The use of evidence in setting and maintaining standards in GCSEs and A levels

Discussion paper

Tom Benton & Tom Bramley

Assessment Research and Development (ARD)
Cambridge Assessment

28th January 2015

University of Cambridge
Local Examinations Syndicate

1. Introduction

This paper argues the following:

- Although there are reasons not to rely upon expert judgement as the sole source of information on standards, recent reviews of the literature have overemphasised the unreliability of this source and tended to ignore more positive evidence.
- Several sources of evidence can legitimately contribute to the standard setting and maintaining process. They broadly fall into three categories: i) evidence about exam difficulty; ii) evidence about examinee ability; and iii) evidence about the quality of examinee work. These sources of evidence should be derived as independently of each other as possible.
- The totality of the evidence can then be used to derive grade boundaries, with a rationale for how much weight to give to each source. Although, at least initially, it would be prudent to give greater weight to statistical sources of evidence, wherever the decision was made to deviate from these the procedure by which such a decision had been reached should be transparent.
- Comparative judgement of relative script quality is preferable to absolute judgements of grade-worthiness in awarding. It can be made relatively independent of other sources of evidence by requiring examiners to make judgements about the relative quality of different scripts from different years blind to the number of marks that were awarded.

2. Critique of widespread reviews of the literature

Recent explanations by Ofqual of their approach to maintaining standards (e.g. Ofqual, 2014b) draw heavily upon (and to a large extent directly repeat) a section from a book chapter written by Baird (2007). The ways various pieces of research were summarised by Baird (2007) are given below. Alongside this we present comments on the caveats surrounding these various pieces of research, the different ways in which these research studies could be interpreted and elements of these same research studies that were not reported. As can be seen, the standard interpretation of the research evidence as showing that "it is difficult for awarders to make grade boundary judgements accurately and consistently simply by looking at students' work" (Ofqual, 2014a) is not universally supported. Furthermore, much of this literature is concerned with the use of expert judgement in the form of examiners directly grading scripts with different numbers of marks. As such, many of the findings would not apply to a comparative judgement exercise where examiners are rather directly instructed to consider the *relative* grade-worthiness of different groups of scripts.

Cresswell (2000)

Prevailing interpretation: When grade boundaries are set by expert judgement alone, year on year changes in pass rates are substantially bigger than we might expect. These changes are not obviously explicable in terms of changes in the cohort.

Comments: The research also suggests that examiners do have a good sense of the required direction of change in grade boundaries but under the previous system did not go far enough. One possible reason for this might be that they were (quite logically) giving some weight to the existing grade boundaries as a source of evidence as (presumably) exams are designed to be of equal difficulty in each year.

This raises two questions:

1. What role did existing grade boundaries play within the process and is there any possibility that this may have reduced the ability of examiners to provide fully independent judgements regarding appropriate grade boundaries?

2. Would it really be justified to give *zero* weighting to this source of evidence (existing boundaries) in favour of statistical predictions? Is it not reasonable that in many cases a sensible positioning of grade boundaries will fall somewhere between the two (albeit closer to the statistical predictions than was achieved in this paper).

It is also worth noting that this research also contained a substantial critique of the sole use of statistical information. Thus, the research does not recommend that statistics should entirely replace expert judgement but rather states that "the use of statistically informed qualitative judgement to set annual standards is entirely appropriate for the primary purpose of public examinations – the provision of qualifications" (page 98).

Good and Cresswell (1988)
Prevailing interpretation: Examiners cannot adequately compensate in their judgements of students' work for the demands of the question papers
Comments: This paper tested whether examiners could achieve comparability across papers with extreme differences in difficulty - essentially papers from different tiers within the same subject. Although it is clear that examiners struggled with this task it should be noted that:

1. This does not necessarily imply that examiners will be unable to make appropriate adjustments to grade boundaries when the difference in difficulty between two papers is small.
2. This type of "vertical equating" is also a very difficult task using statistical methods and is often somewhat controversial.
3. The study itself was done at a point when tiered examinations were just being introduced with the beginning of GCSEs. The paper itself does not expect the results that were found to be a permanent feature of the examination system but rather says "It is to be hoped that, once awarders had gained experience of differentiated examinations, they would be able to apply the new standards in the usual way" (p279).

Another section of the paper that is less frequently referenced is evidence on the consistency of grading decisions between different teams of examiners when they are employed in more typical awarding situations rather than trying to ensure comparability across tiers. This part of the work concludes "Although the grade awarding teams produced results which were not identical, within the context of public examining, the differences between them were not large." (page 276). This evidence of the frequent reliability of expert judgement is not recorded within the summaries by either Baird or Ofqual.

