# AWARDING GCSE AND GCE - TIME TO REFORM THE CODE OF PRACTICE?

Ben Jones

## 1. ABSTRACT/INTRODUCTION

The theme of this paper is that the process by which national general qualifications – the GCSE and GCE in particular – are awarded is no longer fit for purpose.

It is argued that, although the process has changed little since GCE A levels started in 1951, much of the context in which it is practised has changed substantially, and as such it should be reconstructed, perhaps quite radically. In particular, it is argued that the process has neither adequately kept pace with, nor adapted to, the findings of research projects, developments in technology, and the availability of mass, candidate-level data. Some of its implicit theoretical underpinnings are frequently exposed as ill-founded, it is expensive in terms of both time and money, and is opaque and not conducive to rigorous monitoring and evaluation. Perhaps most concerning of all is the evidence that strict adherence to the statutory *GCSE, GCE and AEA Code of Practice* (CoP) (2009) not only provides no guarantee of the alignment of standards – whether between awarding organisations or years – it may actually serve to undermine that aim[1]. In short, it is argued that the process is in need of a substantial reform.

In section 2 of the paper, the essential features of the awarding process, which are prescribed in part of §6 of the CoP: *Awarding, marking review, maintaining an archive and issuing results*, are described from AQA's perspective; most details of how other awarding organisations interpret the CoP are not known. A brief description of how little the process has changed over at least twenty-five years is also included. Section 3 comprises a critique of the present arrangements' prime emphasis on examiner judgement, by referring both to formal and informal research evidence which casts doubt on its reliability. In contrast, sections 4 and 5 identify the increasingly available, valid and reliable candidate-level statistical data which are currently used, albeit voluntarily and in a piecemeal way, by awarding organisations to inform their awards. The shortcomings of relying solely on such data are also identified. Section 6 notes some structural, procedural and political developments which would also argue for a change in emphasis in the approach to awarding, before the summary and discussion in section 7 which proposes how the system might be changed.

**KEYWORDS:** Awarding; Standards; Code of Practice; Examiner Judgement; Statistical Predictions

[1] The most obvious recent example of this phenomenon was the GCSE Science award in 2008, where each Awarding Body could justifiably claim to have complied with the requirements of the Code of Practice, yet before the Regulator's late intervention, the standards between the awarding bodies were badly misaligned, and afterwards those between years were misaligned.

## 2 CURRENT PRACTICE

### 2.1. Current practice as dictated by the CoP

Every summer, awarding decisions are made nationally for approaching 500 GCSE specifications and 500 GCE (including AS and A level) specifications, affecting some 600,000 and 300,000 candidates respectively. These qualifications have been called "curriculum-embedded" (Cresswell, 1996), indicating that they are taught to, and assessed against, a defined specification including programmes of study, assessment objectives, grade descriptors etc. Candidates' performances in these qualifications are assessed through their work being marked according to a clearly defined mark scheme, and are rewarded according to a grading scheme: A*-G in the case of GCSE and, as from 2010, A*-E in GCE A level.

The process by which candidates' marks in the assessments are translated into grades has, in recent years, been dictated by an - initially voluntary and latterly statutory - CoP with which all awarding organisations offering these qualifications comply. A full prescription of the awarding process is contained in some detail in paragraphs 6.1 to 6.29 of the current version of the CoP. In essence the senior examiners for each specification (Chair, Chief and Principal Examiners and Moderator(s)) meet for usually 1 or 2 days (some meetings last for up to 5 days) to peruse all the relevant evidence, especially ranges of borderline scripts, and recommend grade boundary marks. The nature of the evidence used to inform the award meeting is essentially of two types: qualitative (e.g. current and archive scripts, mark schemes, Principal Examiners' reports and grade descriptors) and quantitative (e.g. mark distributions, item level data, entry patterns and centres' estimated grades).

At each judgemental grade boundary, the committee members independently peruse and vote on a range of scripts, after which the Chair, on their behalf, recommends what s/he considers to be the most appropriate provisional boundary mark. At the end of the whole process, one or more of these provisional boundary marks may be changed in the light of the resulting subject outcome.

### 2.2 Some features of the CoP

It is worth identifying and commenting on some of the explicit and implicit features of the current CoP. The stated purpose of this detailed process is the maintenance of standards: *"The prime objectives are the maintenance of grade standards over time and across different specifications within a qualification type"* (Para 6.2).

In a stable state, the former ambition should be relatively straightforward to achieve: the specification, structure and senior examining team will be the same, increasing the likelihood that the papers and mark schemes will be of similar demand, and the entry size and ability profile are also unlikely to change. The second aim may be less easy to realise, even in a stable state, as awarding organisations may propound various reasons for their specification appearing to be out of line with the others (e.g. the effect of a programme of teacher support, new support materials, INSET etc.). In any case, even in a stable state, the process prescribed in the CoP may not best realise either of these desirable ambitions.

There is a noticeable difference in the extent, detail and relevance of the qualitative and quantitative evidence to be provided, and in the level of prescription with which it is to be used. The implication of the CoP is that the awarders' judgements made on the basis of the

qualitative evidence are more valid and accurate that those provided by the statistical evidence. For example, paragraph 6.16 prescribes in some detail the process for scrutinising scripts around each putative grade boundary and how, on the basis of that scrutiny, a judgement about the recommended boundary mark should be arrived at. *"Awarders must then use their collective professional judgement to recommend a single mark for the grade boundary … which will include consideration of the evidence listed in paragraph 6.15."* As alluded to earlier, this evidence is either qualitative (question papers etc.) or quantitative. Not only is there an implicit imbalance in the weighting which should be assigned to these types of evidence, but such quantitative evidence as is required to be referred to is descriptive in nature - e.g. *"mark distributions for the current and previous series, where available"* - and, in some cases (e.g. centres' estimates), of doubtful reliability. Even when more sophisticated analyses are available, they may not necessarily be a central focus of the meeting.

The structure, organisation and conduct of the award meetings themselves reinforce the impression of the accuracy and importance of the awarders' qualitative judgements. The sheer person-hours spent scrutinising scripts, the way such deliberations are reported and decided on, via a "tick-chart", the discussions about individual marks and individual scripts etc. all serve to give weight to qualitative evidence at the expense of the quantitative. To the uninitiated observer, and probably for most examiners, this display of apparently reliable judgement is both impressive and reassuring.

Nevertheless, statistical information has a stronger, albeit covert, influence on the awarding process than is immediately apparent. Since awarding committees rarely recommend a mark beyond the initial range of scripts they are asked to scrutinise, how the script ranges are selected will largely determine what boundaries are recommended. Again, the CoP seems to give precedence to examiners' qualitative judgement in this selection: *"These ranges must be based on the preliminary ranges of marks proposed by principal examiners"* although, as a secondary consideration, it goes on to state that they *"must also take account of the relevant technical and statistical information"* (emphases added). It is, perhaps, legitimate that the CoP does not prescribe explicit details regarding what information might be relevant in this context, but in not doing so it leaves open the possibility that different awarding organisations, and the same awarding organisations at different times, will use different statistical sources and indicators.

### 2.3 A little history

The above process has changed little over the 60 years since the first GCE awards were made. Although there is little by way of procedure guidance files or codes of practice regarding how GCE (and even GCSE) examinations were initially and subsequently awarded, interesting documentary evidence does exist in the form of three video films commissioned by the Joint Matriculation Board (JMB) in 1985 to describe its processes (JMB, 1985). One of the films, lasting some 45 minutes, is entitled "The Award Meeting" and records the course of the JMB's 1985 GCE A level History award meeting, with a helpful commentary on the procedures. Although there were some noteworthy differences in the organisation of the meeting[2], in essence there is little to indicate that the meeting was a quarter of a century ago;

[2] Some of the more noticeable differences are:

- Principal Examiners worked in pairs when scrutinising and forming judgements on scripts;
- Principal Examiners scrutinised the whole work of candidates in the mark range, not, as currently, one unit/component at a time (using, generally, the work of different candidates for each component).
- There were no tick charts – each pair of Principal Examiners was asked to recommend their preferred boundary mark
- The judgemental grade boundaries were: B/C, C/D and E/O; the grade A boundary was not determined judgementally although it is not made clear how it was set.

the procedures would largely comply with the current CoP. There were, however, some noticeable features in the award, some of which, in retrospect, given the subsequent increased availability of more statistical information, are somewhat ironic. First, it is clear from Table 1 that, in its fourth year, the specification (or syllabus, as it was then known) was in a very stable state. Such stability would serve to make both the statistical and judgemental evidence more reliable[3].

