation in mathematical performance between 1964 and 1968 or between 1996 and 2012. To investigate this further we compared the parameter estimates of A and E grade scripts from 1996 and 2012 (we eliminated 2012 B grade scripts from this analysis as B graded scripts were missing from the 1996 archive).

The mean parameter estimate of grade A and E scripts from 1996 was -0.843 compared to -0.554 for the equivalent scripts from 2012, a mean difference of 0.289 95% CI [-0.540, 1.118]. This difference did not approach significance, $t(22) = 0.724$,

p = .477, and, critically, was in the opposite direction to that predicted by those who believe that standards have declined.[1]

To summarise, we found evidence of a decline in A-level mathematics standards since the 1960s, but this decline appeared to have taken place between 1968 and 1996. Contrary to some suggestions in the popular press, no evidence that standards have declined since 1996 was found.

Question demand

Earlier in the article we discussed a common threat to validity arising from expert judgements being influence by the demand of different test questions (Good & Cresswell, 1988). To explore whether question demand biased the judges in the main study we conducted a follow-up study using the same test questions. The responses, which we produced, were all "perfect", that is every response was worth full marks. This was intended to minimise the bias that can arise due to variation in question demand: there was effectively only one candidate who answered every question accurately and in full. Otherwise the study design was identical to that of the main study.

We expected to replicate the same broad pattern of results as for the main study, thereby suggesting that question demand did not bias the judges and so did not invalidate interpreting the findings as a reflection of changing standards. If, conversely, the Good and Cresswell hypothesis holds, and the outcome of the main study reflected question difficulty rather than candidates' mathematical performance, we would expect a symmetrical outcome (i.e. a reversal of the estimated relative

[1] Note that if a Bonferroni correction were applied to this comparison, then the result would be even further from the conventional level of statistical significance.

standard of each year of examination) from the judgement of responses worth full marks.

Method

Materials. None of the candidate responses in the main study were used in the follow-up study. Instead, for each question used in the main study (38 in total), we constructed a fictitious response based on the mark scheme (where available) and actual student responses that had scored full or almost full marks. Where mark schemes were not available, for the examination papers from 1964 and 1968, we drew on the number of marks available for different question parts and high-scoring student responses where possible. The fictitious responses included an extra ten questions compared to the main study, due to no candidates having responded to certain questions for examination papers that allowed a choice (1964, 1968). In total 48 "perfect" question responses were included. The fictitious responses were handwritten by a researcher for consistency.

Participants. Eighteen of the twenty judges involved in the main study were recruited, and were paid for their time.

Procedure. The procedure was identical for the main study, except that each judge completed 45 pairwise comparisons of question responses. 810 pairwise judgement decisions were collected in total. The judges were blind to the purpose of the study, which took place prior to completion of the online survey confirming that none guessed or inferred the purpose.

Analysis and results

The preliminary analysis steps of the main study were repeated. The SSR was .874, one of the 18 judges was a moderate misfit, one of the 38 question responses was a marginal misfit and one question response was a moderate misfit. These measures suggest the internal consistency of the outcome of the modelling procedure was acceptably high.

To compare the outcomes with the main analysis, the mean of the parameter estimates for a given year was calculated to produce an overall parameter for each year, shown in Figure 3. A comparison of Figure 2 and Figure 3 suggests that the same pattern of results was obtained for the fictitious perfect candidate as for the real candidates at grades A, B and E. Note that the modelling procedure produces relative not absolute parameter estimates, and so only the patterns of changes and not the raw values can be compared across Figure 2 and Figure 3.

To compare the results we calculated the Pearson correlation between the mean parameter estimates of responses to each question in the main study, and the parameter estimates for the same questions in the follow-up study. The correlation coefficient was high, $r = .680$, suggesting the main finding was replicated for the case of the "perfect" responses. To investigate this further we conducted a linear regression analysis using year of examination (1964, 1968, 1996, 2012) as the predictor, shown in Table 3. The model explained 27.9% of the variance in the parameter estimates of the "perfect" responses, $F(1,36) = 13.94$, $p = .001$. As with the main analysis, year was a significant predictor of parameter estimate, $\beta = -0.53$, $p = .001$, and this effect appeared to be driven by changes between 1968 and 1996, not by changes since 1996.