Baird (2000)
Prevailing interpretation: "examiners are setting standards …with reference to their own mental models of the standard" (Baird, 2007, page 142) rather than exemplars or benchmark scripts.
Comments: This finding will be highly dependent upon the way in which the task given to examiners is described. Specifically, if examiners were specifically asked to perform paired comparisons or a rank-ordering exercise then such a result should become far less likely. Furthermore, the results do not imply that examiners will provide unreliable evidence with respect to their own internal standard and thus cannot be trusted.

Baird and Scharaschkin (2002)
Prevailing interpretation: "examiners…make more severe judgements of candidates' work when they judge each question paper independently than when they judge all of their work for A level." (Baird, 2007, page 143).

Comments: This simply implies that comparative judgement between (for example) a single component and a full A level wouldn't work. This does not imply that comparative judgement cannot work at all but rather requires that the process is applied consistently.

Scharaschkin and Baird (2000)
Prevailing interpretation: "examiners are unduly influenced by the consistency of candidates' performances" (Baird, 2007, page 142) in that they award better grades on average to scripts with consistent performance rather than those where performance greatly varies across items.
Comments: This is an interesting finding but no problem to the idea of using comparative judgement to determine examination standards unless one set of scripts being compared displays greater "inconsistency" than another. Bramley (2012) also found an effect of the profile of marks on the perceived quality of performance and recommended that the scripts used in traditional grade awarding meetings should be selected to be representative of all scripts with the same total score in terms of profile of marks (and other features shown to influence judgements). The same advice would apply to selection of scripts for comparative judgement exercises.

Baird and Dhillon (2005) / Forster (2005)
Prevailing interpretation: Correlations between examiners' rank-orders of the quality of scripts and the marks awarded to scripts within a small range were close to zero. Specifically, "none of the 36 correlations calculated were statistically significant" (Baird, 2007, page 143). This implies that "asking awarders to distinguish qualitatively between scripts on adjacent marks is asking them to perform a nigh impossible task" (Baird and Dhillon, 2005, page 27).
Comments: The results in these papers concern the reliability of a single examiner within a tiny mark range. This does not tell us the reliability of examiners as a whole. Both Good and Cresswell (1988, see above) and Novakovic and Suto (2010) empirically show that these can be reliable and that different groups of examiners can often achieve reasonable consistency when their knowledge is pooled. A more extensive critique of the Baird & Dhillon paper can be found in Benton (2014).

Stringer (2012)
Prevailing interpretation: Not used by Baird/Ofqual but of a similar vein. The most interesting evidence comes from the account of an awarding committee failing to spot when the statistical predictions for a particular paper were calculated incorrectly, leading (initially) to badly biased results (although these were corrected before any results were issued).
Comments: It is interesting to note that the same paper also gives an example of where another awarding committee *did* correctly identify the same problem with statistical predictions. This paper perhaps highlights the extent of the current pressure on examiners not to depart too far from statistical predictions rather than their inability to reliably use their judgement when allowed to exercise this free from the influence of statistics. This emphasises the need to keep the sources of evidence relevant to standard maintaining independent of each other.

Impara and Plake (1988)
Prevailing interpretation: Referenced as showing "abundant evidence that examiners are not good at discerning the difficulty of questions" (Baird, 2007, page 141).
Comments: Although the study does indeed show major discrepancies between the percentage of minimally proficient candidates expected (by examiners) to correctly answer each question and the percentage that actually do, this study also shows that in terms of judging the *relative* difficulty of questions, judges actually perform quite well displaying a strong correlation (0.78) between judged difficulty and empirical difficulty. The authors concluded that "Thus, as in previous studies, the teachers' rank ordering of item difficulty was moderately accurate, even though their precision in estimating item difficulty was not" (page 76). Other studies reviewed in Brandon (2004) give similar results. Furthermore, in the

context of another assessment, the same authors state that judges' ratings of item difficulties were "both valid and reliable" (Plake, Impara, and Irwin, 1999).

Summary
As the above summaries show, although there are certainly reasons to avoid using expert judgement as the sole means of maintaining standards, much of the existing research used to decry the use of expert judgement is more positive about its potential than is often acknowledged. Furthermore, some of the criticisms of expert judgement (such as reliance on internal standards) in the above literature would not apply if expert judgement was employed in the form of comparative judgement.

