AWARDING IN THE 21ST CENTURY - A VIRTUAL MODEL

Lesley Meyer, Jo-Anne Baird, Neil Stringer, Lynne O'Sullivan and Carolyn Adams

SUMMARY

Awarding meetings are essentially unchanged since their inception, with Senior examiners coming together to look at statistical information and candidates' work. Modern techniques permit different approaches to this process whilst preserving the validity of the process by incorporating the views of subject matter experts.

A virtual award meeting will be held using Virtual Classroom technology, with Senior examiners contributing to the meeting remotely. Senior examiners scrutinise scripts remotely and their judgements are collected electronically. The Chair of Examiners and staff establish zones of uncertainty and prepare other materials for the virtual award. The normal Awarding Committee are presented with the collated information and discuss it over a synchronised telephone conference.

Scanning of the scripts and remote scrutiny opens up the possibility of including other stakeholders in the qualitative judgments process, such as examiners, teachers and even students. The source of judgements would need to be delineated, but it is possible that standard setting could be made more transparent to stakeholders.

More modern approaches offer the opportunity to make the most of the Senior examiners' expertise and make the process more efficient and robust.

WHY CHANGE?

Senior examiners' time is a scarce resource in the UK assessment industry. There is a finite number of experienced assessors in each subject area and, with modular examinations, there are lots of demands upon their time to write assessments, quality control assessments, standardise examiners' marking, quality control examiners' marking and develop new syllabuses. Demands are spread throughout the year, but there are seasons when the demands are very high, as the turnaround between students sitting the assessments and results being issued is short and the volume is high – 8.5 million assessment marks were processed by AQA in 2005 in about 8 weeks. Standard setting typically takes two day's time for approximately 8 Senior examiners and the travelling takes about a day on average in total.

Current standard setting processes do not utilise Senior examiners' time effectively. Although quantitative information plays a large part in the standard setting process ('awarding') in the UK, most of the time is spent scrutinising a small sample of students' work from the current and previous year's assessments to ascertain the quality of candidates' performances and set the cut scores. There is ample research evidence to indicate that, as in any field (Dowie & Elstein, 1988) these qualitative judgements are not robust (Baird & Dhillon, 2005). Recent research indicates that Senior examiners cannot distinguish well, qualitatively, between the work of students in a small range of marks. Essentially, the task that is required of Senior examiners is not possible even for subject matter experts because students on a

given mark will have reached that mark by different routes through the question paper, scoring highly on some questions and not so well on others, and are therefore variable in terms of their worthiness for the grade in question (Scharaschkin & Baird, 2000). Variability of quality on the same mark is large enough to overlap with quality on adjacent marks. Baird and Dhillon (2005) showed that if the marks are removed from the examination scripts, senior examiners could not successfully rank order them within a seven mark range. Thus, requiring Senior examiners to make these fine distinctions is not a good use of their expertise. Neither does requiring them to travel to face to face meetings use their time wisely.

Marking processes would benefit from more quality control from the Senior examiners. UK assessments largely involve short answer and essay style questions, rather than multiple choice questions. As such, there are significant challenges in assuring the marking process. Our definition of 'true mark' resides in the Senior examiner's standard of marking for a given assessment. Appeals over marking are a small proportion of the overall marks issued (0.78% in AQA) and fewer still are upheld (0.015%), but with more resources, quality control processes during the tight timescale in which the marks are collected from examiners and reviewed could be made more rigorous.

Value for money. It is incumbent upon Awarding Bodies to ensure that they provide a value for money service to the country and the Government has indicated in its 14-19 Implementation Plan (DfES, 2005a) that it will be reviewing the value for money from the UK assessment industry. As such, AQA is re-designing this process with a view to ensuring that standard setting is at least as robust as the current system, but creates efficiencies: potentially in terms of Senior examiners' time, staff time, accommodation costs, travel costs and time for this process to be conducted. Time savings will be essential for future plans for post-qualifications admissions systems to be introduced for UK University entrance – a policy currently being debated in the UK (DfES, 2005b).

