

An Appropriate Role for Professional Judgement in Maintaining Standards in English General Qualifications

Neil Stringer*, Assessment & Qualifications Alliance, UK

Abstract

In the UK, public examination standards are set and maintained using subject experts' (senior examiners) judgement of the quality of candidates' work, informed by statistics about the candidature and the mark distribution of the examination. This model has been called weak criterion referencing: the requirement of (strong) criterion referencing for evidence of specific knowledge, skills, and understanding is relaxed to allow for variations in examination difficulty, requiring only maintenance of the general quality of examination performance. There are several problems with this use of judgement. Firstly, the examiners making the judgement have insufficient information to estimate quantitatively the relative difficulty of two successive years' examinations, so their judgement is impressionistic. Secondly, and consequently, judgemental and statistical evidence are not complementary means of detecting the same "real" boundary mark because they do not share a common definition of difficulty. Thirdly, it has been demonstrated that experienced examiners are unable to distinguish between candidates' work within a small range of marks. To compound this problem, examiners appear biased toward giving candidates the benefit of doubt when deciding grade boundary marks. It is advocated that subject experts' role be changed to verifying statistically recommended boundary marks; trials of such a procedure are reported.

Introduction

In the UK¹, schooling is compulsory until the age of 16. In the final two years of compulsory education, pupils follow a program of study that culminates in them taking General Certificate of Secondary Education (GCSE) examinations in the subjects that they have studied. These qualifications are recognised by employers as well as by schools and further education institutions as entry qualifications for studying for higher-level qualifications, such as the General Certificate of Education (GCE). Like the GCSE, GCEs are subject qualifications and are normally taken during and/or following a one-year (GCE Advanced Subsidiary level) or two-year (GCE Advanced level) course of study. Typically, GCE candidates are seventeen and eighteen year olds studying three or more subjects. Like GCSEs, GCEs are recognised by employers and also by higher education institutions as entry qualifications for higher-level qualifications, such as a bachelor's degree. The results of both types of qualification are reported as grades: A to G for GCSE and A to E for GCE. While these grades are associated with descriptions of candidates' performances they are not, strictly speaking, criterion referenced.

As a rule, GCSE and GCE examiners do not pre-test or reuse examination questions. Instead, they set question papers that are intended to be as difficult as the previous year's papers. Consequently, the difficulty of question papers can be expected to vary from year to year. Criterion referencing does not take account of contextual factors that may cause variations in examination difficulty, so its use would thus be unfair to candidates who happened to take an

* Assessment & Qualifications Alliance, Stag Hill House, Guildford, Surrey, GU2 7XJ

Email: NStringer@aqaa.org.uk

¹ Note that Scotland operates a separate examination system from England, Northern Ireland, and Wales.

examination in a year in which the paper was particularly difficult. Therefore, when the grade boundary marks (cut scores) are decided by committees of senior examiners after the candidates have taken the examination, the examiners are expected to make allowances for the difficulty of the paper. This has been described as 'weak criterion referencing' (Baird, Cresswell, & Newton, 2000) and it is a Qualifications and Curriculum Authority (QCA) Code of Practice requirement for senior examiners to set grade boundaries using their professional judgement of the quality of candidates' work, informed by relevant technical and statistical evidence (QCA, 2008).

The problems with weak criterion referencing

Whilst we all broadly understand what is meant by "difficulty", measuring it requires us to have a precise definition of it in mind. At the individual level it is very subjective: I may find test A harder than test B, and you might find test B harder than test A. A third person may find test A *much* harder relative to test B than I did. Collectively, our responses would indicate that A is harder than B. This is where the use of examiners' judgemental evidence, particularly in combination with statistical evidence, encounters a fundamental problem. How can examiners estimate *quantitatively* and *precisely* the difficulty of an examination paper? The answer is that they cannot, because there is insufficient information to do so. Using their collective judgement, examiners are capable of identifying the sign of a change in difficulty but tend to underestimate its size (Cresswell, 1997). To illustrate, the examiners' task can be considered to be similar to that in Figure 1.



Figure 1. Here is last year's 'Grade C'. Find the shade on this year's grey scale that matches the *original* shade, allowing for any change in the ink that might have occurred through aging over the last year.