3. Sources of evidence that have a legitimate role to play in standard maintaining

Standard maintaining is a difficult task because two things can change from exam session to exam session: the ability of the examinees and the difficulty of the exam papers[1]. If we could be sure that the exam papers were of the same difficulty as before we could use the same grade boundaries as before, and if we could be sure the examinees had the same ability distribution as before we could set the grade boundaries so as to give the same distribution of grades as before. Although this simplistic conception hides considerable problems (Bramley, 2013), it immediately suggests two relevant sources of evidence for decisions about grade boundaries: evidence about the difficulty of the exam papers; and evidence about the ability of the examinees.

However, a further distinction should be drawn between the ability of the examinees and the quality of the work they produce in the examination. This implies that independent evidence about the quality of work produced is also relevant to standard maintenance. For example, we could randomly divide a group of examinees into two (so the two groups have the same ability) and incentivise one group to perform better (e.g. by offering money for each mark gained). It is possible that the quality of work produced by the incentivised group would be higher and so they would 'deserve' better grades … or would they? When the difference between the groups is some factor other than motivation (e.g. the quality of teaching they have had, or the familiarity of their teachers with the exam syllabus) it is not at all clear whether the factor should be taken into account in the standard maintaining process. These problems are discussed in much greater detail by Newton (2010a,b).

Although the separation of examinee ability, exam difficulty, and quality of work seems very neat, we do recognise that it is not possible to completely separate these concepts – they are internally related. But that does not prevent us from categorising different sources of evidence as relating to one or other of them. Possible methods for obtaining these different sources of evidence are discussed briefly below.

1. Evidence about the difficulty of the exam papers
Exam papers are constructed to specifications that are intended to ensure that papers are of broadly equivalent difficulty from one session to the next[2]. Therefore previous grade boundaries are a relevant source of information that should be given some weight (Bramley, 2013). Furthermore, it is possible to use expert judgement of question difficulty to get a sense of at least the direction and possibly the magnitude of any changes in the overall difficulty of an exam paper. This can be done using relative judgements of question difficulty (Curcin, Black & Bramley, 2010), or by Angoff-style judgements of the performance of examinees on the grade boundaries, or by a form of quasi-common item equating using expert judgements of question similarity (Bramley & Wilson, in prep.). And of course, in

[1] Not meaning just traditional written papers, but assessment tasks in general.
[2] Of course, sometimes there is a deliberate intention to make a paper easier or harder than its predecessor, but we concentrate on the usual situation here.

situations where pre-testing is possible we can get statistical information about the difficulty of the questions.

2. Evidence about the ability of the examinees
This source of evidence, in the form of prior attainment (at Key Stage 2 for GCSEs and at GCSE for A level), has become the dominant source of evidence used to set grade boundaries at GCSE and A level in recent years. The idea is that if a group of examinees has better (or worse) prior attainment than the previous cohort, it is reasonable to expect them to get a better (or worse) distribution of grades. The 'comparable outcomes' method as applied in practice (see Benton & Lin, 2011 or Taylor, 2013 for technical details) is statistically very similar to one particular method from the test equating literature, with prior attainment category playing the role of an anchor test score. The implications of this are discussed in Bramley & Vidal Rodeiro (2014).

The comparable outcomes[3] approach was originally used by the regulator to maintain standards in times of change to the exam system (see Bramley, Dawson & Newton (2014) for a discussion of two high profile crises in England caused by different understandings of comparability of standards in times of change). Its purpose was to compensate examinees for the possibly lower quality of work they might produce as a result of being the first group to experience a system change (e.g. a change of syllabus, or a change of assessment style from linear to modular). The comparable outcomes approach gives a great deal of weight to evidence about examinee ability (as opposed to evidence about exam difficulty or quality of work). It is not so clear that this is appropriate in 'normal' circumstances when there is no change to the exam system. It has largely succeeded in reducing or eliminating grade inflation, but at the cost of being described by its critics as sophisticated norm-referencing.

However, considerations of the ability of the examinees have rightly played a part in standard maintaining, long before the comparable outcomes approach appeared. As shown by Newton (2011), A levels and GCSEs have never been strictly norm-referenced, but rather the common sense idea that if the cohort of examinees seems similar to the previous cohort then we might expect the grade distribution to be similar has always been a general principle taken into account when setting grade boundaries. Before the availability of the longitudinal national databases allowing linking to prior attainment, a variety of other indicators of the ability of the examinee cohort were (and in some circumstances still are) used to inform judgements about how similar the cohort of examinees might be. These include information about the proportion of the examinees from different types of school, and the schools' forecast grades (or changes in the distribution of forecast grades). Also, the performance of examinees in the subset of schools taking an exam in two years ('common centres') may be considered, on the assumption that these subsets of examinees are of similar ability.