**TABLE 1. JMB GCE A LEVEL HISTORY SYLLABUS A ENTRIES AND PASS RATES 1981-1985**

| Year | Entries | Grade A % | Grade E % | Grade E boundary mark (*I*200) |
|---|---|---|---|---|
| 1981 | 9426 | 8.9 | 67.1 | 78 |
| 1982 | 9959 | 9.4 | 67.3 | 77 |
| 1983 | 10280 | 9.5 | 66.8 | 76 |
| 1984 | 10231 | 9.9 | 68.9 | 75 |
| 1985 | 9921 | 9.9 | 68.2 | 75 |

Second, the statistical evidence which was adduced were the pass rates of the previous year in particular, and those of the previous four years in general, and results of subject pairs analyses comparing the 1984 candidates' mean History grades with their mean grades in their other subjects. As the commentary says, "*JMB puts considerable store by subject pairs statistics*". Third, the statistical preamble and discussion at the start of the meeting (led by the Secretary of the JMB) lasted the whole morning; no scripts were scrutinised until after lunch. Fourth, at the grade E decision, there was a dispute between the statistical evidence (the equi-percentile outcome from the previous year) and the awarders' evidence. As a perusal of more scripts[4] did not resolve the issue, the Chair was left to make a judgement, which he clearly agonised over, but ultimately chose to give preference to the statistical evidence. It is, perhaps, ironic that, in an era when the statistical data were less valid than they currently are, the Chair trusted their direction against that of the professional judgement of eleven awarders. Fifth, it was noted that in each of the previous years since the syllabus had been revised, the grade E boundary mark had dropped by one mark per year (see Table 1). The following exchange between one Principal Examiner and the Secretary of the JMB is enlightening:

> Principal Examiner: *"They [the boundaries] always seem to move downwards when they move …The raw figures might suggest that there has been a slight drift in standards over the years."*
> Secretary of the JMB*: "… This is a critical point; does this drift in the marks represent no more than a sort of drift we get because, as I say, there is nothing absolute about marks, or is it an unconscious drift down in standards. That is really the question isn't it?"*
> Principal Examiner*: "It is not a question to which we are going to get an answer now, but since 1981 the drift has been down."*

As the exchange identifies, whether this drift in boundary marks reflected the paper becoming gradually more demanding at that level over the period, or whether the committee was

[3] Incidentally, it is interesting to note that the AQA GCE A level History entry in 2009 was 9013 and the pass rates at grades A and E respectively were 28.1 *per cent* and 98.4 *per cent*!

[4] This practice is sometimes employed by Chairs nowadays to try and resolve similar dilemmas. It is, however, discouraged as scripts scrutinised in the knowledge of the dilemma will not be viewed in the same independent manner as those viewed beforehand, and votes might be biased accordingly. Moreover, the Chair should ensure sufficient scripts have been scrutinised before drawing the meeting together to make a boundary recommendation.

gradually becoming more lenient, is the critical question. Notwithstanding the Principal Examiner's final comment, it is the essence of the dilemma which the awarding process is intended to solve, and the question to which it is required to provide an answer.

There were a couple of other features of the award which are worth reporting. Principal Examiners frequently remarked that making decisions was difficult due to lack of balance across the questions and papers. (N.B. Footnote 2 reports that, unlike current practice, candidates' total work was scrutinised at award, not a unit/component at a time.) They also acknowledged that the sample of scripts they looked at was small and possibly atypical. In this vein, a comment from the same Principal Examiner quoted above is pertinent:

> *"I think if we go to [a mark of] 75 we do so on the basis of statistics rather than on the small and probably unrepresentative sample of scripts we saw*."

Second, although candidates' scripts from a range of six marks was pulled, work on only three of those marks appeared to be scrutinised; this same practice has recently been reintroduced in large, stable specifications.

### 2.4 Some features of AQA's current practice

In several ways, AQA's current practice of awarding aims to mitigate some of what it sees as defects of the prescribed CoP procedures, while remaining fully compliant with them.

AQA does not, for example, necessarily accept at face value the Principal Examiners' initial putative grade range (see section 2.2); if it appears to be some distance from expectations, the Principal Examiner is invited to revise it in the light of more objective statistical evidence. More importantly, the range of scripts which is pulled for the awarding committee to scrutinise, while being code compliant in including at least part of the Principal Examiner's range, generally focuses on, and is centred around, the boundary mark which the best statistical evidence suggests would yield the most defensible outcome. This is known as the Statistically Recommended Boundary (SRB). In addition, if boundary marks are recommended which yield outcomes which exceed what are deemed to be acceptable and defensible guidance limits[5], a thorough, and somewhat time-consuming, approval process is required, involving a detailed "Exceptional Report" from the Chair defending his/her recommendations on the basis of the quality of candidates' responses and a meeting with a designated Approver before close scrutiny of the award by the Accountable Officer. If, at any stage in the approval process, an Approver does not consider the award defensible, the Chair can be requested to amend one or more boundary recommendations better to reflect the quantitative evidence. This is a rare occurrence (the summer 2009 series had 5 (/1393) such occurrences in the GCSE group of qualifications and 8 (/1605) in the GCE group), and usually the Chair agrees to the boundary change, although if s/he refused, the CoP stipulates the mechanisms which must be applied, including the involvement of Ofqual and the AQA Council's Awarding Standards Committee (paras. 6.26-6.28).

[5] Guidance limits refer to the number of candidates differing from the predicted percentage. The limits vary according to the subject entry size, the larger the entry, the more stringent the limits. AQA's traditional limits are as follows:

| number of candidates | guidance limit |
|---|---|
| >500 | +/-2% |
| 301 - 500 | +/-3% |
| 201 – 300 | +/-4% |
| <200 | None |

Finally, AQA undertakes what is known as a "pre-results checking procedure" (PRCP) which, as its name suggests, is a check on the outcomes just prior to final publication. The purpose of the check is to ensure that any late marks, or adjustments made to marks, since the award have not overly affected the outcomes, causing them to exceed the appropriate guidance limits. When this does happen, the Chair is (again) asked to sanction a change to one (or more) boundaries recommended at the award meeting, which they almost invariably endorse. (The summer 2009 series had three such occurrences in the GCSE group of qualifications and none in the GCE group.)

These internal AQA procedures, none of which is defined by the CoP, but each of which is compliant with it, indicate that AQA's awarding procedures (rightly or wrongly) are more heavily informed and guided by quantitative statistical evidence in general, and the SRBs in particular, than either the CoP implies or than are practised by other awarding organisations (insofar as their awarding mechanisms are known). One consequence of this is that AQA is able to provide compelling evidence that it has maintained its standards across years, the first requirement of the CoP, albeit often causing it to become more severe in relation to other awarding organisations. Each year, the awarding organisations cooperate in a post-award "statistical screening" exercise, in which their outcomes are compared after the ability profiles of their candidates (as measured by their concurrent/previous GCSE scores for GCSE/GCE outcomes respectively) are controlled for. In cases with large discrepancies, remedial measures are taken in the following series in an attempt to regain inter-awarding body alignment. After the 2008 GCSE summer series, AQA boundaries were relatively severe compared to those of the other awarding organisations in 34 out of 39 cases where a significant difference was flagged. Moreover, this widespread severity was after three or four years of statistical screening during which the awarding organisations had purportedly made remedial adjustments - AQA usually in the direction of becoming more lenient. A second consequence is that the powers of the awarders – as implied in the CoP – have effectively been curtailed and their freedom of deciding which boundary marks to recommend reduced. This is a point not lost on many awarders who argue, not without some justification, that their role has changed from making a judgement based on script scrutiny to confirming that the quality of work seen on the SRB is defensible. Some argue that their role is effectively redundant. Of the 3205 judgemental grade boundary decisions made in AQA in 2009, all but 179 (6 *per cent*) were within one mark of the statistically recommended boundary.