So – why change? In summary, we have identified two problems with the current awarding system, both of which relate to the burdens placed on already busy senior examiners. In order to reduce the time awarders need to spend in awarding meetings, we propose reducing the numbers of scripts examiners are asked to review. In order to reduce the time spent travelling to the meetings and reading through the scripts in the meetings themselves, we propose holding virtual meetings which senior examiners can attend in their homes and providing scripts electronically in advance of the remote discussion.

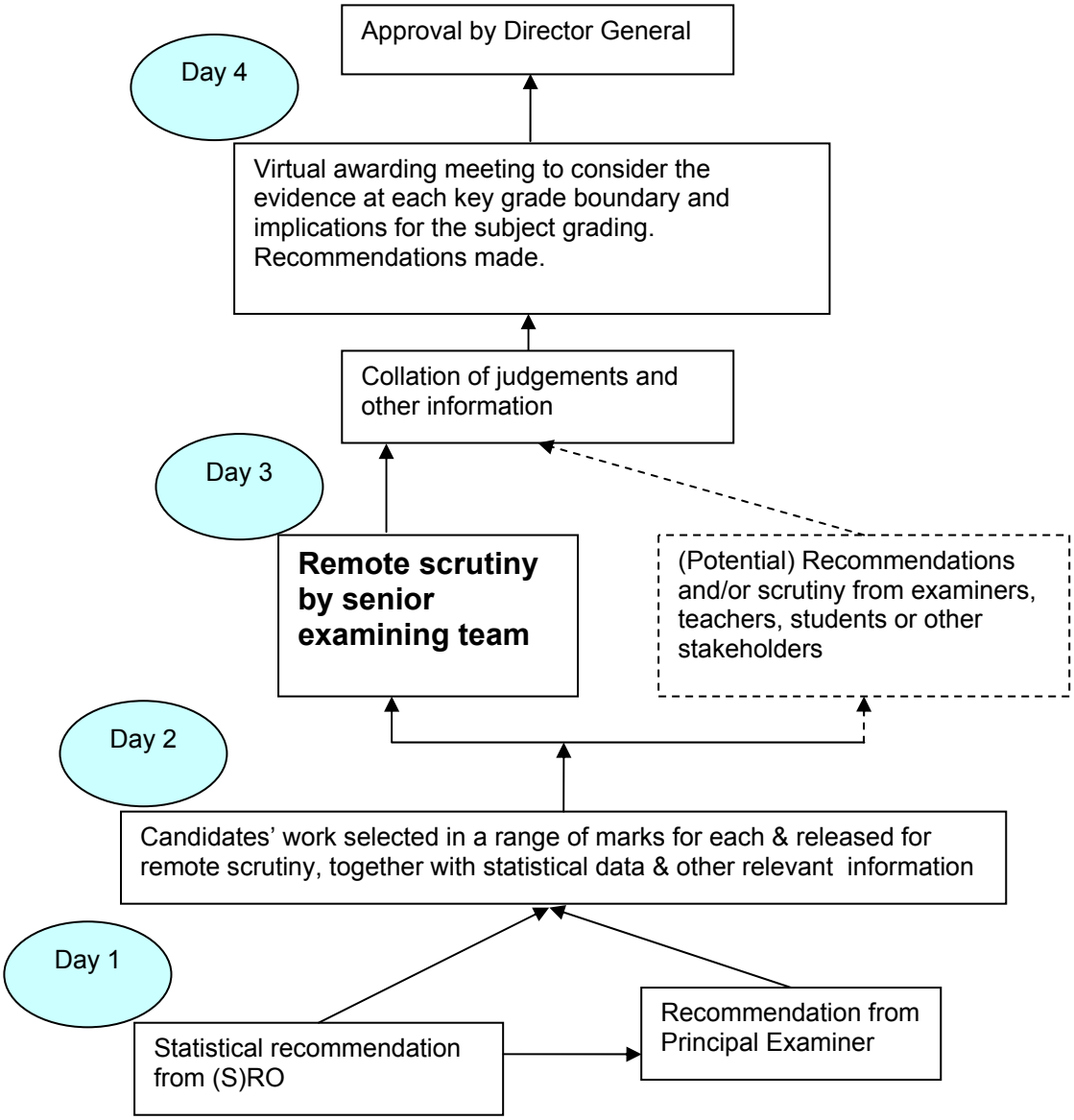
A VIRTUAL AWARDING MODEL

A description of the current awarding model accompanies this paper (Meyer, 2005). A flowchart outlining how the basic process could operate is shown in Figure 1. Essentially, the process could remain largely unchanged, except that

- candidates' work is scrutinised remotely,
- discussion regarding the recommendations for the grade boundaries is conducted using Virtual Classroom technology, and
- the awarding meeting itself would last approximately half a day.