Given some background information about two cohorts of candidates for an examination, it is possible to statistically estimate the difficulty of the two years' examinations. For example, if we know that the two cohorts had performed identically in previous examinations, we might expect them to perform very similarly in subsequent examinations. If they do not, we might attribute the difference between their performances to differences in difficulty between the two years' papers of the new examination. So, if 25 percent of the 2007 cohort gain 45 marks or more in the exam and 25 percent of the 2008 cohort gain 48 marks or more in the exam, we might infer that the 2008 cohort took the easier of the two papers and that the marks of 45 and 48 are comparable in terms of achievement. We might therefore set a grade A boundary at 45 marks in 2007 and at

48 marks in 2008. The Assessment and Qualifications Alliance (AQA) awarding meetings are routinely provided with predicted examination outcomes based on the prior attainment of the current cohort of candidates and the relationship between the prior attainment and the examination outcomes of the previous cohort of candidates. Such an analysis implies that difficulty is an empirical measure that relates to the relative performances of two entire cohorts of candidates, or at least very large samples from them. The problem with adopting such a definition of difficulty is that it is impossible for the examiners to testify to it. Informing examiners of the average difference in prior attainment between two cohorts cannot help them to make comparisons between individual scripts from the two cohorts because the examiners do not know anything about the individuals whose scripts they are comparing. Even if the examiners had access to the necessary information about each of the candidates *whose work they scrutinised* and a computer to analyse these data, such an underpowered analysis would be unhelpful for quantifying the effect, and possibly unreliable at indicating even its sign.

The statistical and judgemental evidence, rather than converging on the same concept of difficulty, point at distinct concepts. The fact that there is often strong agreement between statistical and judgemental evidence perhaps testifies more to the stability of grade boundaries in established examinations and the lack of independence between the two sources of evidence—examiners are aware of the statistically recommended boundary before they scrutinise scripts—than it does to the logical coherency of the process (Figure 2).

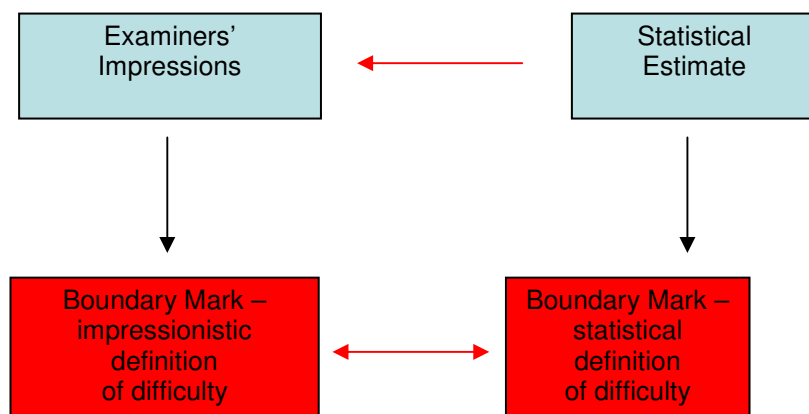


Figure 2. Multiple definitions of difficulty and crosstalk between evidence. Examiners’ judgemental evidence can be influenced by the statistical estimate in which case the judgemental and statistical estimates will lack independence.

Assuming that awarders *could* somehow compute and visualise a model script of the adjusted standard they were looking for, research has shown that they would be unlikely to identify it among a small range of scripts with any degree of precision (Baird & Dhillon, 2005). Figure 3 demonstrates the difficulty of discriminating between stimuli that are adjacent or merely proximate on a scale.



Figure 3. The greyscale from Figure 1 is divided/graded into six equal portions. Hold a piece of paper in each hand either side of any of the grade boundaries. Starting very close to the boundary, move the papers apart slowly, keeping the boundary mark in the middle of the gap between them. How far from either side of the grade boundary do you move the paper before you begin to notice differences between the two grades? If some pixels were randomly reshuffled to simulate marking tolerance, it is likely that you would have to move the two papers even further apart.

So, we are in a position where examiners: 1) cannot make objective quantitative estimates of examination difficulty, therefore they; 2) cannot testify to the same definition of difficulty that statistical predictions do and they; 3) cannot discriminate between scripts on neighbouring marks precisely. Add to this awarding committees' "tendency to choose the lower of two scores when there is a decision to be made about setting the minimum mark for a grade" (SCAA/Ofsted, 1996) and the finding that they are more likely to lower a grade boundary than to raise it, relative to the statistically recommended boundary (Stringer, 2008), and one might be forgiven for wondering why we would want to continue to use professional judgement in the awarding process.