3. Evidence about the quality of work produced by examinees
If examination grades are to have a valid criterion-related interpretation in terms of what examinees know and can do, it is important that evidence about the quality of work they produce plays some part in the standard maintaining process. This evidence, in the form of judgement by acknowledged experts of the quality of a sample of scripts, has long been, and still is, part of the standard maintaining process as described in the Code of Practice (Ofqual, 2011). No-one has claimed that evidence about the quality of work produced is irrelevant or unimportant – the debate has been about whether expert judgement is reliable enough (see above) or whether expert judgement is being captured in the best way. Currently the procedures require experts to make 'top-down bottom-up' judgements about the grade-worthiness of scripts in a range of marks around where the grade boundary is thought to lie. The 'top-down bottom-up' label refers to the process of starting at the top of the range and

[3] Cresswell (2003) first used the term 'comparable outcomes' as opposed to 'comparable performances' to highlight different perspectives on standard maintaining.

working down to a mark where there is some uncertainty, then starting at the bottom and working up to a mark where there is some uncertainty, the two marks thus derived forming a narrower range in which the grade boundary can be set.

There have been two main criticisms of this approach. The first is that the range of scripts considered by the experts is determined by the two sources of evidence described above and thus is not independent of them, virtually[4] guaranteeing that this source of evidence does not conflict with the others. The second is that the process requires absolute judgements of grade-worthiness (e.g. "is this script worthy of an 'A'") and thus requires the experts to share the same idea of an absolute standard. As described below, procedures relying on comparative judgements can overcome both of these disadvantages.

Existing research on the use of comparative judgement for grade awarding
Aside from the literature reviewed by Baird (2007), Cambridge Assessment has produced several pieces of work suggesting that comparative judgement may be an effective method of awarding. Only a minority of these have been acknowledged in the recent reviews by Ofqual (for example, Ofqual, 2014a). This research has shown:

- The method is practicable (Bramley, 2005, 2007)
- The method removes the reliance on internal models of examination standards (Bramley, 2007)
- The method has greater validity than existing methods of applying expert judgement (Black and Bramley, 2008)
- The method can be tested for whether it works appropriately in each individual situation (Bramley and Gill, 2010)
- It is possible to calculate confidence intervals showing the exact reliability of the generated grade boundaries (Bramley and Gill, 2010).

Benton (2014) also shows how the above research on the reliability of expert judgement is compatible with some of the more critical research detailed earlier. He also provides a mathematical model which can be used to design an expert judgement process with a given level of reliability or to calculate the reliability of various alternative methods of applying comparative judgement beyond those explored in the above papers.

The suggested method requires examiners to compare the relative grade-worthiness of sets of scripts (often pairs) from different years. The results of these comparisons can then combined to allow an overall comparison of the relative difficulty of two examinations which, in turn, allows the grade boundaries from one year to be mapped to an appropriate point in the mark scale the following year.

It should be recognised that these comparative judgement methods do not provide an entirely independent source of evidence because they all implicitly rely on the experts making a judgement about the relative difficulty of the questions when they compare quality of work produced on different examinations. However, they are arguably much more independent than the current top-down bottom-up method.

It may of course not be feasible to use such methods in all circumstances – for example because of resource or logistic constraints. But it is worthwhile to have a clear conception of what an 'ideal' process might look like, in order to evaluate the costs and benefits of departing from it to a greater or lesser extent.

[4] But not absolutely – the experts can in theory ask to see scripts in a different range of marks.

4. Combining the three sources of evidence

It may never be possible (or even desirable) to specify some exact formula for weighting the three sources of evidence when making decisions about grade boundaries. However, there are some circumstances where it will clearly be justifiable to give more (or less) weight to particular sources of evidence. For example, in times of major system change it makes sense to give less weight to evidence about quality of work or exam difficulty and to prioritise evidence about examinee ability, even if this is only in the form of a general measure of prior attainment. In times of stability it may make sense to give more weight to evidence about the quality of work; or to the difficulty of the questions.

Considerations of reliability are still important – and this means it will be important to develop or deploy methods that can quantify the reliability of any given evidence, whether it comes from judgements about script quality, or from pre-testing, or from quasi-equating (comparable outcomes). For example, for judgements about script quality it is important to estimate the extent to which different groups of examiners or a different selection of scripts for inclusion in the process would have resulted in a different derived grade boundary. It is also important that the process of combining different sources of evidence acknowledges the conditions under which each is likely to be more reliable and whether the recommendations from them are significantly different.