## 3. DEFICIENCIES OF THE CURRENT METHOD

Although little in the essential conduct of the award meeting has changed over at least the past twenty-five years, there has been, and continues to be, much change in the wider context in which awarding takes place. These changes can be classified into three areas: research on examiner judgment; improved statistical information; and structural, procedural and political developments. The thrust of this paper is that current practice has neither adequately kept pace with, nor adapted to, these developments.

### 3.1 Research on examiner judgement

First, a considerable body of research has built up which indicates that, despite their apparent accuracy and expertise, the judgements of senior and experienced examiners are less reliable, and more susceptible to changes in conditions, than had previously been believed. In an interview at the end of the 1985 JMB GCE History award meeting, for example, the Chair is asked about this apparent accuracy:

> Interviewer*:* *"…..Remarkable consistency over the years*
> Chair: *"…One is very reassured indeed"*
> Interviewer: *"Is this built up purely through experience – it is quite remarkable the way groups [pairs of awarders] quite independently – seem to agree on the level, that a 75 was a pass, a 74 wouldn't be, a 94 might be a boundary but a 93 wouldn't – is that achieved over the years, through experience?"*
> Chair: *"Yes. I don't think you can dispute that point"*

Research in the intervening years has, however, served to cast doubt on such confidence, with the "remarkable consistency" seeming to be more a function of the statistical guidance – whether explicit or implicit through the range of scripts provided for scrutiny – than of awarders' inner knowledge accrued over years of the appropriate standard. The extent to which awarders can make reliable and accurate judgemental decisions on scripts has been shown to be questionable, with many of the recent research findings emanating from AQA and its predecessor bodies.

#### 3.1.1. Formal evidence

Perhaps most significant has been the body of work demonstrating the difficulty that even experienced examiners have in making accurate comparisons of work produced under different demand conditions. The introduction of the GCSE in 1988 posed a particular problem in this regard, as it embodied a differentiated paper assessment design requiring the same standard to obtain at grade C on both the higher and lower tiers. As part of a substantial research study to inform the introduction of the GCSE – *the Novel Examinations at 16+ Research Project* – Good & Cresswell (1988) found this intention was not being realised:

> *"For the purposes of grading differentiated papers, it is suggested that grades can be defined as comparable if they are reached by the same proportion of a given group of candidates. This definition was not, however, consistent with the grade awarders' judgements of comparable performances. The awarders tended to consider fewer candidates to be worthy of any given grade on harder papers or, alternatively, that more candidates reached the required standards on easier papers."* (p. vii).

Although this research was undertaken in the context of differentiated paper assessments, the same principle - of awarders' making inadequate compensation for differences in assessment demand - can be, and has been, applied elsewhere, particularly to examinations in adjacent years.

Principal Examiners are, by definition, experienced in the various functions of examining, and are charged each series with producing examination papers and mark schemes of comparable demand. They trust their judgement in being able to undertake that task, as evidenced by the general consistency of their initial recommended grade boundary marks.

However, were they able perfectly to set papers and mark schemes of equivalent demand between years, award meetings would be unnecessary as the same boundary mark would carry forward the standard. That they cannot do so, despite the aid of a rigorous question paper evaluation process, is the reason award meetings are necessary. (It is, incidentally, somewhat perverse that the known difficulty in judging and setting papers of identical demand is relied on in the awarding process which is designed to rectify it.)

Applying the approach to inter-year standards, Cresswell (2000) reported on an analysis of outcomes in two adjacent years (1990-1991) of thirty-eight, undifferentiated GCE A level subjects, which had been awarded purely by examiner judgement. He initially posited, then demonstrated the inadequacy of, three possible explanations for the differences in the subject outcomes between the two years: the candidature as a whole had improved/deteriorated; the balance of entry between centre types and/or genders had changed; this year's new (missing) candidates were better (worse) than the rest. He then investigated a fourth explanation: that a relationship exists between examiner judgements and the statistical features of the candidates' marks. More specifically, he argued that the differences in subject outcomes between years were at least partially due to fluctuations in the awarders' judgemental standard and, in particular, demonstrated that, whilst awarders tend to be able to identify *if* the demand of a paper/mark scheme differed from the previous year, they fail adequately to compensate for such differences, typically adjusting the grade boundary some 40 *per cent* of the distance (in either direction) to where it should be moved to maintain the standard.

Awarders' failing adequately to compensate for differences in demand when recommending grade boundaries has come to be known as the "Good and Cresswell effect", and even nowadays it can be seen many times in each awarding series, although for any individual case one can never be totally sure if one of Cresswell's initial explanations might be justifiable.

Cresswell (1997) replicated the above analyses using 1993-1994 data, by which time awards were provided with, and guided by, more statistical information. The results of the analysis showed that now about 90 *per cent* of the compensation which the statistical modelling suggested was required was allowed for by the awarders' recommended boundary marks. Baird & Morrissey (2005) also replicated and extended this type of analysis for 1998-1999 GCE and GCSE data, and close relationships between the actual boundaries and those implied by the statistical predictions were again found. Similarly, and more recently, in his work on the efficacy of using statistical approaches to guide awarding, Stringer (2008) reported astonishingly large correlations between the predicted and actual cumulative percentage outcomes at GCE A level between 2003 and 2007. Not only are the coefficients large, even in 2003 (0.974 and 0.919 at grades A and E respectively), at both grades they have been increasing monotonically during that period; in 2007 the corresponding coefficients being 0.993 and 0.989.

Baird & Morrissey (2005) suggested that there are at least two – not necessarily mutually exclusive - interpretations of results of this type: either the statistics are very influential in the standard setting process and/or there is a high degree of (independent) agreement between the judgemental and statistical evidence. From the evidence of Cresswell's first study and other research reported here, the former appears to be the more likely explanation.

Part of a research project designed to investigate this issue was undertaken by Baird & Dhillon (2005). Their research question was, *"Can awarders discriminate in terms of grade worthiness between work on adjacent marks?"* Using scripts from GCSE English and GCE

Physics examinations, awarders were given fourteen mark-free scripts in a seven mark range from around a grade boundary, and were then asked to rank order them and assign a grade to each on a seven point scale. Overall the results were very poor: for few examiners did the rank order correlate significantly with the actual marks, and comparisons of their judged grade boundaries with the actual boundaries were also disappointing. At grade C in GCSE English the awarders were moderately successful in their given tasks, but even so, in over 25 *per cent* of cases their judgementally determined boundaries were more than two marks adrift of the actual boundary. The report concludes:

> *"For each of the other grade boundaries investigated, the success rates were so low that it has to be concluded that the task of distinguishing grade-worthiness of scripts that vary by only a small range of marks is impossible, even for highly expert judges"* (p11-12)

Whilst the ability of awarders to judge grade boundaries reliably in the absence of substantial statistical guidance is thus questionable, other research has indicated how their judgement is influenced.

Scharaschkin & Baird (2000), for example, investigated the extent to which consistency of candidate performance in two GCE A level subjects affected awarders' judgement of their grade-worthiness. This is an important issue, especially at the lower grades where performance is typically unbalanced. They found that inconsistent performance in biology produced fewer judgements of grade worthiness than consistent or average performance, and very consistent performance in sociology was preferred to average consistency. Thus, in both cases a feature of examination performance, other than stipulated by the mark scheme, affected grading decisions and the authors thus concluded *"that examiners' judgements of standards should be supported by other sources of evidence, such as examination statistics."*

One item of non-statistical evidence with which awarders are required to acquaint themselves before awarding is the bundle of archive scripts. In an experiment investigating the effect of these on awarders' decisions, Baird (2000) found that "*in both subjects (GCE A level English and Psychology) no effect of archive scripts was found when examiners judged the position of the grade boundary marks"*, and that this is consistent with the idea of *"an internalised standard being used by both sets of examiners when carrying out the boundary judgement task."* This internalised standard can be thought of as a loose prototype, some of whose characteristics may be common to a comparator. However, the qualities required for performance at a given grade cannot be exemplified because of the large number of routes to a given mark.