Why we should continue to use professional judgement in the awarding process

The current emphasis on judgemental evidence in awarding body procedures places unrealistic expectations on examiners, but that does not mean that examiners' judgments are not an essential part of the system. Although examiners' judgements are poor at indicating the relative difficulty of successive examination papers, statistical approaches to accounting for examination difficulty cannot tell us anything about the quality of candidates' performances. Unlike cohort referencing (William, 1996), where (approximately) the same proportion of candidates from successive cohorts would receive each grade, the predictions used to guide awarding meetings make allowances for changes in the distribution of ability between cohorts, albeit using a measure of prior attainment as a proxy. What the predictions do not—and *cannot*—do is allow for changes between the predictor variable, e.g. mean GCSE score, and the outcome variable, e.g. GCE A level Psychology grade (Baird, 2007; Goldstein & Cresswell, 1996). A system that does not allow for the possibility of a change in the value added between, for example, GCSE and GCE, is potentially unjust. Only expert human judgement can (or at least attempt to) tell us whether this relationship has been maintained in reality. Related to this is the importance of public confidence in the examination system: the currency of examination results depends on it. Even if a statistical procedure for maintaining GCSE and GCE standards that did not require the participation of examiners was developed, using it might seriously damage the perceived validity of a qualification.

Confirmatory awarding

Given that script scrutiny cannot contextualise accurately script quality in terms of relative examination difficulty, AQA has trialled a confirmatory method of awarding that incorporates a more appropriate role for awarders, shifting the emphasis from identifying a specific mark from

within a range toward verifying the statistically recommended boundary. Script scrutiny in the confirmation method serves to ascertain whether candidates' work on the statistically recommended boundary displays the minimum general quality expected of candidates achieving that grade. Each awarder looks at scripts on the statistically recommended boundary and indicates whether he or she thinks the work is:

2 = Certainly worth the grade

1 = Borderline worth the grade

0 = Certainly *not* worth the grade

The examiners' judgements are recorded in a chart and an index of uncertainty (median rating) is calculated. It requires more than half the committee to vote in the same direction to move from the statistically recommended boundary. In other words, the index must be 2 or 0 to warrant considering the mark below or above the statistically recommended boundary, respectively. Values of 0.5, 1, and 1.5 indicate that the balance of opinion is that the statistically recommended boundary is "borderline worth the grade", i.e. comparable to the established standard of work at the grade boundary, thus the statistically recommended boundary is confirmed.

Mark /Awardeer	Chair of Examiners	Principle Examiner	Chief Examiner	Awardeer 1	Awardeer 2	Index
38						
37						
36 (SRB)	1	2	1	0	1	1
35						
34						

Figure 4. A tick chart using the confirmation method on which the statistically recommended boundary (SRB) is confirmed.

Trialling and evaluating confirmatory awarding

The confirmation method was trialled in one written paper unit of each of eight GCE awards in February 2007 and all of the written paper units of each of four GCE awards in February 2008. Evaluating the method from an operational perspective was straightforward: the outcomes of the awards in the trials were compared with the outcomes of awards using standard procedure. The views of the examiners who participated in the trials were collected and analysed and any concerns could either be addressed through revisions to the method or answered without the need to revise the method. A summary of these analyses is presented below, although neither of them tells us whether the confirmation method gives a more accurate answer than the standard procedure.

Cross-validating the trial procedure is certainly not straightforward and perhaps not even possible. If we devised a method for visually estimating the weight of elephants, we could validate the measure by comparing the estimates produced by the method with the measured weights of a sample of elephants. If we could not weigh the elephants, we could at least compare our outcomes with those of another established and validated means of estimating elephants' weights. Weight, though, is a clearly defined standard; examination standards are

not. They are very far from having a standard definition, not least because the standard that is maintained often depends on what the results of the examination are used for. For example, we can maintain statistical outcomes – 10% of each cohort are awarded a grade A; we can maintain performance outcomes – grade A candidates must be able to do x, y, and z; we can maintain the consensus of the examiners that the performances on the grade A boundary are equivalent to those on the grade A boundary in the previous year; we can maintain the statistical relationship between prior attainment and examination outcomes; and so on. The grade boundary marks required to maintain each of these things would be likely to vary.

If a procedure for maintaining standards is deemed socially acceptable, the standard is effectively defined by what that procedure is intended to maintain. It is a value judgement as to whether one procedure or another does a better job of maintaining standards because it depends entirely on what is intended to be maintained. By introducing confirmatory awarding, we might subtly change what is being maintained by placing greater emphasis on statistical equivalence and less emphasis on judgemental equivalence. Is that better than the status quo? If you are persuaded that statistical analyses provide the more objective and precise way of estimating the difficulty of an examination paper and that examiners' contributions should be to verify the statistical recommendation, then you might conclude that it is. If, on the other hand, you are unconvinced by the arguments for estimating examination difficulty statistically and think that examiners should decide what mark constitutes a particular grade performance based on scrutinising examination scripts across a range of marks, then you are likely to conclude that the confirmation method is not an improvement on the standard procedure. Within the current QCA Code of Practice, there is considerable scope for awarding bodies to decide how much emphasis is placed on statistical recommendations and examiners' judgements.

