EVALUATION OF THE FEBRUARY 2008 CONFIRMATION METHOD AWARDING TRIAL

Chloe Davenport

SUMMARY

This report presents the findings of the Confirmation Method awarding trial carried out on four subjects in the February 2008 series. The Confirmation Method requires awarders to look at a script on the SRB and decide if it is certainly worth the grade, borderline worth the grade or certainly not worth the grade. It has been argued that as awarders are unable to differentiate between scripts within a small range of marks, and because they already rely on statistical evidence provided to them before recommending a mark away from the SRB, they should need to find compelling evidence against accepting the SRB. Quantitative analyses of the data show that there was a non-significant difference in the outcome of recommended boundaries between the method trialled and traditional awarding methods highlighting that awarders are already highly influenced by the SRB. Awarding time, as well as the width of zone of uncertainty, was found to be greatly reduced using the Confirmation Method. Feedback was gathered from the awarding committee at the end of the trial and is discussed. Smooth running of all meetings was reported. Some committees commented that they would have felt more comfortable had they been allowed to look at more scripts on each mark; the increased comparisons with archive material was considered to be beneficial and also focus in meetings was more easily maintained. The report concludes that looking at two scripts on each mark might be an improvement and a development of the method in future trials.

BACKGROUND

The rationale for the Confirmation Method is that the relative difficulty of two examination papers is what primarily accounts for the average difference in performance between two populations of candidates taking two versions of an exam when differences in their population characteristics have been accounted for. Therefore, the relative difficulty of the two papers (defined as it is here) cannot be estimated quantitatively through script scrutiny.

Awarding using the Confirmation Method involves asking awarders to decide whether scripts on the Statistically Recommended Boundary (SRB) are of a suitable standard to be consistent with the same grade in previous series. Awarders are asked to look at only one script on the SRB and compare it to archive scripts to ensure comparability. Awarders need to decide if the script they look at on the SRB is certainly worth the grade, borderline worth the grade or certainly not worth the grade. It is argued that without the statistical information provided, awarders would be unable to decide accurately a script's quality in terms of examination difficulty between two successive years, which is precisely what must be done to maintain standards. This is also because there is evidence showing that awarders cannot discriminate between scripts within a small range of marks sufficiently well to be able to identify the boundary precisely (Baird and Dhillon, 2005). The Confirmation Method streamlines the awarding process which, it is argued, will facilitate large scale capitalisation of other modernisation initiatives within AQA, such as electronic marking (CMI+). At present, it would be impossible to implement CMI+ across all components as script scrutiny would require

facilities that AQA cannot accommodate such as considerable IT and space, to allow on-screen scrutiny, or the mass printing of scanned scripts. These problems would be reduced through the use of the Confirmation Method, since the method requires awarders to scrutinise fewer scripts, Modernisation will also be critical for the viability of the Government's proposed move towards a system of Post-Qualifications Applications (PQA) to higher education by 2012.

Stringer (2007) trialled the Confirmation Method in eight subjects, on one unit from each. It was found that the Confirmation Method greatly reduced awarding time; in total it was estimated that if the Confirmation Method was to be used in all subjects the number of awarding days would be cut from 316 to 204, a saving of 35 *per cent*. The width of zone of uncertainty also reduced by an average of 1.22 scripts per grade in each unit. The distance of the recommended boundary from the statistically recommended boundary was not found to alter significantly between the traditional method of awarding and using the Confirmation Method in awarding. Generally, awarders tend to go with the mark of the SRB or a mark adjacent to this. This did not alter using the Confirmation Method.

QCA granted permission to deviate from the Code of Practice where the trial procedures required, for example the requirement to establish upper and lower limiting marks. QCA also waived the need for an awarding meeting to be reconvened if a grade boundary is subsequently moved to a mark on which scripts had not been scrutinised during the awarding meeting, providing that the mark would have been within the normal five-mark range scrutinised under normal procedures.

THE PROCEDURE

The following section details the Confirmation Method instructions given to awarding committees this series.