At a seminar on Chairs' Exceptional Report writing at the 2008 AQA GCE Chairs' Conference, scepticism was expressed about their duty objectively to justify an exceptional grade boundary recommendation by comparing subject-specific characteristics on current and archive scripts. It was argued that it would be equally easy to muster evidence showing that the quality of the archive scripts were worse, equal to, or better than those on the current boundary.

Despite, or perhaps because of, this notion of a loose prototype, awarders are vulnerable to sources of bias such as those described above. Another source was described by Stringer (2008), who reported that where awarders recommend boundaries other than the SRB, they tend to be generous, i.e. selecting marks below, rather than above, the SRB. Dismissing alternative explanations, he concluded that the bias was in the examiner judgement (as

opposed to the statistical prediction) and was due to awarders' tending to give candidates the benefit of any doubt. This practice, while small in any given year, will introduce an incremental, structural bias, ensuring that standards are slightly compromised over time. This point was also made by a 1996 SCAA/Ofsted report in which stated that, *"the emphasis given to awarders' judgement of the quality of candidates' work rather than to statistical data, coupled with a tendency to choose the lower of the two scores when there is a decision to be made about setting the minimum mark for a grade, may have allowed small, unintended but cumulative reductions in grade standards in successive years.*" (p15).

### 3.1.2 Informal evidence

Apart from the formal research in this area, there are other reasons to doubt the reliability of awarders' judgements.

First, very occasionally the data on which predictions are calculated, and used as the basis for pulling script ranges, have been faulty. Because of their rarity and impact these occurrences tend to stick in the memory. However, a concerning feature of such cases is that the awarders do not usually recognise from their script inspection that anything is awry; they are content to recommend a boundary mark somewhere in the middle of the erroneous range. (Section 5 indicates, however, that occasionally they do identify a problem, and thus, if for no other reason, a scrutiny of scripts is a vital, albeit not fully reliable, failsafe mechanism in the system.) If awarders confidently identify a putative grade standard so far from where it should be (and is subsequently set), it is hard to trust their decisions in choosing a boundary between two or three marks.

Second, an underpinning assumption of current practice is that an individual awarder's decisions within a script range are made independently of each other. When scripts in, say, only a five mark range are scrutinised, this is difficult to achieve, and in order for awarders to achieve internal consistency for themselves, and in the eyes of their colleagues, there will be pressure on them to ensure their judgements are consistent with the marks, especially if scripts are viewed in mark sequence (a practice AQA does not endorse).

Third, for both practical and theoretical reasons, a candidate's full work is not seen by the awarders (unlike in 1985 and previous years). As such, they cannot claim to know "the A level standard", which resides at subject level. It is thus possible for awarders accurately to maintain unit level standards, but for the subject standard to shift considerably (or *vice versa*) due, for example, to changes in inter-unit correlations. (An illustration of this effect is the difference in the statistically recommended unit boundary before and after it has been adjusted to ensure that the subject outcome meets its prediction.) Awarders should always know the impact of their decisions on the subject outcome and are usually willing, if not always happy, to realign their judgements at unit level in the service of the (unseen) subject standard.[6]

Fourth, the CoP requires that, *"After the marking and moderating period, principal examiners and principal moderators must propose preliminary ranges of marks for each component/unit as a basis for locating key grade boundaries on the basis of judgement*" (paragraph 6.4). Apart from being another instance of examiner judgement appearing to have precedence over statistical information (there is no overt requirement for a statistically generated range of

[6] The introduction of the new specification GCE AS units and subject awards in 2009 provided some dramatic examples of this. In order to achieve a better balance between the coursework and written unit outcomes – while maintaining the AS subject standard – substantial changes to unit standards were often made, with the agreement of the Chair.

marks to be derived), this exercise is increasingly difficult. Even Principal Examiners have often only marked a relatively small sample of scripts, which is likely to be biased, for example being dominated by scripts from a large, particularly well or poorly performing centre. Moreover, the increasingly widespread introduction of e-marking, whereby examiners no longer mark whole scripts, but samples of questions, not only makes the reconstitution of a set of scripts for the Principal Examiners to mark an artificial activity, but means they see considerably fewer (whole) scripts than previously. Typically Principal Examiners would have marked at least 50 scripts as their own allocation, in addition to samples from their team leaders as part of the marker standardisation process. With e-marking, Principal Examiners only mark fifteen reconstituted scripts – hardly enough to produce accurate grade boundary ranges. (Incidentally, as examiners no longer mark candidates' complete scripts, they too will become further de-skilled in building up a broad understanding of candidates' overall performance, albeit only at unit or component level.)

It is indicative of the unreliability of these proposed ranges that they are often a substantial distance from the final recommended mark. Sometimes this can be anticipated by Principal Examiners' being asked if, in the light of statistical data, they would like to revise their proposals, which arguably undermines the purpose of the exercise. Similarly, Principal Examiners often try and derive statistical information from their own (sometimes biased) sample of scripts - or, more tellingly, request it from the AQA - in order to inform their proposals. All of this casts doubt on the reliability of the Principal Examiners' proposals and on why they are currently deemed so important. If even the Principal Examiner of a paper is neither confident about, nor competent in, estimating even the range of likely boundary marks, how much confidence can be had in the other awarders' ability to judge within a mark or two?

Finally, all the English awarding organisations have already adopted, or are about to adopt, remote awarding procedures, whereby awarders do not attend an award meeting but scrutinise, and record their judgements about scripts, online. Such a change in technology also makes possible many changes in procedure, including how judgements on scripts are made. In particular, it is relatively easy to design the process such that independence of individual judgements is enhanced. Edexcel, which was the lead awarding body in this development, noted that the range of marks on which awarders were, individually and collectively, uncertain - sometimes called the "grey area" or "zone of uncertainty" – widened considerably. Apart from providing more informal evidence of awarders' unreliability of judgement, and the power of social processes in face-to-face award meetings, it suggests that recourse to a statistical indicator, if robust, is desirable.

## 4. IMPROVED STATISTICAL INFORMATION

Whilst the predictive statistical information in the 1985 JMB GCE History award was confined to the equi-percentile outcomes from the previous year and subject pairs analyses[7], increasingly more sophisticated, valid and reliable data have become available in the interim which are of particular value for informing awarding. (Descriptive statistics in the form of paper mean marks and standard deviations, and previous year's boundary marks, were also available but these are of limited use in indicating the position of the boundary marks.)

In the era of stable entry linear specifications (see Table 1), using the previous year's equi-percentile outcomes as the basis of statistical guidance was a legitimate and largely valid approach. Indeed, it is still adopted in AQA for calculating the SRBs of large and stable entry GCSE specifications. However, the desirability of adjusting these raw percentages to allow for changes in the ability profile of the candidature between years became increasingly apparent. Various approaches were adopted to achieve a better predictive statistic for the current year's outcome, primarily what became known as the "delta index" and "super delta index", which respectively used candidates' centre type classification and their centre's position in the Examination Performance Tables as the basis of the prediction, and common centres' analyses, which essentially applied the equi-percentile approach at centre level, rather than overall.

Although these approaches represent improvements to the overall equi-percentile approach, they are not ideal as their control variable, being at centre, not candidate, level, is a crude measure of individual ability (or, indeed, any group effect). It was noted, for example, that the variance of students' achievement within the 'Comprehensive' centre type (the largest category) was similar to that between all the centre types, and that the super delta approach did not produce predictions much different from the simple delta analysis.