The confirmation method is intended to limit examiners' intervention to cases where they collectively have strong misgivings about the statistically recommended boundary mark. It is also intended to actively prevent them from deliberating over the choice between the statistically recommended boundary mark and a mark adjacent to it; a discrimination that evidence shows they cannot make genuinely (Baird & Dhillon, 2005). It is only in the latter respect that the confirmation method differs from AQA's standard procedures, which already require strong justification for decisions that deviate significantly from the statistical recommendations. However, the current QCA Code of Practice permits recommendations based largely on examiners' judgements and, given that we know they tend to under-compensate for changes in examination paper difficulty between years (Cresswell, 1997), this potentially undermines the ethos of weak criterion referencing which is to allow candidates of equal ability to gain equal grades regardless of the difficulty of the particular examination paper they took. Whether this possibility is objectionable depends on how weak we would like our weak criterion referencing.

Operational evaluation

The quantitative evaluation² essentially consisted of comparing the confirmation method with standard procedure in terms of the number of marks on which scripts were scrutinised and the distance of the final recommended boundary mark from the statistically recommended boundary mark. The outcomes on the units in the trial were compared with the outcomes on the same units in the previous year's award and the remaining units in the same award. According to the QCA Code of Practice, to which the standard procedure conforms, the range of uncertainty is delineated by the upper and lower limiting marks, these being, respectively, the highest and lowest marks that the examiners' judgmental evidence supports as potential grade boundary

² Data from February 2007 trial.

marks. Using standard procedure, this range was on average 2.7 marks wide; using the confirmation method, examiners looked at an average of 1.3 marks per boundary. The distance of the final recommended boundary mark from the statistically recommended boundary mark was on average 0.4 marks using standard procedure and 0.1 marks using the confirmation method. These figures suggest that there was typically enough certainty about the quality of the statistically recommended boundary for examiners to confirm it as their recommended boundary. In fact, out of sixteen decisions, only one was to recommend a mark other than the statistical recommendation: in this case the examiners moved up two marks from it in the light of clear qualitative evidence.

After the trials, the examiners involved were provided with forms to give written feedback on their experiences of using the confirmation method. Several examiners stressed the importance of looking at a range of scripts, including several on each mark close to and on the statistically recommended boundary. Two Chairs of Examiners thought that scrutinising scripts on one mark was unacceptable, with both commenting that comparison is the essence of awarding. The same Chairs advocated scrutinising the statistically recommended boundary and one mark above and below it; however, neither explained what information they glean by comparing scripts within a range of marks. Presumably, examiners look within the range for a script on which they start to see significant amounts of the skills and knowledge associated with the grade boundary in question. Although it is clear how this might work across a larger range of marks, it is doubtful that scripts one mark above or below the statistically recommended boundary will be very much different *overall* from scripts on the statistically recommended boundary itself. This is almost certainly why research has shown that examiners cannot tell apart scripts within such close proximity of each other (Baird & Dhillon, 2005). Furthermore, because examiners cannot quantitatively estimate the relative difficulty of a paper compared with the reference year paper, they cannot estimate the precise levels of knowledge and skills that they should find for the grade in question. If they look for fixed levels, then they are attempting strong criterion referencing, which is not how GCE standards are intended to be maintained.

If examiners *could* quantify the knowledge and skills they were looking for in the current year's scripts, identifying them through script comparisons would remain problematic. A paper with many (part-) questions provides many routes to the same total mark (Scharaschkin & Baird, 2000), even where there is no choice of questions. This presents a problem for examiners comparing performances on two scripts: a good performance on one question can compensate for a poor performance on another, so an uneven performance across a paper can be as valuable as an even one, even though Scharaschkin and Baird found that examiners rate even performances more highly than uneven ones. Examiners sometimes claim to have found qualitative shifts in candidates' responses between marks but, in the context of Scharaschkin and Baird's work, and given Baird and Dhillon's (2005) finding, it is more likely that they simply have found an even script on the mark above a less even script.