Overall, the findings for a fictitious "perfect" candidate replicated those for the real grade A, B and E candidates in the main study. This supports the main finding that judges perceived candidates who sat papers in the 1960s to have performed better than candidates who sat papers in the 1990s and 2010s. Specifically, it reduces the possibility that the results of the main study were distorted due to candidates of similar ability being perceived as markedly different due to the effects of question demand. Rather, the results appear to have arisen due to genuine changes in standards over time.

Discussion

We investigated whether mathematics experts perceived a decline in standards of A-level mathematics examinations over the previous five decades. Our results suggest that higher grades were awarded for perceived lower mathematical performance in the 1990s and 2010s compared to the 1960s. Furthermore, the results provide an indication of the extent and period of the perceived decline in standards. A candidate who achieved a grade B in 1996 or 2012 appears to have been perceived by experts to have performed approximately at the level of a candidate who achieved a grade E in 1964 or 1968. However, we found no evidence of a perceived overall decline since the mid-1990s.

These findings offer some consistency with the picture presented by previous research into standards in A-level mathematics over previous decades. Specifically, the lack of evidence for a decline in standards from the mid-1990s to the near present is consistent with government-sponsored reviews (QCA, 2001; QCA, 2006b; Ofqual, 2009). It is also consistent with research suggesting some decline before the mid-

1990s (Bramley, Bell & Pollitt, 1998; Quinlin, 1995), although our script sample does not allow us to comment when this might have occurred since 1968. However, while our results support the direction of change reported from the 1980s into the 2000s by Coe (2007) and Lawson (2003), they cast doubt on the suggested steepness of the decline. Whereas our results support a decline of around three grades over almost five decades, we found no evidence to support a decline of more than this between the late 1960s and mid-1990s.

One reason for this difference may be that standards rose and declined at different times since the 1960s. If so, then our finding is not necessarily inconsistent with Coe (2007) and Lawson (2003). However, another reason may be that the present study avoided some of the concerns raised by Coe and by Lawson about validity of their methods. The common test measures reported by Coe (2007) and Lawson (2003) are proxies for performance (Murphy, 2007), whereas the approach adopted here makes use of direct evidence of performance and collective expert judgement (Bramley, 2007). Furthermore, a novel design enabled the minimisation of threats to validity that have limited the interpretation of findings from previous studies, as described above.

Few will be surprised at the perceived decline in standards we have reported between the 1960s and the mid 1990s. However, in light of high-profile concern about standards, some may be surprised at the lack of evidence of a perceived decline since the 1990s. In fact, despite these high-profile concerns (e.g. Willis & Paton, 2009), confidence in A levels among teachers, students, parents and the general public has remained stable or even increased over recent years (QCA, 2006c). The regulator of examinations in England, Ofqual, recently concluded that the A-level system needed to be redesigned because the contents of the examinations needed to be revised to

ensure they were consistent with current requirements from Higher Education and employers (Ofqual, 2014). Modularity, resitting and the culture of teaching to the test are key points of issue (Simpson & Baird, 2013), but the majority of people "agree that most students taking A levels get the grades their performance deserves" (QCA, 2006c, p. 9).

Our results support a perceived decline in standards, but we collected no data as to why the decline might have occurred. Various mechanisms have been put forward. For example, Stringer (2012) suggested that examiners give candidates the benefit of the doubt and that this might have cumulatively led to lower performance receiving the same grades in subsequent years. Others have suggested market and accountability pressures conspire to incentivise examination publishers to design piecemeal, predictable examinations that are easier to prepare for (e.g. Mansell, 2007). In all likeliness these and other mechanisms contributed, in part at least, to the perceived decline in standards reported here.

