12th Annual Conference of the Association for Educational Assessment-Europe (Belfast 2011), 10-12 November 2011

Presentation by Graham Hudson DRS Data Services Limited, UK

Title: New Mechanisms for Assuring Marking Quality – A Framework to Support Secure Assessment Outcomes for Learners

ABSTRACT

The presentation will build upon research presented to the AEA-Europe in Olso, 2010, in which markers who may require additional guidance and training to keep to agreed marking standards can be identified early in the marking process.

The quality control framework, based upon percentage double marking, has been applied to a further 50 examination components with a view to developing a common framework within which marking quality can be managed more consistently. The outcomes of the further work will be presented.

Marking reliability in high-stakes assessments is fundamental to the use of assessment data for summative and formative purposes. As the majority of the UK Unitary Awarding Bodies provide information for learners taken from examination performance, it is vital that this information can be relied upon. Electronic marking not only provides the means to capture the data but also robust processes to manage the quality.

PRESENTER

The presentation will be given by Graham Hudson, Director of Electronic Assessment for DRS Data Services Limited, UK.

Background

DRS has successfully implemented electronic marking with a number of awarding body clients in the UK, the largest of which is AQA. The general benefits of using electronic marking are becoming more widely recognised both within the UK and internationally.

Key to the approach adopted by DRS and its clients is the focus on improving the quality of marking through the use of technology. Marking judgements made by senior examining personnel, combined with sophisticated algorithms, enable those marking standards to be built into a marking process that continuously checks marking standards with a regularity that could not feasibly be achieved in a paper-based system.

In addition, those awarding bodies that have embarked upon exploring electronic marking have found that the change programmes initiated have led to a wider review of operational processes, leading to further streamlining and improvement that may not have been envisaged when considering electronic marking initially.

This paper provides an update to AEA-E members of further research work building upon the paper presented at the 11th Annual Conference of the Association for Educational Assessment in Oslo in 2010 which set out a framework for assessing the quality of markers' marking making use of percentage double marking. The conceptual framework described at that conference has been worked up into a potential operational model.

Further detail and examples will be provided during the conference presentation.

Key benefits of electronic marking

Electronic marking makes use of scanned images of candidates' examination and test scripts to support the marking process. Images of candidates' scripts are held securely and distributed as questions, or parts of questions, to markers for marking across the Internet. Marks are captured at the time of marking and checking of marking standards takes place in real time.

Use of the images of candidates' answers now provides many more degrees of freedom to support more rapid processing of marks and a variety of quality control measures. Paper-based systems are constrained by the physical limitations of the scripts – which can only be in one place at a time.

By dividing the candidates' scripts into segments, electronic marking provides significant improvements over conventional marking by:

- removing marking bias, related to the leniency or severity of a marker's judgement for an individual candidate and for groups of candidates;
- enabling markers to focus on topics related to their expert knowledge;
- allowing markers to focus only on marking and not be diverted by administrative or procedural matters;
- marking that does not meet the appropriate quality tolerances can be identified in real time and markers stopped from marking that item and provided with further training;
- removing clerical errors (such as addition errors by markers and transposition errors to marksheets) inherent in a paper-based system.

In addition, logistical benefits can be derived, such as:

- reduction in script movement and risk of loss before marking;
- ensuring that all answers are marked;
- removal of clerical errors and inputting of marks manually;
- supporting candidate reviews, re-marking and awarding though the access to images of candidates' scripts;
- high level of visibility of how marking is progressing and meeting of key deadlines.

The most fundamental improvement, however, is enabling the regular checking of marking quality.

In addition, other processes can be supported, such as providing an electronic training resource to markers to augment or substitute the current marker standardisation meetings that take place prior to marking. This electronic process is commonly known as e-Standardisation.

Implementation in the UK and internationally

During the past 9 years, all UK Unitary Awarding Bodies that provide school and further education qualifications have piloted or implemented electronic marking. A number of other, professional awarding bodies have also followed suit.

In the UK, at least 9m candidates' scripts have been scanned, imaged and marked on a PC by markers during the summer 2011. Organisations, such as DRS, have worked with awarding bodies to put in place the necessary technical infrastructure, change management, training and programme management to support the annual increase in the number of scripts processed in this way.

As a result, all major UK awarding bodies are committed to this approach and have seen the benefits identified above realised with the examiners, schools and colleges, candidates and parents.

Interest has also been expressed internationally, with DRS conducting marking pilots in Australia, the Caribbean, West Africa, Malaysia and Poland.