The common centres analysis (which is still sometimes used) has its own deficiencies related to shifting entry patterns between centres. (An analysis of AQA GCSE centres in 2007 showed that only 25 *per cent* were stable, as defined by having a 10 *per cent* or smaller change in their entry from 2006. Even by broadening the definition of stability to 40 *per cent,* only some 65 *per cent* of centres became included.) Thus a similar centre entry profile between years might reflect a substantially different candidate entry profile, especially if centres' entry policies are based on candidate ability. Special examples of this difficulty occur when centres split a subject entry between two specifications – for example, the less and more able groups being entered for modular and linear specifications respectively or, in GCSE, when a centre's overall entry remains stable but the balance between tiers changes significantly.

In recent years, however, due to the JCQ awarding organisations' having been granted access to candidate level, national curriculum Key Stage 3 (KS3) test results[8], and their sharing of their GCSE result data, matching individual candidates' achievements at GCSE and GCE to their prior attainment at KS3 and GCSE respectively (with whichever awarding body they entered for) has allowed much more reliable, accurate and valid predictions to be made. In order not to presume a linear relationship between mean KS3 and the GCSE score, and to ensure that the relationship for the mid-ability candidates does not disproportionately

[7] It is not clear from the film how the script ranges for inspection were determined. At one point the commentary simply states: "Statistical evidence is used to determine where the grades might fall [i.e. the SRB] and samples of scripts either side of that mark are brought up for reassessment by the Chief [i.e. Principal] Examiners." It is likely, however, that the "SRB" would have been determined by the equi-percentile outcome from the previous year.
[8] The availability and usefulness of KS3 result data ceases, of course, in 2010.

affect the outcomes of candidates at the ends of the distribution, the individual level data are grouped in deciles for the purpose of calculating the predictions. A full regression approach would encounter both problems. However, little is lost by way of predictive power by adopting the grouped approach, which is also more manageable to operate.

The approaches are not without some minor problems, for example independent centres have tended not to take KS3 tests, thus eliminating them from the analysis. However, because they are predicated on individual candidate level data, these predictions explain far more of the variation in candidates' outcomes and thus form far more reliable (and valid) predictors than those based on centre level measures.

The increasing use of these type of predictions has, however, precipitated some discussions about the robustness of the predictions and the appropriateness of the guidance limits used to accompany them. The issue was raised, for example, at the final plenary session of the 2008 AQA GCE Chairs' conference.

A recent substantial technical investigation was undertaken by Pinot de Moira (2008) into the robustness of the predictions and the appropriateness of the guidance limits. It had been assumed that the main factor affecting the accuracy of the predictions was the correlation between the predictor and outcome variables. However, using a proportional odds model to predict grade outcome from prior achievement, Pinot de Moira found that this had relatively little influence. Instead, the confidence limits within which the prediction lies were found to be wider to the extent that: the distribution of prior achievement across the range was skewed, the entry size was small, and the nearer the prediction was to the midpoint along the scale. Moreover, she suggested that the guidance limits adopted by AQA were, if anything, on the generous side while concluding that *"the incumbent model is not without limitations but these, or similar, limitations would be present in any alternatives"* (2008, p. 28).

Nevertheless, different awarding organisations have used the predictions yielded by these analyses, especially those relating to GCSE outcomes, to varying degrees. For the new specification GCE awards, however, and with firm encouragement from Ofqual, they now all subscribe to using (and actively meeting, within Ofqual's imposed guidance limits[9]) predictions at subject level based on the national subject outcome the previous year, adjusted for variation in candidates' ability measured by their mean GCSE scores. Hitherto, a similar, inter-awarding body approach has only been employed at GCSE for the Science suite.

Despite this research and inter-awarding body collaboration, there is a residual, but growing concern in some quarters about the technical validity of the predictions, and the premium increasingly being put on them. Most recently these were expressed in a paper by Sofroniou and Pierce (2009) in which he argued for more evidence of robustness, testing of the underpinning assumptions and application of tolerances more appropriate to the uncertainties in the model, amongst others. Such research is to be welcomed, the outcomes of which would be used to refine - not reject - the model and its approach.

[9] Ofqual's guidance limits in 2009 were as follows (cf information in footnote 5):

| number of candidates | guidance limit |
|---|---|
| 1000+ | +/- 1% |
| 500 – 999 | +/- 1.5% |
| 201 – 499 | +/- 3% |
| 200 or fewer | None |

Using the national outcome as the basis for predictions has a double benefit if all awarding organisations apply them, yielding both inter-year and inter-awarding body comparability. It does mean, though, that the dominant awarding body in any subject in effect determines the national standard, with which the smaller ones have to comply. However, this would occur regardless of how the predictions were calculated, unless each awarding body was given equal weighting regardless of its entry, and that would cause swings in national outcomes following changes in centres' entry policies.

## 5. WHAT IS RIGHT WITH THE CURRENT METHOD?

Although all the research on examiner judgment has cast doubt on its reliability, there are also good reasons for being wary of approaches based solely on statistical predictions. Cresswell (2000), for example, identifies two major conceptual problems with such an approach. The first is that it will not control for all the determinants of candidate attainment, especially those such as motivation and quality of teaching which are likely to be linked to specification design. This raises the issue of what it is legitimate to take into account when predicting outcomes. For example, not rewarding improved performance in a particularly well taught or well supported specification because it does not fit the statistical model clearly challenges what we understand standards to mean. If the statistical model is adjusted to compensate, loss of absolute objectivity over time is incurred. The second problem relates to identifiable subgroups of candidates; if, for example, males and females added value differentially between the base and outcome measure, a shift in entry pattern between the genders would not be adequately reflected in the predicted outcome, even if the overall base profile measure remained unchanged.

Baird, Cresswell & Newton (2005) tackle these issues further. They argue it is not possible, nor possibly even desirable, to try and agree a single definition of standards, and proceed to describe four: weak criterion referenced; cohort referenced[10]; catch-all; sociological, of which the first best describes the current position. Essentially it is a criterion referencing approach but with appropriate allowance made to take into account the demand of assessment. Much of the research described above demonstrates how difficult examiners find this task, thus explaining the progressive use of supportive and guiding information, especially statistical data, to assist them. Indeed, many argue, with some justification given the results of recent research, that awarding is now, effectively "adjusted cohort referenced". If it is desirable to control for extraneous factors when providing statistical guidance to awards, it is surely better to do so as best and as fully as one can. The "catch-all definition", in which all relevant candidate and centre level variables are controlled for, is thus appealing. However, even this raises intractable, value-laden questions regarding possible interaction effects between the examination and control variables. In addition, not only are many of the required control data unavailable, even if they could be obtained the practicality of analysing them routinely for each specification and examination series would be prohibitive. More fundamentally, and linked to Cresswell's points referred to above, such a purely statistical approach lacks validity, hence compromising public credibility. At some point in the process expert judgement, taken on behalf of society, is needed to endorse the standards awarded (i.e. the sociological perspective).

The current approach as defined by the CoP thus embodies some vital features which must be retained. Primarily, the fact that groups of experienced examiners endorse the quality of work on given boundary marks, and thus the standard, confers a large degree of public

[10] Sometimes wrongly referred to as "norm referenced"

credibility, confidence and trust in the system, which is vital to its survival. Were the standards to be set wholly statistically, public confidence could soon be undermined. As Cresswell (1996) argued, the awarders, as a "guild of experts", undertake their role on behalf of society, which entrusts them to make good decisions. The argument is not that they should not be involved, and seen to be involved, in the process, but that the reliability of their judgements has to be assessed and balanced against that of the other information which can be brought to bear. The crisis of confidence in GCE in 2002 serves as a reminder both of the need for standards to be set using, and evaluated against, the best statistical evidence and also for expert judgement to endorse decisions.