Introduction

The Confirmation Method is not simply the standard AQA awarding procedure with the difference that awarders consider one mark at a time. The rationale for the Confirmation Method is that the relative difficulty of two examination papers is what accounts for the average difference in performance between two populations of candidates taking two versions of an exam when differences in their population characteristics have been accounted for. Therefore, the relative difficulty of the two papers (defined as it is here) cannot be estimated quantitatively through script scrutiny. The function of script scrutiny in the Confirmation Method is to ascertain whether the statistical estimate of difficulty that has produced the statistically recommended boundary appears to be sensible. Accordingly, it should be stressed to the awarding committee that they should only vote against the SRB if they believe that there is significant disagreement between the standard of work at the SRB and the established standard of work at the grade boundary being considered, i.e. that demonstrated by the archive.

Procedure

Script scrutiny begins with each awarder looking at one on the SRB. They each indicate whether they think the work is:

Certainly worth the grade (coded "2")

Borderline worth the grade (coded "1")

Certainly not worth the grade (coded "0")

Their judgements are then recorded on a tick chart (see Table 1) where the index of uncertainty (median rating) is calculated. It requires more than half the committee to vote in the same direction to move from the SRB. In other words, the index must be 2 or 0 to warrant looking at the mark above or below the SRB, respectively. Values of 0.5, 1, and 1.5 indicate that the balance of opinion does not warrant a move from the SRB.

In the case shown in Table 1, the SRB is confirmed because the index of uncertainty indicates that the work of the mark 36 is broadly comparable to the established standard of work at the grade boundary.

Table 1:

Mark /Awarder	Chair	PEx	Chief	Awarder 1	Awarder 2	Index
38						
37						
36 (SRB)	1	2	1	0	1	**1**
35						
34						

Where the index of uncertainty indicates the same direction of change, as in table 2, the process should be repeated using scripts on the appropriate adjacent mark; in this case 35. The process concludes when the committee reaches a mark with an index other than 0 or 2.

Table 2:

Mark /Awarder	Chair	PEx	Chief	Awarder 1	Awarder 2	Index
38						
37						
36 (SRB)	2	2	1	2	1	**2**
35	1	2	1	1	1	**1**
34						

It should be noted that the procedure set out above differs slightly from that used in an earlier trial. Feedback from the February 2007 trial revealed an objection to the use of the word "substantially" before the terms of stronger or poorer. It was argued there were never going to be substantial differences between scripts on marks around the SRB so coding was changed to certainly worth, borderline worth or certainly not worth the grade. Another change made to the procedure this time was that committees were guided to go with a majority rule by taking the median rating from the tick charts.

UNITS INVOLVED IN THE TRIALS

Table 3 shows the subjects involved in this trial. In the previous February 2007 trial only one unit from each subject was trialled, however this time the trials were carried out on the whole subject awarded in February. Subjects were selected from those that had trialled it previously so they were already quite familiar with the method. In the subjects used, the committees had not experienced problems in the 2007 trial.

Table 3: The subjects and units involved in the trials

Subject	Units	Office
GCE PE	PED1, PED 2, PED4.	Guildford
GCE Law	LAW1, LAW2, LAW3, LAW4, LAW5	Guildford
GCE Spanish	SP01, SP02, SP3T	Harrogate
GCE French	FR01, FR02, FR3T	Harrogate

It should be noted that for the GCE French and Spanish Speaking units (FR3T and SP3T) grade boundaries are usually carried forward from the previous series because the assessment has not (arguably) changed, SRB's are typically the same as the carry-forward mark and awarders rarely recommend a change from these.

QUANTITATIVE ANALYSES[1]

Width of the Zone of Uncertainty

Table 4 shows the width of zone of uncertainty using the Confirmation Method. Using the Confirmation Method the zone of uncertainty is defined as the range of marks on which scripts are scrutinised. Using the standard AQA procedure the zone of uncertainty is the range of marks between and including the upper and lower limiting marks on the tick chart. This is required by the Code of Practice, paragraph 6.15 (QCA, 2005).

[1] Information used to produce the statistics for the width of zone of uncertainty and the distance of the recommended mark from the SRB was taken from the awarding documentation.