Quality control and electronic marking

The most common types of examination papers fall into two categories:

- candidates write their answers onto the question paper in spaces left for prose, mathematical formulae, diagrams or graphs (*constrained answer booklets*);
- candidates write their answers in free-form essay style onto a lined answer booklet without specific structure (*unconstrained answer booklets*).

Segmenting answers in a constrained answer booklet is straightforward, and all recognised electronic marking systems support this approach. Segmenting answers in an unconstrained booklet is more difficult as it is not possible to pre-determine where a candidate will begin and end an answer, although DRS has devised an approach to achieve this.

The approach to quality control between constrained and unconstrained will need to be different, as free-form answers tend to be longer, cover several pages and include more judgemental elements to mark. This is unlike the constrained answers which are shorter and tend to have more structured marking guidelines. The benefits of this approach enable accuracy of marking to be checked regularly by the system and any marking deficiencies addressed promptly.

Quality control for 'unconstrained answers'

For unconstrained answers, which tend to be long and cover several pages, an approach to quality control needs to take into account the time taken for a marker to read an answer and to make a judgement. A 'seeding' approach, of pre-marking items which are delivered subsequently to markers to check marking standards would not be appropriate for practical reasons.

As a result, DRS has developed a set of algorithms and associated business rules that will combine the benefits of regular quality checking with those of double marking.

In so doing, a number of issues have had to be addressed, such as:

- against what standard will markers' marking be compared;
- if quality control is gauged by checking marking standards between markers, what happens to a marker when no other markers are marking;
- if mark difference exist between markers, which marker is deemed to be 'correct';
- and how does poor marking ultimately be identified and a marker stopped.

Applying the quality control framework

Annex 1 provides a summary of the terminology used during the remainder of this paper.

Annex 3 provides further details of the work carried out previously and has proposed a means for identifying *'a poor marker'* at an early stage in marking[1].

This work has been extended in the latest study to focus on producing a set of guidelines that can be used to help in deciding:

- how much double marking should be carried out (ie what level to set the *percentage double marking* level), and
- how many discrepant marks should a marker be allowed before they are prevented from marking any further (ie exceeded the *penalty cap*)?

So, first how often a discrepant marks are found in an *average marker's* allocation of work needs to be established. This had been identified using multilevel modelling on the data sets available from markers who have been monitored using *percentage double marking*.

[1] *Building a quality control framework for electronic marking*, Graham Hudson and Tom Benson, 11th Annual Conference of the AEA-E, Oslo, 2010

For each marker the items that he or she has marked that have been checked by a peer marker are identified. At this point, there are now three possibilities:

1. The score given by the marker is within tolerance of the score given by the peer reviewer. This is defined as **discrepancy level 0**.

2. The score given by the marker is outside tolerance of the score given by the peer reviewer but when further reviewed by a senior marker is found to be within tolerance of the score the senior marker awards. This is defined as **discrepancy level 1**.

3. The score given by the marker is outside tolerance of the score given by the peer reviewer and under further investigation by a senior marker is found to be outside tolerance of the score the senior marker awards. This is defined as **discrepancy level 2**.

Multilevel modelling[2] is now used to explore the relationship between the level of discrepancy as determined above and the maximum number of marks available for the item, the tolerance level and whether the item had related parts.

The multilevel aspect of the model was used to account for the fact that certain markers may be better at staying within tolerance than others.

The coefficients from multilevel modelling are shown within the table below. Within this model the outcome variable is the level of discrepancy with possible values of 0, 1 or 2.

Table 1: Results of multilevel logistic regression

Effect	Coeffi-cient	Standard Error	Signifi-cance	Odds Ratio
Fixed Effects				
Intercept - Out of tolerance with peer (2 mark item)	-2.196	0.094	0.000	0.111
Intercept - Out of tolerance with senior (2 mark item)	-3.199	0.096	0.000	0.041
Number of marks available on item	0.424	0.034	0.000	1.528
Number of marks available on item squared and divided by 10	-0.049	0.006	0.000	0.952
Adjustment for 1 mark items	-1.275	0.286	0.000	0.279
Tolerance	-1.614	0.129	0.000	0.199
Random Effects				
Variance between markers	0.093	0.027	0.001	

Results show that the probability of marks awarded being out of tolerance depends upon the maximum numbers of marks available for the item. In order to improve the fit of the model and to more fully capture the nature of this relationship, a quadratic term was also included along with an adjustment for items with just one mark available.