Limitations

We sought to minimise sources of bias and other threats to validity as detailed above. However, as with any standards comparability study, these could not be entirely eliminated and the findings require careful interpretation (Baker, Sutherland & McGraw, 2002).

A key limitation of our study is that we were only able to take occasional snapshots of performance due to the paucity of the historic archive. As such very few papers were analysed within the time points studied. More papers would enable more confidence in our reported findings. A further consequence of this paucity is that our

regression line presents a smoother picture than may be the case. For example, it brushes over the A-level mathematics problems in 2002 brought about by a radical change in the syllabus structure, commonly referred to as *Curriculum 2000*. More data points may reveal a more nuanced picture.

Another limitation is that comparative judgement harnesses contemporary perceptions to determine the quality of candidates' work. That which the community of experts, in our case research mathematicians, value now may differ from what they valued in the past. As such, there is no objective way in which comparisons over time are possible (Cresswell, 1996). Therefore all statements of changes in examination performance must be seen through the lens of contemporary, if expert, value judgements.

This limitation means it is important to be clear about what comparative judgement methods can and cannot demonstrate when applied to monitoring standards over time. We have provided evidence based on a group of contemporary mathematicians' perceptions of the relative performance of a sample of candidates experiencing different syllabuses and assessment arrangements. This enabled us to draw conclusions about standards over time as perceived from a contemporary vantage point, but not to fully detangle how contemporary values and changes in syllabuses and assessment arrangements may have shaped the final outcome.

Finally, we should address Goldstein's (1979) objection, that pursuing questions of standards over time is a pointless exercise, not only because questions cannot be answered satisfactorily, but also because the answers are uninteresting. Surely the most important question to be asked of a qualification is whether it is fit for purpose now. However, in the absence of rigorous studies in this area the examination system is left to the mercy of opinion and speculation. Standards are important, and

given teachers and parents express concern about the way the debate is discussed in the popular media (QCA, 2007), impartial and rigorous evidence, along with important caveats about interpretation, is important. The worth of such evidence is less the development of an understanding of whether the currency of qualifications have changed, and more the way in which a suspicion of drifting standards in the past has real implications for the way in which the examination system is run today.

Conclusion

We have presented evidence based on perceived changes in standards over time to A-level mathematics. Ideally, a more complete archive of graded scripts would have been obtained, enabling a more detailed picture to emerge. It is possible such an archive exists, and if so a follow-up study would be very worthwhile, but to the best of the authors' knowledge further graded scripts filling the gaps are unlikely to be found. This highlights the importance of systematically archiving sample scripts (Robinson, 2007) to enable comparative judgement studies into standards over time in the medium- to long-term future.

Finally, the approach used here can readily be applied to other mathematics qualifications, and other subject disciplines. If this is done, then it is important that similar steps to minimise threats to validity are adopted to maximise confidence in results, and that findings are cautiously interpreted.

Acknowledgements

This work was supported by a Royal Society Shuttleworth Research Fellowship to IJ, a Royal Society Worshipful Company of Actuaries Research Fellowship to MI, and AQA. We would like to thank Dr Michelle Meadows of Ofqual for commenting on

early drafts of this work, and Ofqual for providing access to their historic archive of examination scripts.

References

ACME (2011). *Mathematical Needs: Mathematics in the Workplace and in Higher Education*. London, UK: Advisory Committee on Mathematics Education.

Adams, R. (2007). Cross-moderation methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 212-245). London: Qualifications and Curriculum Authority.

Baker, E., Sutherland, S. & McGraw, B. (2002). *Maintaining GCE A Level Standards: The Findings of an Independent Panel of Experts*. London: Qualifications and Curriculum Authority.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 246-294). London: Qualifications and Curriculum Authority.

Bramley, T., Bell, J. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives, 25*, 1–24.