As noted in section 3.1.2, on rare occasions, faulty data are used to produce the predictions. Although awarders do not usually recognise anything awry about the statistical prediction from their script inspection, on occasion they have done and remedial action taken in the form of fresh script ranges pulled. Were a purely statistical approach to be adopted, this safeguard, generally unreliable though it is, would be lost entirely.

Finally, not all predictions are equally valid or reliable. For example, in subjects with small entries, or where there has been a change in entry pattern between tiers – as has happened in the large entry GCSE modern foreign language subjects over the past decade – awarders' supposed familiarity of judgement has to be more trusted to ensure awards comparable to those of the previous year are made.

## 6. Structural, Procedural and Political Developments

Apart from questions related to the reliability of examiner judgement and the availability of statistical predictions, there are several structural, procedural and political issues which argue for a change in emphasis in awarding procedures.

First, grade boundary recommendations are made "at award", when not all of the mark data have been processed. Of course, if the examiner judgements were reliable it should not matter what proportion of marks were on the distributions. However, Ofqual's requirement that at least 85 *per cent* of marks must be fully processed for an award to be deemed valid suggests that it does not have full confidence in that reliability. In fact, AQA's research (Dhillon *et al* (2004)) indicates 75 *per cent* to be appropriate operationally in this regard, but even with 85 *per cent* of marks on the distributions, the outcomes (and, if defined in a certain way, standards) can change between the award and the final outcomes. Part of AQA's own practice, not required by the CoP, is the Pre-Results Checking Procedure (see section 2.4), on the basis of which, and with the Chair's agreement, boundary marks are sometimes changed to ensure the subject standard is maintained and yield a defensible award.

Second, the whole apparatus of award meetings is both expensive and time-consuming. In the 2009 summer series, AQA alone hosted 243 award meetings (lasting between one and five days: see section 2.1) comprising 3205 judgemental decisions. (Although this was something of an exceptional year, with separate award meetings for legacy and new specification GCE AS awards, the corresponding figures in 2008 were 183 and 2726 and, in any case, qualifications and specifications are revised regularly.) AQA recently calculated that the cost of a typical face-to-face award meeting, excluding awarders' fees, was £2,500. In an environment of increasing pressure to schedule later examination dates, the length of the current awarding season is a major obstacle to providing earlier results, especially if post qualification admissions (PQA) is introduced. Until recently, little had been heard about the government's aspiration to introduce PQA by 2012. However, in her speech to the

Headmasters' and Headmistresses' Conference, Kathleen Tattersall (Chair of Ofqual) (2009) gave strong support to the idea, suggesting GCE examinations could start directly after the Easter break. Were this intention to be realised, the pressure to mark, process and award GCE examinations earlier, and presumably in a shorter period, would be intense. The combination of electronic marking and remote awarding, with the primary focus on using statistical evidence, would allow results to be made available more speedily.

Third, features of the current system are arguably subtly productive of grade inflation. For example, such is the prevailing ethos among awarders of not wanting unfairly to disadvantage candidates, in both their individual and collective decisions they are likely to give the candidates the benefit of the doubt (Stringer, 2008). Other sources relate to, for example, the effects of marker review and special consideration cases, along with the inclusion into the system of late marks. (The latter effect used to be larger when examiners marked whole scripts because the better scripts tended to take longer to mark.)

Fourth, general qualification assessments are currently undergoing a major structural change from terminal, linear examinations to modular, unitised tests. This largely began with the introduction of Curriculum 2000 for the GCE and will largely be completed with the new specification GCSE specifications for first (full course) award in 2011. There is general, but not unanimous, agreement that standards reside at subject level in these modular assessments, which makes both statistical and judgemental approaches to awarding extremely difficult. Not only do awarders not see candidates' full work (in common with legacy linear specifications), at most awards they will not know the implications at subject level of their decisions without a strong statistical steer. Although this steer itself will necessarily be weaker than usual and subject to caveats regarding the effects of, for example, resits and candidate maturation, to retain any hope of maintaining subject standards in these unitised specifications, those standards (as well as the encompassing unit standards) must be defined by strong statistical evidence.

Fifth, Ofqual is shortly to be invested with increased powers and responsibilities as a fully independent regulator, accountable to Parliament. As such, it will have more powers and more responsibility for ensuring standards of public examinations are maintained. In order for it to perform its function it will presumably need a transparent mechanism for evaluating national awards, and ensuring comparability between years and awarding organisations. Recent history has demonstrated that the procedures required by the CoP alone do not meet that requirement, and that further approaches and criteria are needed.

Sixth, the requirements of the CoP appear to give predominance to the responsibility of awarding organisations to maintain standards between years, the requirement for maintaining inter-awarding body standards being more muted. However, in terms of fairness to candidates and centres, and the efficient working of the selection processes to higher education and graduate recruitment, it is arguably the latter which is the more important. Compliance with the CoP might ensure maintenance of an awarding body's standards, at least to some degree (although even this may be overstated), while inter-awarding body differences arise and persist. Ofqual's own recent practice has seemed to demonstrate agreement with this perspective. In 2008, for example, it requested AQA substantially to adjust its GCSE Science standards downwards in order for them to be aligned with the other awarding organisations, knowing that this precipitated a change in AQA, and national, standards from the previous year.

Seventh, the market for examination entries for general qualifications is increasingly competitive, with the ethos and strategies of even the awarding organisations with charitable status now being more business orientated. Originally, this market was largely geographical or sector based according to awarding organisation, but now those informal distinctions have disappeared, the market is more open. As a consequence there is an incipient pressure for awarding organisations to lower their standards in order to attract more entries from centres eager to raise their position in the DCSF Performance Tables. Whether or not such pressure is yielded to - and there is no evidence from any awarding organisation that it has been - the system is vulnerable to charges of "dumbing down", despite their being almost invariably unsubstantiated. However, in such an environment, awarding organisations need to be able, with Ofqual's support, individually and collectively, to be able robustly to defend their standards against such a charge.

Finally, and following on from the above, events in the last year or two have conspired to force CoP procedures to be overturned and, in effect, national and awarding body standards to be set primarily on the basis of statistical evidence. Apart from GCSE Science, as mentioned in section 4, in 2009 Ofqual required awarding organisations to produce awards in the new GCE AS specifications which were tightly constrained to the matched candidate, mean GCSE based predictions: +/-1 *per cent* from the predictions for the larger entry specifications (see footnote 9). It was also very stringent in only allowing two exceptions to this requirement: "It is impossible to obtain an outcome within tolerance" and "It is impossible to meet a prediction as a result of outcomes in February"[11]. This gave very little – if any - freedom for awarders' judgement within a unit – effectively the boundaries were established statistically and confirmed by the awarders. If awarders did want to recommend a boundary away from the SRB, they often had to compensate in the other unit in order to ensure the subject outcome was within the tight guidance limits. (In any case, given the evidence adduced earlier, doubt has to be cast on awarders' ability reliably to differentiate between two adjacent marks.) The consequence of this approach was that the post-award statistical screening exercise - generally agreed to be the most valid indicator of inter-awarding body standards – threw up not a single specification from the English and Welsh awarding organisations which was out of line with the national standard.

## 7. Discussion and Recommendations

The theme of this paper has been that current procedures for awarding general qualifications as prescribed by the CoP are no longer fit for purpose as they do not guarantee standards between years and, particularly, between awarding organisations. The argument has particularly hinged on the increasing realisation that awarders' judgements are unreliable compared to statistical predictions.

Implicit in the argument is the notion of multiple definitions of standards, as alluded to by, for example, Cresswell (1996) and Baird *et al* (2000), even within a solely statistical context. This paper has contrasted standards as defined by professional expert judgement (i.e. performance standards) and standards as defined by the best available statistical predictions[12] (i.e. outcome standards). Whilst having a degree of face validity, the former

[11] From Appendix 1 of a letter from Isabel Nisbet (Acting Chief Executive, Ofqual) to all Accountable Officers, June 29 2009.