Face to face awarding meetings typically last two days, so more award meetings could be scheduled in a particular time period using the same amount of resources as currently. However, this change in process model permits a range of other possible alterations to current procedures, which are discussed below.

Figure 1: Potential model for virtual awarding



POSSIBLE CHANGES TO THE QUALITATIVE JUDGEMENTS PROCESS

1. How many candidates' examination scripts should individuals scrutinise remotely?
Scrutiny of a large sample of candidates' work is unnecessary, as the work has already been marked. The purpose of the scrutiny is to give the senior examiners a qualitative impression of the candidates' work on given marks. To reduce the burden of this process upon senior examiners, it would be desirable if, individually, examiners scrutinised fewer scripts than is currently the case. There are various ways in which that could be achieved without undermining the quality of information gained from this part of the process. The following questions cover some possible scenarios for scrutiny at each judgemental grade boundary:

a) Should individuals scrutinise more than one script per mark point in the range?
Often at present, members of the awarding committee scrutinise more than one script per mark point for reassurance, but this may not be practicable (or, indeed, necessary) in a remote situation and in the context of a re-designed awarding process. As long as, like

now, the full range is covered by the scrutiny team as a group there may be no need to make each individual scrutinise more than one script per mark.

b) Should individuals work through the full range?
Requiring each individual to scrutinise one script per mark in the seven mark range would enable all individuals to cover the range fully, but is probably excessive.

c) Should individuals scrutinise alternate marks?
Making individuals scrutinise scripts on alternate marks across the range would enable all individuals to work systematically across the range, but reduces the scrutiny time from that in option b) and could therefore be an attractive option. If carried out literally, some would receive three scripts, others four, but that is not necessarily a problem.

d) Should individuals be allocated a set number of scripts, selected at random from the range?
Individuals could be allocated an agreed number of scripts (four, for example), sampled at random from the range. Some individuals could, at random, be allocated four high-scoring scripts and others four low-scoring scripts for a particular boundary, but between them the scrutiny team would cover the full range. Practically, four is probably the maximum. For example, consider

i) a typical future A-level
4 scripts per judgemental grade boundary x 2 judgemental grades (A and E) x 4 question papers = 32 scripts per individual;

ii) a typical GCSE with two written papers
Foundation tier: 4 scripts per judgemental grade boundary x 2 judgemental grades x 2 components = 16 scripts per individual;
Higher tier: 4 scripts per judgemental grade boundary x 3 judgemental grades[1] x 2 components = 24 scripts per individual; 40 scripts in total per individual.

Allocating only one script to each individual per grade boundary decision would be a risky option. Not only would it place overwhelming emphasis on the provision of 'suitable' scripts (e.g. scripts on which candidates had a reasonably balanced performance across the question paper, etc.), but more importantly the scrutineer would not be presented with much opportunity to compare candidate performance in *this year's* scripts (they would only be able to compare the single script with the archive).

2. In what form should scrutineers' judgements be recorded?
It would be perfectly adequate to continue to follow the standard tick chart format used for recording awarders' judgemental decisions within AQA, but there are alternatives.

a) Standard tick chart notation
Each individual would have to send their decisions electronically, on all the scripts they scrutinised at each judgemental grade. A grid adapted from the current TICK CHART RECORD (Appendix 1) could be used for this purpose by the examiner concerned and the existing template could also be used by the AQA Officers when combining each individual's judgements onto an overall tick chart summarising decisions for all the examiners involved in the scrutiny for each boundary decision. The standard notation used could be continued, i.e. ✓='in the grade', ✘ ='not in the grade', ?='unsure'. Once compiled, the zone of uncertainty would need to be drawn, presumably in the standard manner, at a meeting involving the Chair of Examiners and senior AQA staff. Further in-depth discussion of the judgemental evidence, statistical data (and any other recommendations from teachers, HE experts etc. if obtained, see again Figure 1) would then lead to a final recommendation for each grade boundary.