Table 4: Width of zone of uncertainty

Subject	Unit	A	E
GCE PE	PED1	1	1
	PED2	1	1
	PED4	1	1
GCE Law	LAW1	1	1
	LAW2	2	1
	LAW3	1	1
	LAW4	1	1
	LAW5	1	2
GCE Spanish	SP01	1	1
	SP02	1	1
	SP3T	2	2
GCE French	FR01	1	1
	FR02	1	1
	FR3T	1	1
Mean		1.14	1.14

In the trial carried out in February 2007 the zone of uncertainty was compared to that of the same units from February 2006. It was found that there was a reduction in the zone of uncertainty of approximately 1.34 marks, down to an average of 1.32 marks. The results this time reflect similar findings with an average of 1.14 marks[2]. These figures suggest that in most cases the awarders were happy with the SRB. The two units where awarders looked at scripts other than the SRB occurred in GCE Law. In unit 2 awarders looked at the mark below the SRB, and in unit 5 awarders looked at the mark above the SRB. In both cases awarders did not recommend the SRB and decided upon a mark adjacent which they looked at. In no unit did awarders need to look at scripts on more than two marks – the SRB and 1 mark +/- from this.

It is a Code of Practice requirement that, if a boundary is moved outside of the range considered in the awarding meeting, the meeting is reconvened so that the awarding committee can provide qualitative evidence at that mark. Ordinarily, a proposed boundary change would be highly unlikely to move outside of the range of scripts scrutinised in the meeting. However, a proposed change of one mark could invoke this procedure if the meeting had used the Confirmation Method. For the purpose of these trials, QCA waived this requirement.

[2] Since different units were included in the February 2007 and February 2008 trials any comparisons in outcomes are not direct.

Distance of the Recommended Mark from the Statistically Recommended Boundary

Table 5 contains the number of marks the recommended boundary lies from the SRB. The sign in brackets indicates whether the change from the SRB was above or below it[3]. Signs are not included in the calculation of means.

Table 5: Distance of recommended mark from the SRB.

Subject	Unit	A	E
GCE PE	PED1	0	0
	PED2	0	0
	PED4	0	0
GCE Law	LAW1	0	0
	LAW2	(-) 1	0
	LAW3	0	0
	LAW4	0	0
	LAW5	0	(+) 1
GCE Spanish	SP01	0	0
	SP02	0	0
	SP3T	0	0
GCE French	FR01	0	0
	FR02	0	0
	FR3T	0	0
Mean		0.07	0.07

As can be seen in most cases the majority of decisions were the same as the SRB, with the only two units that differed being the adjacent mark to the SRB. The mean distance for both grades was 0.07 marks. This result is similar to the findings of the February 2007 trial in which the average distance of the recommended mark was 0.13 marks from the SRB; the effect of a difference of two marks in one unit. When these results were compared to the differences of the same units conventionally awarded in February 2006 the differences were negligible suggesting that the Confirmation Method does not have a large impact on how close the recommended mark is to the SRB. Awarding committees already tend to recommend the SRB or an adjacent mark: in the Summer 2007 award season of the 1208 GCE judgemental grade decisions required to be made, 90.9 *per cent* were the statistically recommended boundary or +/- 1 mark from the SRB.

Time taken to make decisions

Table 6 shows the time taken by each awarding committee to decide on the recommended marks. Support Officers in each meeting were asked to record the time taken for awarders to scrutinise scripts plus any discussion and decision making.

[3] A positive number indicates a recommendation above the SRB and a negative number indicates a recommendation below the SRB.

In the case of FR3T and SP3T, the speaking units for GCE French and Spanish respectively, the boundaries are typically carried forward and awarders came to the meeting having already listened to two candidate's tapes on the SRB at home, which could explain why the times for these units are much lower than for any other units. This was a deviation from the procedure, which stated that awarders should scrutinise one script (or tape) at the SRB.