Bramley, T. & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, *25*, 293–317.

Christie, T. & Forrest, G. M. (1980). *Standards at GCE A-level: 1963 and 1973*. London: Macmillon Education.

Coe, R. (2007). *Changes in Standards at GCSE and A-level: Evidence from ALIS and YELLIS*. Durham: Centre for Curriculum, Evaluation and Management, Durham University.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, *25*, 271–284.

Cresswell, M. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, Developments and Statistical Issues* (pp. 57-84). Chichester, UK: John Wiley.

Cresswell, M. (2003) *Heaps, Prototypes and Ethics: The Consequences of Using Judgements of Student Performance to Set Examinations Standards in a Time of Change*. London: Institute of Education.

Croft, A. C., Harrison, M. C., & Robinson, C. L. (2009). Recruitment and retention of students: An integrated and holistic vision of mathematics support. *International Journal of Mathematical Education in Science and Technology, 40*, 109–125.

Engineering Council. (2000). *Measuring the Mathematics Problem*. London: Engineering Council.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, *12*, 1–12.

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

Goldstein, H. (1979). Changing educational standards: A fruitless search. *Journal of the National Association of Inspectors and Educational Advisers*, *11*, 18-19.

Hawkes, T., & Savage, M. D. (2000). *Measuring the Mathematics Problem*. London, The Engineering Council.

Isaacs, T. (2014). Curriculum and assessment reform gone wrong: The perfect storm of GCSE English. *Curriculum Journal*, *25*, 130-147.

Lawson, D. (2003). Changes in student entry competencies 1991–2001. *Teaching Mathematics and Its Applications*, *22*, 171–175.

Mansell, W. (2007). *Education by Numbers*. London: Politicos Publishing.

Murphy, R. (2007). Common test methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 301-323). London: Qualifications and Curriculum Authority.

National Committee of Inquiry into Higher Education (1997). *Higher Education in the Learning Society: Report of the National Committee of Inquiry into Higher Education.* London: HMSO.

NCES (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments*. Technical Report NCES 2006-029. Washington, DC: National Center for Education Statistics.

Oates, T. (2011). Could do better: Using international comparisons to refine the National Curriculum in England. *Curriculum journal*, *22*, 121-150.

Ofqual (2009). *Review of Standards in GCE Mathematics in 2004 and 2007,* Technical Report Ofqual/09/4154. Coventry: Ofqual.

Ofqual (2011). *GCSE, GCE, Principal Learning and Project Code of Practice,* Technical Report Ofqual/11/4850. Coventry: Ofqual.

Ofqual (2014). *Perceptions of A Levels, GCSEs and Other Qualifications in England – Wave 12*, Technical Report Ofqual/14/5518. Coventry: Ofqual.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, *19*, 281–300.

QCA (2001). *Five Year Review of Standards: A level Mathematics*, Technical Report QCA/01/764. London: Qualifications and Curriculum Authority.

QCA (2006a). *QCA's Review of Standards: A Description of the Programme,* Technical Report QCA/06/2348. London: Qualifications and Curriculum Authority.

QCA (2006b). *Review of Standards in Mathematics: GCSE 1999–2004 and A level 1998–2004,* Technical Report QCA/06/2348. London: Qualifications and Curriculum Authority.

QCA (2006c). *GCSEs and A level: The Experiences of Teachers, Students, Parents and the General Public*. London: Qualifications and Curriculum Authority.

Quinlin, M. (1995). *A Comparability Study in Advanced Level Mathematics. A Study Based on the Summer 1994 and 1989 Examinations*. London: University of London Examinations and Assessment Council.

Robinson, C. (2007). Awarding examination grades: Current processes and their evolution. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 97-123). London: Qualifications and Curriculum Authority.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Simpson, L., & Baird, J. A. (2013). Perceptions of trust in public examinations. *Oxford Review of Education*, *39*, 17-35.