[12] Outcome predictions have usually been based on candidates' prior attainment (viz. Section 4), although in this context, the definition is widened to include other possible approaches, e.g. common centres' analyses, candidates' concurrent attainment or even equi-percentile outcomes from the previous equivalent examination series. The most valid and reliable available statistical method for predicting outcomes, whatever it is, must be adopted.

definition, in also embodying low levels of reliability, lacks true validity. Consequently, it is recommended that the CoP adopt the latter definition of standards, and its awarding procedures be revised accordingly.

Although this seems like a radical proposal, it would, in fact, be largely formalising what has gradually become current practice in spite of the letter and spirit of the current CoP. It is tacitly assumed, indeed now even endorsed by Ofqual, that candidates taking a new or revised specification should not be penalised for relatively poor performance. As referred to earlier, identifying differences in performance while allowing for changes in demand is difficult to do, and especially so when - as is currently the case with the new GCSE suite - the structure of the specifications are radically altered and attempts are made to re-align coursework and written unit outcomes. In such circumstances, awarding using a purely judgemental approach on the basis of performance tends to create what is known as a “saw tooth” pattern of outcomes, the peaks of which occur towards the end of a specification’s life, and the troughs towards the start. If it is deemed acceptable to override the performance definition of standards in the early series (how many series?) of a specification in the interests of fairness to adjacent age cohorts of candidates, why would it not be acceptable to disregard apparent changes in performance due to other, non-legitimate factors? Even if it were possible for awarders to identify individual factors, which is highly doubtful, being able to distinguish between the effects of the legitimate and non-legitimate, and to measure accurately the degree of such effects, would be nigh on impossible; these would be, at best, very fine judgement calls.

The acceptance of statistical definition of standards is tacitly now becoming embodied in awarding practice, and also officially endorsed. As noted earlier, under the guidance of Ofqual, the 2009 GCE AS awards were heavily driven by statistical expectations, leaving little or no room for awarders’ professional judgement to demur. Similar plans are in place for the GCE A level awards in 2010 and the GCSE awards in 2011 (although the availability of appropriate predictor data for the GCSE awards is a concern). The GCSE Science awards in the last two years also provide examples of where awarders’ judgement was overridden in the light of statistical evidence, although these were by no means isolated cases. If such an approach continues to be applied, as seems likely and, AQA would argue, desirable, then the current method of awarding looks to be at best inefficient and at worst obsolete.

In 2007 Stringer (2007) successfully trialled two alternative, more statistically driven, awarding procedures on behalf of AQA: the “confirmation” and the "zone of uncertainty” methods. In the former approach awarders are initially presented with scripts on the SRB and asked to confirm (or otherwise) whether they appear to be of acceptable borderline quality. They should only vote against the SRB if there is substantial evidence that the standard of work differs considerably from that in the previous series. In the latter approach, a range of marks at the borderline is scrutinised as currently, and the SRB becomes the recommended boundary by default if falls within the zone. Stringer concluded that *“the confirmation method is undoubtedly the better of the two methods trialled in February 2007, as it is consistent with the argument that awarders should not be trying to find the boundary mark themselves but endorsing – where they can – the statistical estimate”.* Although both approaches are quite different from what the CoP currently requires, they are both CoP compliant to the extent that the awarders have seen, and collectively approved, the recommended mark as being of an acceptable grade standard.

In view of the arguments made in this paper, and the way in which current awards are required to be made to satisfy Ofqual’s requirements, it is proposed that the current awarding

procedures as described by the CoP be reconstructed around a more statistically driven approach. This would require the production of valid and reliable predictors which would be nationally coordinated and agreeable and applicable to all awarding organisations. The approach would not be applicable to small entry specifications but any predictive model and its parameters (e.g. guidance limits) could, and should, be refined in the light of emerging research evidence.

It is not the purpose of this paper, nor indeed the CoP itself, to catalogue the technicalities of the procedures and how they would work in practice (e.g. the basis of the agreed national predicted outcomes, the size of the guidance limits, the criteria for exemptions and how they would be awarded etc.). Rather it is changes to the broader principles and procedures which are at issue. In fact, these can be effected by omitting or amending relatively few of the paragraphs in Section 6 of the statutory CoP: *Awarding, marking review, maintaining and archive and issuing results* (although some minor adjustments should be made to other paragraphs – for example a "range of marks" should include the possibility of a single mark and "awarding meeting" would better be described as "awarding process".) However, to be confident of comparable standards both between awarding bodies within a given year, and between years within a given awarding body, it is essential for there to be a central production, stipulation and evaluation of expected outcomes. The current proposals (or something similar) implicitly assume such a mechanism ("a nationally agreed predicted outcome" – see later) as they will only be successful, in terms of greater transparency and trust in the system, if such a scheme is implemented.

The Appendix contains the relevant five paragraphs of Part 6 of the CoP, together with proposals as to how they might be changed better to reflect the argument of this paper and, increasingly, the spirit and actuality of current practice.

Finally, an objection to adopting this approach would likely be that genuine improvements (or deteriorations) in performance improvements would be masked. However, as the description of the 'saw tooth' phenomenon above indicates, even the current approach method does not allow for them, and in any case because genuine changes in performance tend to be marginal and incremental, awarders would be unlikely to be able to identify them between two consecutive series. Over longer periods standards become increasingly relative in nature as the nature of subject requirements, and what is deemed to be rewardable behaviour, change. Using grades to measure absolute performance standards over a long period of time is, therefore, largely meaningless. In any case, as was pointed out earlier, the main focus for candidates and centres is not comparability of standards between series, but between awarding organisations in a given series.

Thus, although this paper is couched in terms of an argument for a change of procedure, it implicitly addresses wider issues. These not only relate to how standards are defined, but more controversially, to the extent - or more accurately, the limits - of what the current system can be expected to deliver in terms of monitoring improvements (or deteriorations) of educational attainment, and what that notion means in any case. The formal impossibility of objectively comparing educational standards over time are well rehearsed, see, for example, Jones (1999). (There is, perhaps an argument for another paper to accompany this one, entitled something like: "Awarding GCSE and GCE - time to come clean about what the Grades mean?")

Currently, critics of the system – typically sections of both the press and politicians – argue that the standards of general qualifications are being lowered, and that the awarding

organisations are at least complicit in what they consider to be a conspiracy. Interestingly, when referring to standards in this context they have in mind levels of educational attainment, although they use rising outcome rates as evidence for their case[13]. In response, defenders of the system (typically the awarding organisations the regulators) attempt to counter such criticisms on the ground from which they were made – ultimately relying on the awarders' 'tick chart' evidence as a defence - rather than acknowledge, and engage in a discussion about, what the system is/is not capable of delivering and intended for. It is confidently hoped, of course, that what this paper proposes would spike the guns of the critics' artillery by yielding greater stability of outcomes between years and awarding organisations. Nevertheless – and perhaps because of that – the implementation of this proposal would provide a good opportunity for the awarding organisations to engage with the press and politicians in a frank and open debate about these matters. Awarding organisations should not, however, abdicate their responsibility to engage with, and undertake research into, the quality of educational standards over time. However, insofar as it is possible to conduct rigorous research in this area, awarding meetings are not appropriate forums. A secondary benefit of the proposed approach might, therefore, be that more resources could be channelled to undertaking such research.

B E Jones
November 2009

## REFERENCES


Baird, J. (2000). Are examination standards all in the head? Experiments with examiners' judgments of standards in A level examinations. *Research in Education*, 64, 91-100.

Baird, J., and Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact.* AQA Research Paper, RPA_05_JB_RP_077.

Baird, J., Cresswell, M.J., and Newton, P. (2005). Would the real gold standard please step forward? *Research Papers in Education,* 15(2), 213-229.

Baird, J., and Morrissey, M. (2005). *The association between statistical recommendations and grade boundary judgments.* AQA Research Paper, RC101.

Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein and A. Heath (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.

Cresswell, M.J. (1997). *Examining judgments: Theory and practice of awarding public examination grades*. Unpublished PhD thesis, University of London Institute of Education.