b) Numeric coding of decisions

[1] An important point to consider as part of the development of a remote scrutiny process is how grade D on Higher tier could be dealt with (and, if still appropriate, grade B on Intermediate tier), as for these grades the standard setting procedure currently requires comparison with *current* scripts on the same grade on another tier, rather than with archive scripts. This hurdle is documented as a problem for future consideration here and in Point 6, Table 2.

An alternative to the standard coding would be to record decisions numerically. This is attractive as it enables the decisions to be manipulated arithmetically, for example an average score of scrutineers' decisions could be calculated for every mark in the range. These average scores could be directly interpreted and could therefore more fully inform the decisions regarding the zone of uncertainty.

An example of how this could operate is shown in Table 1. Eight examiners, working remotely, have each been allocated three scripts, taken across the mark range, to scrutinise for this particular grade boundary. For each script, the examiner records one of three numeric scores:
2 = definitely worth the grade
1 = borderline
0 = definitely not worth the grade

For each mark in the range, the mean average, median and mode are calculated across the examiners. As shown in the example, average scores could also be calculated for each examiner (but, on only three scripts, is not particularly informative). Further, there is no reason why this approach could not be used in (currently more standard) situations where each examiner scrutinises more than one script per mark. Modelling of this method on archive 'tick charts' has shown that the boundary decisions may be clearer and would not be different from those actually reached.

Table 1: alternative, numeric summary of examiner decisions for each script scrutinised in the range

	Examiner										
Mark	Exr 1	Exr 2	Exr 3	Exr 4	Exr 5	Exr 6	Exr 7	Exr 8	Mean	Median	Mode
70	2		2	2		1			1.75	2	2
69		2			1		2	2	1.67	2	2
68	1	2		1					1.33	1	1
67		1	1		1	1			1.00	1	1
66	1						1	0	1.00	1	1
65				1			1	0	1.00	1	1
64			0		0	0			0.00	0	0
Mean	1.33	1.67	1.00	1.33	0.67	0.67	1.33	0.67	SE(Mean)	0.14	

For a typical borderline mark, it would be expected that the mean, median and modal averages would all be equal to (or close to) 1. If the mode was zero for any mark, the majority of examiners would have felt it was definitely not worth the grade. Alternatively, if the mode was 2 the majority would have felt it was definitely worth the grade, whereas the aim, of course, is to find the mark which the majority felt to be borderline (i.e. for which the mode was 1). Likewise, if the median for a mark was not equal to 1 it could not (or should not according to judgemental evidence) be the borderline mark, as this would imply the majority of examiners scored it either as definitely out (0) or definitely in (2). The more the mean average creeps above 1 the greater the tendency for scrutineers to feel that the mark is in, and vice versa as the mean average sinks below 1.

Thus, discussions at the Approval Meeting[2] would sensibly centre on the marks in the range for which the mode and median are both 1 (marks 65-68 in the example above), and the main focus would probably tend towards the mark(s) where the mean average also is as close as possible to 1 (65-67, which could be deemed the zone of uncertainty). Obviously this interpretation (and the zone) would also need to be informed by the statistical evidence and any other judgemental evidence from HE experts, teachers and the Press, if obtained. Provisional and Final approval would follow as necessary (Figure 1).

3. *Alternative scrutiny process which could be used to shorten the scrutiny time and help determine a 'Chair's zone of uncertainty' (for standard awarding meetings):*

[2] between the Subject Officer, Support Officer and Chair and Approver

a) Each examiner would be expected to work down from the top of the range, scrutinising scripts (probably) mark by mark until they reach the point where they are no longer certain that the script is in the grade. They would be expected to work individually and at their own pace, thus not necessarily all scrutinising the same mark at one time. They would then fall to the bottom of the range and would work up, mark by mark, until they reach the point where they are no longer certain that the script is unworthy of the grade. At this point the individual stops scrutinising scripts, having determined his/her own zone.

b) A tick chart would be drawn up by the Officers as normal (assuming, for simplicity, that the standard ✓, ✘ and ? convention is used on the TICK CHART RECORD form). This could take a form similar to Table 3.