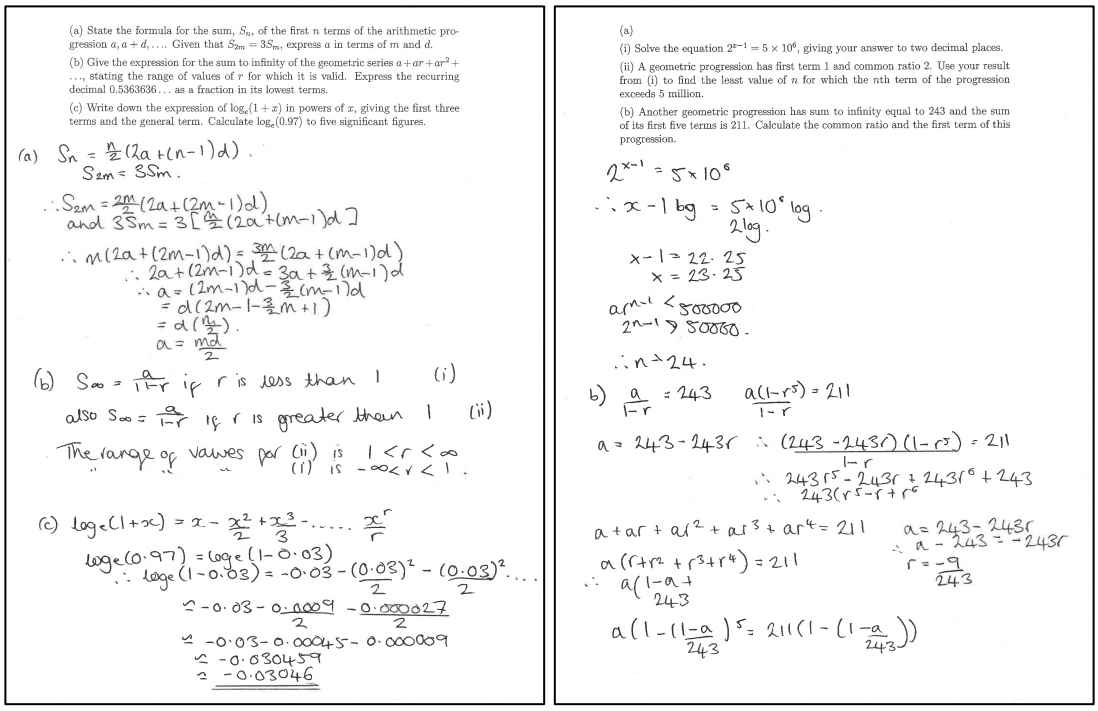
Stringer, N. S. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, *27*, 535–554.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Walport, M., Goodfellow, J., McLoughlin, F., Post, M., Sjøvoll, J., Taylor, M., & Waboso, D. (2010). *Science and Mathematics Secondary Education for the 21st Century: Report of the Science and Learning Expert Group*. London: Crown.

Wheadon, C. (2013). Using modern test theory to maintain standards in public qualifications in England. *Research Papers in Education*, *28*, 628–647.

Willis, A. & Paton, G. (2009). A-levels: Row over maths standards. Retrieved from http://www.telegraph.co.uk/education/6067901/A-levels-Row-over-maths-standards.html



(a) State the formula for the sum, S_n, of the first n terms of the arithmetic progression $a, a+d, \ldots$. Given that $S_{2m} = 3S_m$, express a in terms of m and d.

(b) Give the expression for the sum to infinity of the geometric series $a + ar + ar^2 + \ldots$, stating the range of values of r for which it is valid. Express the recurring decimal $0.5363636\ldots$ as a fraction in its lowest terms.

(c) Write down the expression of $\log_e(1+x)$ in powers of x, giving the first three terms and the general term. Calculate $\log_e(0.97)$ to five significant figures.

(a) $S_n = \frac{n}{2}(2a+(n-1)d)$.
$S_{2m} = 3S_m$.