The confirmation method's emphasis on direct comparison with the archive was welcomed by many examiners, who expressed that they often neglected this using the standard procedure. One (acting) Chair of Examiners commented that this method focussed the awarding committee and that the numerical system was effective in obliging individuals to make clear decisions about individual scripts. Some examiners readily acknowledged that there is always a degree of uncertainty about the recommended grade boundary and that the confirmation method is no different in this respect. A number of committees were positive about the confirmation method but believed that the mark below the statistically recommended boundary should be checked, even if the statistically recommended boundary appeared sound. However, doing so would be

unsatisfactory because, whatever the statistically recommended boundary, some scripts on the mark below it might well be as good—if not better—because of marking tolerance. Alternatively, scripts on the mark below the statistically recommended boundary might *look* as good as the statistically recommended boundary simply because examiners cannot make fine distinctions between adjacent scripts. Likewise, some scripts on or just above the boundary will be worse than some scripts below it. Although this is undesirable, short of severely restricting the forms of assessment used in general qualifications, for example to objective tests, there will frequently be an element of interpretation involved in applying a mark scheme to a script.

Feedback was positive concerning the timesaving aspect of the process: the boundary decisions took 57 percent of the time required for the standard procedure (Stringer, 2007). Whilst the operating costs of holding awarding meetings represent a relatively small proportion of the cost to awarding bodies of administering examinations, shortening the time spent awarding may have a number of significant benefits in other areas. For one, the Government's proposed move toward a system of Post-Qualifications Applications (PQA) to higher education by 2012 will be more feasible if the awarding season can be substantially shortened. Examiners commented during the trials that they were less tired by the time they scrutinized scripts on the final boundaries than they would normally be. This is not to say that decisions are currently compromised by fatigue—there is no evidence to suggest this—but the risk of it occurring must be lower if examiners are less tired.

Various awarding body staff and examiners noted that the confirmation method might be suitable for large, established specifications, but probably not for new or small ones. This seems a sensible position: firstly, although statistical outcomes are always a consideration when setting standards in a revised syllabus, the exercise is essentially to set grading standards according to criteria, therefore an emphasis on qualitative judgement is appropriate and; secondly, the size of the confidence intervals associated with statistical predictions vary inversely with the size of the entry on which they are based (Pinot De Moira, 2008).

Conclusions

Weak criterion referencing as it is embodied in the QCA Code of Practice, and therefore in awarding bodies' procedures, has several problems. At a conceptual level, the examiners making the judgement have insufficient information to estimate quantitatively the relative difficulty of two successive years' examinations, so their judgement is impressionistic. Consequently, judgemental and statistical evidence cannot be complementary means of detecting the same "real" boundary mark because they do not share a common definition of difficulty. At an empirical level, it has been demonstrated that experienced examiners are unable to distinguish between candidates' work within a small range of marks. To compound this problem, examiners appear biased toward giving candidates the benefit of doubt when deciding grade boundary marks. Nonetheless, examiners' professional judgement is an essential part of any form of criterion-referenced examination system. It is advocated that subject experts' role be changed to verifying statistically recommended boundary marks. This proposed change of emphasis in favour of the statistical evidence asserts the necessity of examiners' qualitative evidence in awarding procedures whilst acknowledging its limitations.

References

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

- Baird, J.-A., Cresswell, M. J., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, 15(2), 213-229.
- Baird, J.-A., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact* (No. RPA_05_JB_RP_077). Guildford: Assessment and Qualifications Alliance.
- Cresswell, M. J. (1997). *Examining judgements: theory and practice of awarding public examination grades*. University of London, London.
- Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education*, 22(4), 435-442.
- Pinot De Moira, A. (2008). *Statistical predictions in award meetings: how confident should we be?* (No. RPA_08_APM_RP_013). Guildford: Assessment and Qualifications Alliance.
- QCA. (2008). *GCSE, GCE and AEA Code of Practice*. London: Qualifications and Curriculum Authority.
- SCAA/Ofsted. (1996). *Standards in public examinations 1975-1995*. London: SCAA.
- Scharaschkin, A., & Baird, J.-A. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, 26(3), 343-357.
- Stringer, N. (2007). *Evaluation of the February 2007 alternative awarding procedure trials* (No. RPA_07_NS_RP_039). Guildford: Assessment and Qualifications Alliance.
- Stringer, N. (2008). An appropriate role for professional judgement in maintaining standards in English general qualifications. Assessment and Qualifications Alliance.
- William, D. (1996). Meanings and consequences in standard setting. *Assessment in Education: Principles, Policy & Practice*, 3(3), 287-308.