Table 3: Example tick chart for an alternative scrutiny process for standard awarding meetings (preferable result)

	Mark	Awarder 1	Awarder 2	Awarder 3	Awarder 4	
Upper limiting mark 69 ⇨	70	✔	✔	✔	✔✔	
	69	✔	✘✔✔	✔✔✔	?✔	
	68	✔?	✔?✔	✔✔✔✘		⇦ Chair's upper limit 68
Lower limiting mark 66 ⇨	67			?✔?		
	66		✔?	✘✔✘		⇦ Chair's lower limit 66
	65	✔✘	✘✘	✘	✔✘	
	64	✘✘	✘	✘	✘	

Awarder 1 read scripts down to a mark of 68 before becoming uncertain of the worthiness of the scripts for the grade. He/she then started at 64 and worked upwards, stopping at 65 as they became uncertain whether the scripts on that mark were 'out'. Awarders 2 and 4 had similar experiences but came up with their own areas of uncertainty. However, Awarder 3 reached 67 before becoming uncertain. Moving upwards from 64, he/she reached 66 before doubt set in. Thus this awarder covered the full range – which would not be the aim of this process.

The Chair and committee would then draw the zones of uncertainty based on this grid (for example, 66 could be the lower limit and 69 the upper limit in the example). The committee would then discuss the best mark to recommend within the range 66, 67, 68 or 69, with a view to the statistical evidence. Having determined the mark the Principal Examiner for that question paper would be asked to select acceptable archive scripts on the selected mark.

For example, if the statistical evidence suggested that 67 was the appropriate boundary mark, the discussion could follow the lines of, "given that the zone of uncertainty spans 66-69, there is no reason not to recommend 67 as the grade boundary mark" (assuming that scripts on 67 were confirmed as acceptable). If the statistical suggestion was 65 (unlikely as the initial range should then have gone lower, but for the purposes of illustration), the discussion could be, "65 is not considered acceptable judgementally but, as the zone of uncertainty spans 66-69, 66 would be the appropriate compromise" (again assuming scripts on 66 were confirmed as acceptable).

The aim of this process would be to reduce the length of time taken for scrutiny. Occasionally the Chair may feel that the original zone of uncertainty can be narrowed – for example here to 66-68 rather than 66-69, following discussion and incorporating all other available evidence. This revised zone could be termed the 'Chair's recommended zone'. However, it is easy to envisage a situation where all awarders still find themselves covering the full range (see Table 4 below), which would defeat the object of changing the scrutiny process. Only permitting awarders to scrutinise one or two scripts per mark point could ensure the awarders reach areas of uncertainty before covering the full range, but this could cause the awarding teams to feel that the rigour of the process was being sacrificed for speed.

Table 4: Second example tick chart for an alternative scrutiny process for standard awarding meetings (unsatisfactory result)

Mark	Awarder 1	Awarder 2	Awarder 3	Awarder 4
70	✔	✔	✔	✔✔
69	✔	✘✔✔	✔✔✔	✔ ✔
68	✔✔	✔? ✔	✔✔✔✘	✔
67	✔✔✘ ✔	✘✔ ?	? ✔ ?	✔✘ ?
66	✔✔ ?	✘✘ ?	✘✔✘	✘✘
65	✘✔✘	✘✘	✘	✘✔✘
64	✘✘	✘	✘	✘

4. *A 'confirmation' method is also possible:*
   In circumstances in which the Principal Examiner and the statistics are pointing towards similar marks, the committee could simply be asked to scrutinise scripts on a particular mark to ascertain that it is an acceptable boundary mark. Rejecting the proposed boundary mark would have to be a possible outcome of the script scrutiny, so it would have to be feasible for examiners to request additional scripts to scrutinise. This method would be most efficient of all those proposed, but may not be suitable in cases where there are low numbers of candidates sitting the examination, as the statistics may not be dependable. Further, where the assessments have changed substantially, the statistics may not indicate any changes in the examination performances of candidates, due to changes in the teaching of the subject in the context of the new syllabus and examination structure, including the preparation for teachers associated with those changes.