$\therefore S_{2m} = \frac{2m}{2}(2a+(2m-1)d)$
and $3S_m = 3\left[\frac{m}{2}(2a+(m-1)d\right]$

$\therefore m(2a+(2m-1)d) = \frac{3m}{2}(2a+(m-1)d)$
$\therefore 2a+(2m-1)d = 3a+\frac{3}{2}(m-1)d$
$\therefore a = (2m-1)d - \frac{3}{2}(m-1)d$
$= d(2m-1-\frac{3}{2}m+1)$
$= d(\frac{m}{2})$.
$a = \frac{md}{2}$

(b) $S_\infty = \frac{a}{1-r}$ if r is less than 1 (i)
also $S_\infty = \frac{a}{1-r}$ if r is greater than 1 (ii)
The range of values for (ii) is $1 < r < \infty$
" " " (i) is $-\infty < r < 1$.

(c) $\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \frac{x^r}{r}$
$\log_e(0.97) = \log_e(1-0.03)$
$\therefore \log_e(1-0.03) = -0.03 - \frac{(0.03)^2}{2} - \frac{(0.03)^2}{2} \ldots$
$\simeq -0.03 - \frac{0.0009}{2} - \frac{0.000027}{2}$
$\simeq -0.03 - 0.00045 - 0.000009$
$\simeq -0.030459$
$\simeq -0.03046$

(a)
(i) Solve the equation $2^{x-1} = 5 \times 10^6$, giving your answer to two decimal places.
(ii) A geometric progression has first term 1 and common ratio 2. Use your result from (i) to find the least value of n for which the nth term of the progression exceeds 5 million.
(b) Another geometric progression has sum to infinity equal to 243 and the sum of its first five terms is 211. Calculate the common ratio and the first term of this progression.

$2^{x-1} = 5 \times 10^6$
$\therefore x-1 \log 2 = 5 \times 10^6 \log$.
$x-1 = 22.25$
$x = 23.25$

$ar^{n-1} < 5000000$
$2^{n-1} > 5000000$.
$\therefore n > 24$.

b) $\frac{a}{1-r} = 243$ $\frac{a(1-r^5)}{1-r} = 211$
$a = 243 - 243r$ $\therefore \frac{(243-243r)(1-r^5)}{1-r} = 211$
$\therefore 243r^5 - 243r + 243r^6 + 243$
$\therefore 243(r^5 - r + r^6$

$a + ar + ar^2 + ar^3 + ar^4 = 211$
$a(r+r^2+r^3+r^4) = 211$
$\therefore a(1-a+$
243

$a = 243 - 243r$
$\therefore a - 243 = -243r$
$r = \frac{-9}{243}$

$a(1-(1-\frac{a}{243})^5 = 211(1-(1-\frac{a}{243}))$

Figure 1: Example pairing of questions and responses. Experts were tasked to decide which candidate was "the better mathematician" of many such pairings.