The main thing is to have a defensible rationale for the decisions taken. For the sake of public confidence it is important that both the regulator and exam boards are seen to be tackling “grade inflation”. This means that we would not want to return to a situation where pass rates are ever increasing without strong evidence of genuine improvements in performance.

In future, one source of evidence about the ability of the examinees may come from the currently proposed National Reference Test. However, until this is in place and has been subject to extensive evaluation it is not clear how reliable the evidence from this source will be. Furthermore, the National Reference Test will only provide evidence regarding achievement in English and Mathematics and could not reasonably be seen as a strong source of evidence for genuine changes in performance in other subjects.

5. Final comment

Most of the intended uses of examination results rely upon a form of criterion referencing. That is, they require that the exam certifies whether pupils have acquired the same knowledge and skills as pupils awarded the same grades in the past. As such, our method of grade awarding should aim for a criterion referenced system (albeit weakly criterion referenced) whilst at the same time recognising the extreme difficulty in achieving this. In such circumstances sources of evidence which are not directly aimed at criterion referencing may still have a crucial role. However, the central aim of identifying whether students have acquired a given set of knowledge and skills should never be discarded. This requires that expert judgement plays a continuing role in grading examinations. Since the most transparent and most accurate form of expert judgement is comparative, this should be incorporated in the setting of examination standards.

References

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

Baird, J., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: Valid, but inexact.* Internal report RPA 05 JB RP 077. Guildford: Assessment and Qualifications Alliance.

Baird, J., & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A level examination performances. *Educational Studies*, 28, 143–162.

Benton, T. (2014) Comparing the reliability of standard maintaining via examiner judgement to statistical approaches. Paper presented at the International Assocation for Educational Assessment (IAEA) conference, Singapore, 25th-31st May 2014.

Benton, T., & Lin, Y. (2011). Investigating the relationship between A level results and prior attainment at GCSE. Coventry: Ofqual.

Black, B., and Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23*(3), 357-373.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement, 6*(2), 202-223.

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

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment Publication, 13*, 18-26.

Bramley, T. (2013). *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., and Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, 25(3), 293-317.

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.

Bramley, T., Dawson, A., & Newton, P. E. (2014). *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. & Wilson, F. (in prep). Estimating grade boundary location by expert judgement of question difficulty.

Brandon, P. R. (2004) Conclusions About Frequently Studied Modified Angoff Standard-Setting Topics, *Applied Measurement in Education*, 17(1), 59-88.

Cresswell, M.J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 69-104). Oxford: Oxford University Press for The British Academy.

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

Curcin, M., Black, B., & Bramley, T. (2010). *Towards a suitable method for standard-maintaining in multiple-choice tests: capturing expert judgment of test difficulty through rank-ordering*. Paper presented at the Association for Educational Assessment-Europe (AEA-Europe) annual conference, Oslo, Norway.

Good, F.J., & Cresswell, M.J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, 14, 263–281.

Forster, M. (2005). *Can examiners successfully distinguish between scripts that vary by only a small range of marks?* Unpublished internal paper, Oxford Cambridge and RSA

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69–81.

Newton, P. E. (2010a). Thinking about linking. *Measurement: Interdisciplinary Research and Perspectives, 8*(1), 38-56.

Newton, P. E. (2010b). Contrasting conceptions of comparability. *Research Papers in Education, 25*(3), 285-292.

Newton, P. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 20-26.

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

Ofqual (2014a). *Setting GCSE, AS and A level grade standards in summer 2014*. Coventry: Ofqual. Downloaded from http://ofqual.gov.uk/documents/setting-gcse-level-grade-standards-summer-2014/ on 17th July 2014.

Ofqual (2014b). *Consultation on setting the grade standards of new GCSEs in England*. Coventry: Ofqual. Downloaded from http://ofqual.gov.uk/documents/setting-the-grade-standards-of-new-gcses-april-2014/ on 17th July 2014.

Plake, B.S., Impara, J.C. and Irwin, P. (1999) *Validation of Angoff-based predictions of Item Performance*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 19-23 1999.

Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, 26, 343–357.

Stringer, N. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, 27 (5), 535-554.

Taylor, M. (2013). *GCSE predictions using mean Key Stage 2 level as the measure of prior attainment.* Report to Joint Council on Qualifications (JCQ) Standards and Technical Advisory Group (STAG). Revised 26/06/13.