5. *Pros and cons of remote scrutiny*

While there are many good reasons to consider remote scrutiny, there are also some disadvantages. In particular, there are specific procedural issues which would need to be overcome (see Table 2).

Table 2: Advantages and risks of a remote scrutiny process which requires awarders to review fewer scripts than in the current model

Advantage		Risk/Problem
Faster scrutiny (fewer scripts scrutinised per examiner)	But…	Puts more weight on each examiner's qualitative judgements, rather than interaction between examiners
Promotes a more focussed final discussion between fewer parties	But…	Limited opportunity to develop team spirit as part of the award meeting
Creates potential for extending the qualitative involvement to other external sources (HE experts etc.), thus more democratic?	But…	Senior Examiners may feel their expertise is less valued
Costs saved on awarding meetings, photocopying, travel, accommodation, etc. Awarding process is modernised alongside innovations in assessment & marking processes	But…	Procedures for awarding coursework and portfolios would need to be considered Fact to face meetings are probably necessary for new subjects and we are re-developing all of the A levels currently. There would be entirely new question papers for the first two years of an A level.

CONCLUSION

Changes to the way in which the qualitative judgements are made, collected, recorded and analysed are not necessarily dependent upon the virtual awarding process. Trials of the confirmation method and zone methods could be conducted in ordinary paper-based awarding meetings and this is currently being given consideration within AQA. There are plans for the trialling of virtual award meetings in each of our three offices in live awards this summer. Contingency plans are clearly being given a priority, as it is essential that the grading of the examinations is not put in jeopardy by these innovations to the process. Evaluation of the pilots will be conducted, with a focus upon the impact of the technology,

including the effect of this change upon participants in the meeting and its impact upon the robustness of the standard setting process.

References

Baird, J. & Dhillon, D. (2006) *Qualitative expert judgements on examination standards: valid, but inexact.* AQA Internal Report, Paper RPA_05_JB_RP_077.

DfES (2005a) *14-19 Education and Skills*. Department for Education and Skills publication. http://www.dfes.gov.uk/publications/14-19implementationplan/docs/14-19%20Implementation.pdf. Accessed 3 March 2006.

DfES (2005b) *Improving the Higher Education Applications Process – A consultation paper.* http://www.dfes.gov.uk/consultations/downloadableDocs/Equality%20Impact%20Assessment.doc. Accessed 3 March 2006.

Dowie, J. and Elstein, A. (1988), *Professional judgment: A reader in clinical decision making.* Cambridge University Press.

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

Meyer, L. (2005) *A basic guide to standard setting.* http://www.aqa.org.uk/over/pdf/guidetostandardsetting.pdf Accessed 3 March, 2006.

Lesley Meyer, Jo-Anne Baird, Neil Stringer, Lynne O'Sullivan & Carolyn Adams

AQA

APPENDIX: TICK CHART RECORD

COMPONENT: _____

Grade:	Last year's agreed mark and cum %:		SRB mark and cum %:		Cum % predicted this year:	
	PEx/PMod recommended range:	to	Zone of uncertainty (LL to UL):	to	Recommended mark:	

Initials / Mark	Cum %																

COMPONENT: _____

Grade:	Last year's agreed mark and cum %:		SRB mark and cum %:		Cum % predicted this year:	
	PEx/PMod recommended range:	to	Zone of uncertainty (LL to UL):	to	Recommended mark:	

Initials / Mark	Cum %																

To assist the approval process, please indicate the Chair (C) and the Principal Examiner (PE) or Principal Moderator (PM) on each tick chart, next to their initials.

Leave the full set of awarding documentation with the Director's PA who will organise for it to be scanned and placed onto the awarding database. (Harrogate: take the original to the Processing Department who will then organise photocopying and scanning for approval).