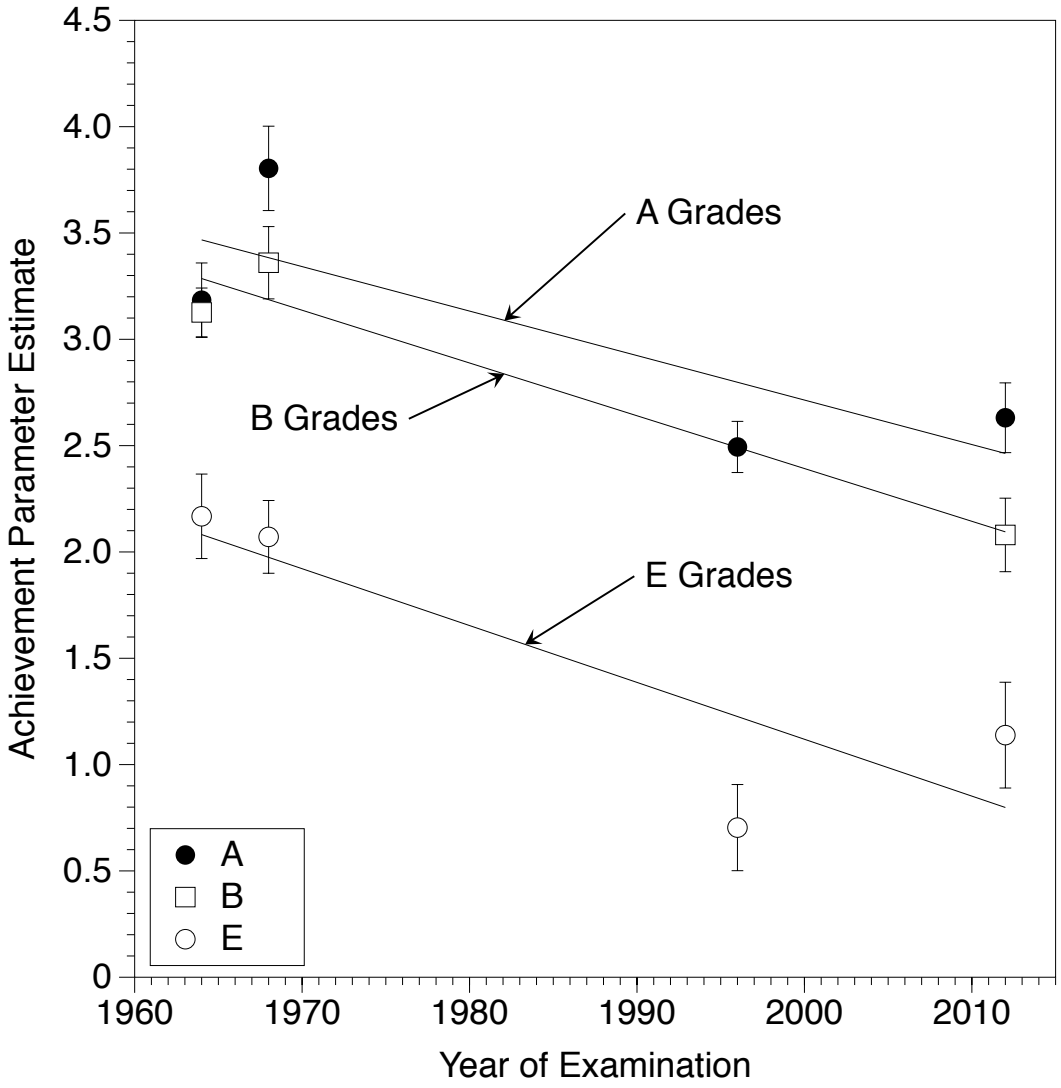


Figure 2: Mean parameter estimates for the six scripts at each grade and year. Error bars show ±1 SE of the mean.