Table 6: Time taken[4] to decide on recommended boundaries

Subject	Unit	A	E	
GCE PE	PED1	13	13	
	PED2	12	10	
	PED4	12	8	
GCE Law	LAW1	9	9	
	LAW2	17	10	
	LAW3	7	11	
	LAW4	11	13	
	LAW5	14	20	
GCE Spanish	SP01	15	10	
	SP02	15	10	
	SP3T	5	5	
GCE French	FR01	6	7	
	FR02	8	11	
	FR3T	2	2	
Mean		9.73	9.27	9.50

This method was found to greatly reduce time spent on script scrutiny and boundary decision making compared with the traditional AQA procedure creating a more efficient awarding process. The mean time taken to make all judgemental grade decisions this year using the Confirmation Method was 9.50 minutes, comparable to that of last year where the mean times were 11.6 minutes for the Confirmation Method and 27.3 minutes using the traditional method. A reduction in the mean of this size in both 2007 and 2008's trials demonstrates that using this method, awards that would normally take more than one day could now require fewer days. Reducing the number of days awarding could create a number of options including compressing the awarding period which could lead to the possibility of results being released earlier. This will be critical for the Government's proposal of Post Qualifications Applications (PQA) to higher education to be successful. The earlier release of results would aid PQA as it means candidates would be able to start looking for alternative courses sooner and universities will have a longer period during which to select those candidates applying late. Nevertheless, other considerations regarding earlier release of results would need to be made including the number of expert markers available in subjects and position in the examination timetable. Those subjects which are large yet have a shortage of expert markers and those towards the end of the exam series might struggle to meet the earlier deadlines

[4] In minutes.

due to a shortage of the necessary data, meaning compressing the awarding timetable might be more problematic than it first appears to be. Increased pressure would also be placed on staff such as those extracting and copying scripts to be used in awarding meetings. An alternative option would be to keep the awarding period at the same length of time but having less overlap of awards relieving the time pressure of approvers and those employees extracting scripts and preparing for award meetings; this would however prevent the early release of results.

QUALITATIVE ANALYSES

In this trial all members of the committee were given feedback sheets to complete at the end of awarding. Those participating were asked to provide feedback on their experiences of the Confirmation Method in this trial. These included the Support Officer, the Subject Officers and the Awarding Committee. In this section, their feedback will be discussed.

Many awarders felt that the Confirmation Method required and facilitated a greater focus on comparisons with archive scripts. With the traditional awarding process this can sometimes be neglected and awarders might start to compare live scripts rather than looking for a script at a certain grade and comparing to archive scripts. Greater focus on comparisons with archive scripts also led to more confidence in the recommended boundaries as it is felt decisions will be more accurate in terms of maintaining standards from a previous series. As the Principal Examiner for GCE Law (Unit 2) stated, "*This exact comparison with archive scripts will mean complete consistency with the standard of the previous session*." *S*ome people felt the Confirmation Method might be most useful for large, established specifications, but not so much for new or small specifications, where grade boundaries tend to be less stable over time.

In all the subjects in which the Confirmation Method was trialled, committee members were satisfied that decisions remained at the same level of quality as traditionally despite the reduction in time spent on script scrutiny. Some of the benefits of the Confirmation Method included not being distracted by scripts that are clearly not 'in', a reduction in time pressure and, due to the shorter time, a greater ability to maintain concentration than when using standard procedure. It is a danger that if awarders have to look at many scripts on many marks in awarding meetings that by the final unit they are tired and decisions made will not be as thorough and detailed as they might have been at the beginning. The Principal Examiner for GCE PE (unit 6) commented, "*Due to the shorter nature of this process I found it easier to maintain focus throughout and apply equal rigour across papers*". The Chief Examiner of GCE PE believed that this method lends itself to online awarding.