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

Dhillon, D., Eason, S. and Pascoe, J. (2004) *Percentage of marks on file at Awarding: Consequences for "post-awarding drift" in cumulative grade distributions*. AQA Research Paper, RC258.

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

Jones, B. E. (1999) *Response of the AQA Research Committee to two QCA research reports: Five Yearly Review of Standards in 16+ Examinations 1976-1996 and GCE A level standards: a review*. AQA Research Paper, RC21.

Pinot De Moira, A. (2008). *Statistical predictions in award meetings: how confident should we be?* AQA Research Paper, RPA_08_APM_RP_013.


[13] There was a notorious occasion in 1984 when the pass rates in A levels had increased again, causing the Daily Telegraph's headline to read "*Standards falling*". A week later the GCSE pass rates fell marginally and the Daily Mail also reported "*Standards falling*"!)

Office of the Qualifications and Examinations Regulator (Ofqual), Department for Children, Education Lifelong Learning and Skills (DCELLS), and Council for the Curriculum Examinations and Assessment (CCEA). (2009). *GCSE, GCE and AEA Code of Practice*. London: Qualifications and Curriculum Authority.

School Curriculum and Assessment Authority and Office of Standard and Education. (1996). *Standards in public examinations 1975-1995*. London: SCAA.

Scharaschkin, A. and Baird, J. (2000). The effects of consistency of performance on A level examiners' judgments of standards. *British Educational Research Journal*, 26 (3), 343-357.

Sofarianu, N., and Pierce, G. *Review of methodology used for predictive modelling of GCE AS awards 2009.* Joint Ofqual/JCQ Standards and Tehnical Issues Group meeting, 16 September 2009.

Stringer, N. (2007) *Evaluation of the February 2007 Alternative Awarding Procedure Trials.* AQA Research Paper, RPA_07_NS_RP_039

Stringer, N. (2008). *Are we successfully maintaining GCE A Level standards?* AQA Research Paper, RPA_08_NS_RP_018.

Tattersall, K. Speech to the Headmasters' and Headmistresses' Conference (HMC), 6 October 2009. Adelphi Hotel, Liverpool.

*The Awards meeting*. (1985) [Video]. Middlesex, England: Focus in Education Ltd.

# APPENDIX. EXTRACTS FROM PART 6 OF THE STATUTORY CODE OF PRACTICE: AWARDING, MARKING REVIEW, MAINTAINING AN ARCHIVE AND ISSUING RESULTS, AND RECOMMENDATIONS FOR CHANGES

6.4 *After the marking and moderating period, principal examiners and principal moderators must propose preliminary ranges of marks for each component/unit as a basis for locating key grade boundaries on the basis of judgement. Full details of the key grade boundaries are given in appendices 3 and 4. All other grade boundaries are determined arithmetically. These proposals must be made following consideration of sufficient candidates' work (marked scripts and/or internally assessed material) to gain a feel for candidates' performance.*

Recommendation: Omit this paragraph. It is not necessary for Principal Examiners to suggest ranges.

6.11 *The process of awarding must be conducted by establishing the range for each key grade boundary and, subsequently, marks at each grade boundary for each externally and internally assessed unit/component. A record of the preliminary ranges of marks proposed by the principal examiners and principal moderators must be included in the report of the awarding meeting. Where the nature of particular specifications requires modifications to be made to the procedures set out below, arrangements will be agreed between the regulators and the relevant awarding body.*

Recommendation: Omit the second sentence of this paragraph (see 6.4. above).

6.12 *The awarding committee must consider candidates' work, selected on the basis of the range for each key grade boundary. These ranges must be based on the preliminary ranges of marks proposed by principal examiners and principal moderators and also must take account of the relevant technical and statistical information. If necessary, marked scripts and internally assessed material outside the preliminary ranges must be included to ensure that work of the appropriate standard is considered.*

Recommendation: The second sentence of this paragraph should read: "These ranges must be centred around putative boundary marks which would yield a subject outcome which matched the nationally agreed predicted outcome."

6.15 *The awarding body must provide the awarding committee with procedures that are consistent with this code. These must be used to conduct the award and may be set out as an agenda for the meeting. The following must be used as appropriate, to inform the determination of marks at key grade boundaries:*

*Qualitative*

*i. copies of question papers / tasks and final mark schemes*
*ii. reports from the principal examiner(s) / principal moderator(s) on how the question paper functioned.*

iii. *archive scripts and examples of internally assessed work (including, in appropriate subject areas, photographic or videotaped evidence) at the relevant grade boundaries, together with relevant question papers and mark schemes*
iv. *samples of current candidates' work (marked scripts and/or internally assessed material) distributed evenly across key boundary ranges for each component, with enough representing each mark to provide a sound basis for judgement so far as the size of entry and nature of work permit. The material should be selected from a sufficient range of centres where work has been marked/moderated by examiners/moderators whose work is known to be reliable.*
v. *any published performance descriptions, grade descriptions and exemplar material, where available*
vi. *any other supporting material (such as marking guides for components where the evidence is of an ephemeral nature)*

*Quantitative*

vii. *technical information, including mark distributions relating to the question papers / tasks and individual questions for the current and previous series, where available*
viii. *information on candidates' performance in at least two previous equivalent series, where available*
ix. *details of significant changes in entry patters, choices of options, and prior attainment, where available*
x. *information on centres' estimated grades for all candidates including:*
   – *qualification-level estimates for linear (including linear unitised) specifications*
   – *unit-level estimates for externally assess units in all other unitised specifications*[14]
xi. *information about the relationship between component/unit level data and whole-subject performance, where available*

*Regulatory reports*

xii. *relevant evidence from the regulators' monitoring and comparability reports.*

Recommendation: The extent of the proposed changes to this paragraph depends on the interpretation of the third sentence: "The following must be used, as appropriate, to inform the determination of marks at key grade boundaries" (emphasis added). If this means that awarders are not required to scrutinise all this information, as ultimately it is the Chair who determines the key grade boundaries, the paragraph can remain as it is. However, to the Quantitative section must be added the provision of putative boundary marks which, when taken together, would yield a subject outcome which would match the nationally agreed prediction. If, however, the paragraph means that all awarders must have access to this information, then the Quantitative section should be a "could have" rather than a "must have" requirement. Awarders' judgements of quality of work should be made irrespective of, for example, the shape of the mark distribution from which it is sampled.

[14] For units that are entirely composed of multiple choice questions, information on centres' estimated grades for all candidates will only be collected for a period that includes to summer awarding series.

6.16 *Awarders must consider candidates' work in the range for each key boundary, ensuring that a sufficient amount of candidates' work is inspected. They must consider each mark in turn, as follows.*

*i. First, working down from the top of the range, awarders must identify the lowest mark for which there is a consensus that the quality of work is worthy of the higher grade of the boundary pair. This forms the upper limiting mark.*

*ii. Next, working up from the bottom of the range, awarders must identify the highest mark for which there is consensus that the quality of work is not worthy of the higher grade. The mark above this forms the lower limiting mark.*

*Awarders must then use their collective professional judgement to recommend a single mark for the grade boundary, which normally will lie within the range including the two limiting marks. This judgement will include consideration of the evidence listed in paragraph 6.15. All awarders must have considered candidates' work at the recommended mark.*

Recommendation: Omit this paragraph and replace it with one requiring the putative grade boundary (SRB) to be recommended if it were confirmed by the awarders, or was in their zone of uncertainty. Although this amendment would allow more freedom in unit awarding procedures – either the confirmatory or zone of uncertainty approach could be used – it would effectively dispense with current practice and make the Chair's recommendation a formality. However, even within the constraints of the current CoP, different awarding organisations interpret this paragraph slightly differently and thus adopt different procedures.

A less stringent version would allow the Chair also to recommend the grade boundary, in effect allowing current practice to be retained. However, the Chair's role in this area would be more clearly defined than currently; the current CoP is somewhat ambiguous about the degree to which the Chair is a representative or delegate of the committee and its views. Under this proposal, the Chair would be a delegate and fully own the recommendation.