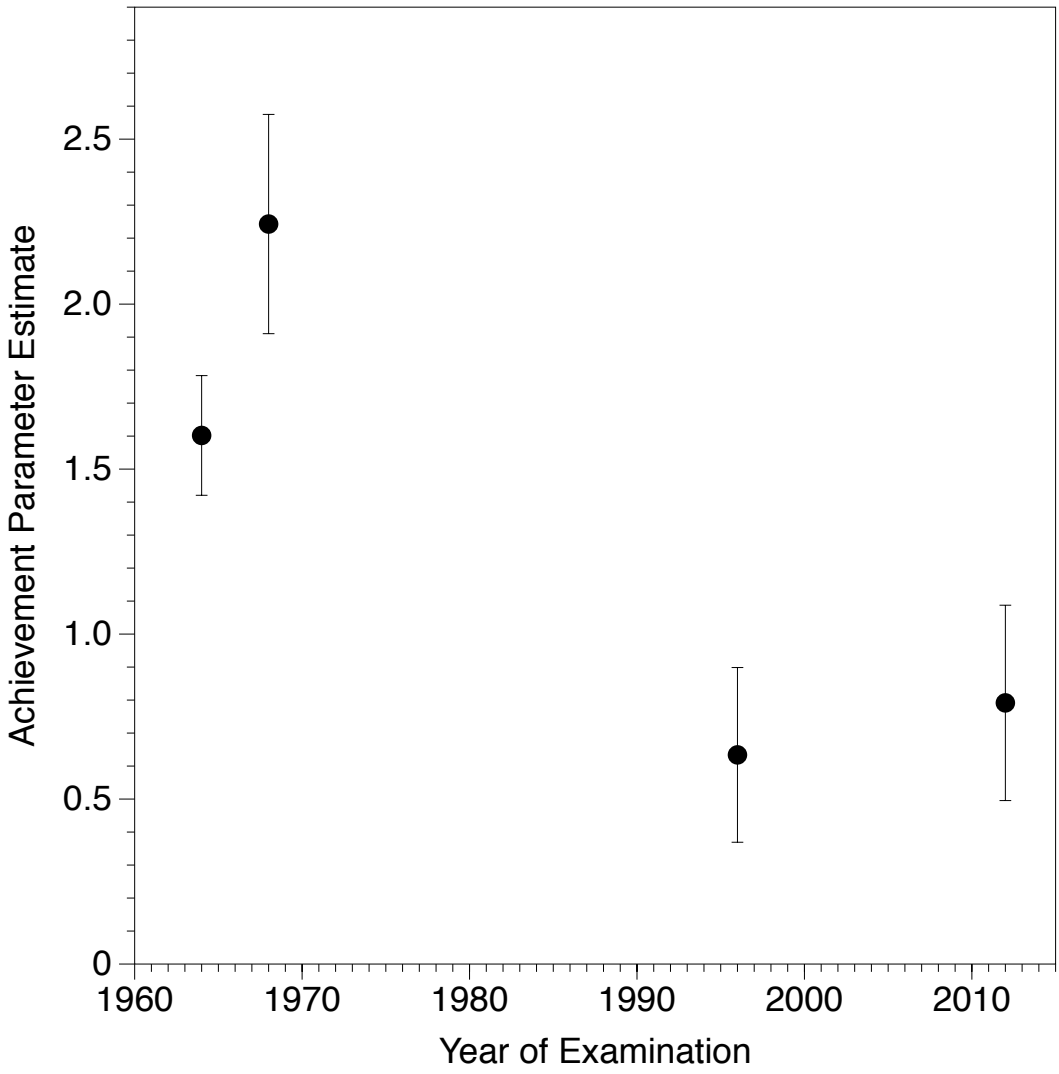


Figure 3: Mean parameter estimates for the "perfect" responses to the examination questions at each year. Error bars show ±1 SE of the mean.

Year	Exam board	Number of questions	Marks available	Grade		
				A	B	E
1964	JMB	Choose 7 of 12	98	6	6	6
1968	JMB	Choose 7 of 9	98	6	6	6
1996	AEB	13	100	6	-	6
2012	AQA	8	75	6	6	6

Table 1: Number of scripts obtained for the study at each year and grade. Note that grade B scripts were not available for 1996.

Variable	*B*	95% CI for *B*	*β*
Constant	47.5***	[35.4, 59.7]	
Year	-0.025***	[-0.031, -0.018]	-0.507***
Grade	0.389***	[0.318, 0.460]	0.690***
R^2	0.748		
F	93.56***		

Table 2: A regression model predicting the scripts' mean parameter estimates. CI = confidence interval, ***p<.001.

Variable	*B*	95% CI for *B*	*β*
Constant	58.7***	[26.8, 90.5]	
Year	-0.030***	[-0.046, -0.013]	-0.528***
R^2	0.279		
F	13.9***		

Table 3: A regression model predicting the perfect scripts' parameter estimates. CI = confidence interval, ***p<.001.