Similar to the February 2007 awarding trials, committee members were positive about the method and how it greatly reduced meeting time; however they did suggest the method could be improved by looking at two scripts on a grade boundary. In GCE French it was stated that, although looking at one script avoided the confusion which sometimes occurs when looking at a range of scripts i.e. awarders being less certain of their decision after looking at a range of scripts, it could be helpful to look at two scripts. Awarders said looking at one script did not allow for much discussion and they sometimes needed the option to look at more than one script to make overall judgements. This is especially true at the start when it is difficult to get a flavour for the quality of scripts at a mark, resulting in a great reliance on archive scripts. The Chief Examiner for GCE Law believes there still remains a considerable difficulty in establishing with confidence relative standards based on the variable eccentricities of one or

two scripts. Committee members stated that, even when there was not a problem with the SRB, they would have felt happier having another script on the SRB to look at to confirm the boundary. The Principal examiner in GCE Law (unit 3) noted that if a marker had been harsh and a script is particularly good for a boundary there was a feeling the boundary was too high – a problem that could be reduced by looking at an extra script on the mark to see if this was actually the case.

As with the feedback from the February 2007 trials, some awarders suggested looking at one mark either side of the SRB as a safety check. The problem with checking the mark below the SRB is that, whatever the SRB, some scripts on the mark below it might be as good, or better, simply because of marking tolerance. Similarly, some scripts on or just above the boundary will be worse than some scripts below it. This is undesirable but, short of severely restricting the forms of assessment used in general qualifications, for example to objective tests, there will frequently be an element of interpretation involved in applying a mark scheme to a script. What is more, scripts on adjacent marks may *look* as good as those on the SRB simply because the awarders cannot make such fine distinctions between adjacent scripts (Baird and Dhillon, 2008). Looking at scripts on adjacent marks when the SRB appears satisfactory would defeat the purpose of the Confirmation Method.

CONCLUSION

Overall, the trials of the Confirmation Method were successful. The unit and subject outcomes of the awards involved were satisfactory and no different to those we might have expected had the standard procedure been used. The awarding meetings were completed in shorter times than is customary, but at no point did any of the personnel involved express concerns about the quality or thoroughness of the process; in fact, many awarders expressed confidence in the decisions they reached, thanks to a greater awareness of the archive scripts during script scrutiny.

The feedback received from awarders was, on the whole, more positive than that received in February 2007. This is perhaps because the committees involved had the benefit of experience on this occasion. Many awarders commented that looking at more scripts on a mark could improve the Confirmation Method of awarding; this was also an outcome of the February 2007 trial. Looking at two scripts on a mark could be an improvement, especially in subjects with small awarding committees and with papers containing many optional questions, where awarders may feel uncomfortable seeing few or no examples of work on some of the questions. As such it is recommended that any future trial, or operational use, of the Confirmation Method involves scrutiny of two scripts per mark per awarder.

The time data from this trial suggest that the summer awarding schedule could be reduced greatly, with the average time to recommend each grade boundary being just under ten minutes compared with almost thirty minutes using the standard procedure in February 2007 (although looking at two scripts will increase - though not double - this time). Savings on this scale, when translated into days spent awarding, could play a critical part in achieving the Government's proposed introduction of a system of Post Qualifications Applications to higher education by 2012. Also, in 2009 and 2010, several new qualifications will be awarded for the first time, including the new suite of GCEs and the Diploma, both of which comprise portfolio units that will take a great deal of time to scrutinise. These awards will be made in addition to existing specifications, including overlap with the current suite of GCEs. If the Confirmation Method were used to award the existing GCEs, which generally have well-established grading

standards and stable grade boundaries, the time saved could be devoted to the more demanding task of setting standards in the new qualifications.

The operational viability of the Confirmation Method, does, however, depend upon a practical amendment to rules set out in the Code of Practice concerning late boundary changes. It is currently a requirement if a proposed boundary change is outside the range considered in the meeting, the meeting be reconvened. As it stands, the reduction in the marks at which scripts are scrutinised brought about by using the Confirmation Method could increase the likelihood of invoking this procedure.

Chloe Davenport
Research and Policy Analysis
May 2008

References

Baird, J. A., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact* (RPA_05_JB_RP_077). Guildford: Assessment and Qualifications Alliance.

Stringer, N. (2007). Evaluation of the February 2007 Alternative Awarding Procedure Trials (RPA_07_NS_RP_039). Guildford: Assessments and Qualifications Alliance.