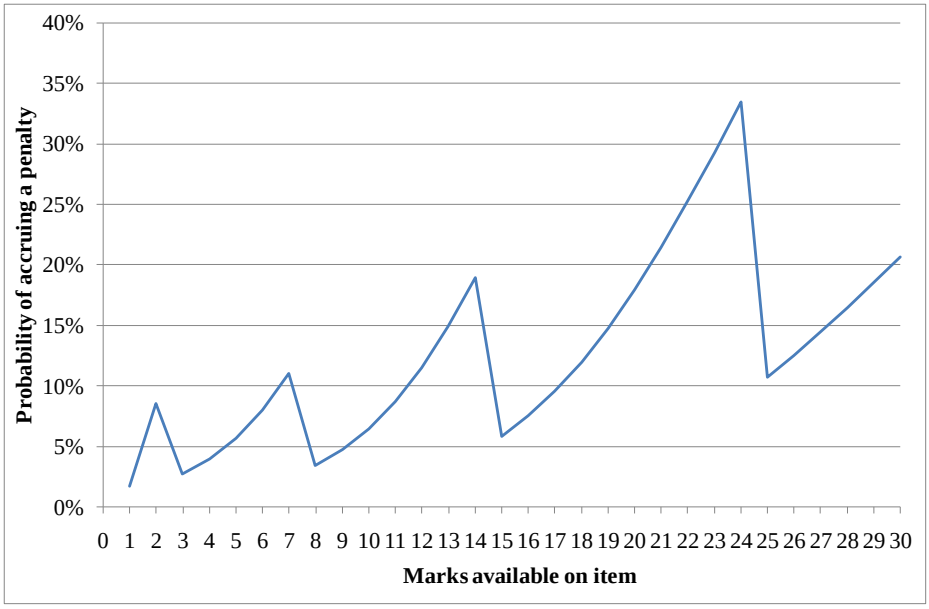
Results also show that increasing the tolerance for any given item decreases the probability of markers being out of tolerance with one another. Analysis showed no significant relationship between whether an item had related parts and the chances of an average marker being out of

[2] To be precise a proportional odds logistic multilevel model was used.

tolerance with peer and senior markers and so this piece of information was not included within the final model.

The model also identified significant variation between markers with certain markers being more likely than others to be out of tolerance.

The results in **Table 1** are summarised in the chart below. This shows how the probability of an average marker accruing a penalty on any item changes according to the number of marks available on an item. As can be seen, generally the probability of accruing a penalty increases with the total marks available on the item although troughs occur corresponding to increases in tolerance for items with larger numbers of marks available[3].



Given the probability calculations described above attention can now be turned to calculating the optimal penalty cap. This is done by looking at the probabilities of an *average marker* and a *poor marker* exceeding the *penalty cap*. As described in **Annex 2**, the aim is to determine the *optimal penalty cap* that will maximise the difference between the probability of an average marker exceeding this cap and a poor marker exceeding this cap (that is, it will maximise the marking confidence). This can be done mathematically.

The outcome of this computation has shown that there is a fixed ratio between the *optimal penalty cap* and the number of items that have been peer reviewed. The fixed ratio is dependent only on the probability of an average marker accruing a penalty and the probability of a poor marker accruing a penalty (which in turn can be calculated from the definition of a poor marker). Thus for any given amount of peer review, the *optimal penalty cap* can be calculated.

[3] In order to complete this chart it was necessary to specify the tolerances associated with items with different numbers of available marks. For the purposes of this chart it was assumed that items with 1 or 2 marks had a tolerance of zero, items with 3 to 7 marks had a tolerance of one, items with 8 to 14 marks had a tolerance of two, items with 15 to 24 marks had a tolerance of three and items with 25 to 30 marks had a tolerance of four.

The percentage double marking tool

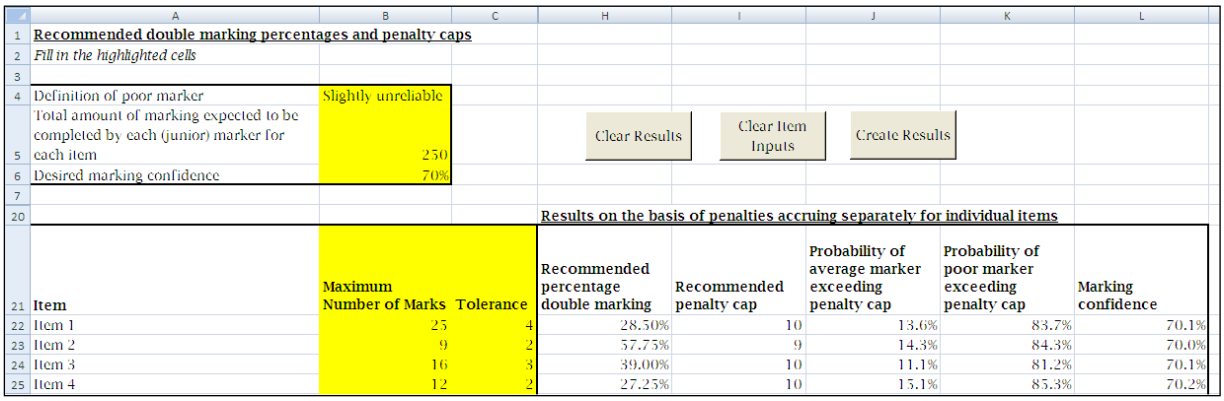
From these results, a 'proof-of-concept' tool has been developed allows the input of:

- the definition of a poor marker for that examination component;
- the total number of items to be marked, and
- the desired level of marking confidence.

The output from the tool will provide, for each item in the test:

- the recommended double-marking percentage;
- the recommended (optimal) penalty cap, and
- the level of marking confidence achieved.

A screen shot of the tool is shown below:



	A	B	C	H	I	J	K	L
1	Recommended double marking percentages and penalty caps							
2	*Fill in the highlighted cells*							
3								
4	Definition of poor marker	Slightly unreliable						
5	Total amount of marking expected to be completed by each (junior) marker for each item	250		Clear Results	Clear Item Inputs	Create Results		
6	Desired marking confidence	70%						
7								
20				Results on the basis of penalties accruing separately for individual items				
21	Item	Maximum Number of Marks	Tolerance	Recommended percentage double marking	Recommended penalty cap	Probability of average marker exceeding penalty cap	Probability of poor marker exceeding penalty cap	Marking confidence
22	Item 1	25	4	28.50%	10	13.6%	83.7%	70.1%
23	Item 2	9	2	57.75%	9	14.3%	84.3%	70.0%
24	Item 3	16	3	39.00%	10	11.1%	81.2%	70.1%
25	Item 4	12	2	27.25%	10	15.1%	85.3%	70.2%

In this example, the *poor marker* has been defined as 'slightly unreliable'. This equates to twice as likely to exceed the penalty cap as an average marker. The *marking allocation* is set to 250 of each item and a *desired marking confidence* has been set to 70%.

For item 1, the mark tariff and tolerance has been input already and the tool shows that the recommended settings that result are:

- percentage double marking, 28.5%;
- penalty cap, 10
- giving a marking confidence of 70.1%.

The tool also allows for adjusting settings to model the effects of double marking and penalty cap values.

In practice, the tool would be used to set an appropriate *percentage double marking* rate for each item in a test. As mark tariffs and tolerances vary, *percentage double marking* rates will also need to vary – which can be supported in the e-Marker® application. Until the development of the tool, making the settings had been based largely on judgement and the outcomes of previous marking exercises. The tool now allows for a more consistent approach to making settings which are now driven from empirical information from previous marking series. This does not remove the application of professional judgement. It does mean that the management of marking quality is given a real boost in terms of a data

framework that is straightforward to use and explain and reduces the reliance of large numbers of more senior examiners making individual judgements about markers.

From an examination board and awarding body perspective, this approach provides more visibility and transparency to marker quality management processes which should provide additional reassurance to national education ministries and regulators that progress in a judgemental area is being made. As a result, confidence in ensuring that candidates get the results that they deserve should be increased.

Further work will be carried out to generalise the underlying data to include more subjects and, it is anticipated, improve the outputs. In fact, as more marking using this approach takes place, the more the tool can be refined and assured. In addition, the tool itself will be developed to a level that can be used in an operational examination context and potentially be integrated into the marking software itself.

Summary

The development of the approach to dynamically checking the quality of marking has progressed since its introduction into the DRS electronic marking system.

By careful application of research findings, it has enabled a practical and useful approach to decision-making for users of electronic marking to be developed. It has taken theoretical modelling and begun to place it in the examination providers' hands.

There are significant benefits of this approach, which include:

- an empirically-based, consistent application of double-marking rules that enable poor marking to be detected with some confidence and within known parameters;
- added confidence for the regulator that well-thought through approaches to improving the quality of marking are being put in place;
- added confidence for candidates that the results they receive are based upon marks subject to a quality control system that has a sound basis and can be applied clearly and consistently.

DRS remains committed not only to bringing operational system and process benefits of electronic marking to high-stakes examination users, but also to ensure that the management of the quality of marking is constantly reviewed and the technology leveraged in every way possible to improve it.

Acknowledgement

The research analysis and framework construction described in Sections 6 and 7 of this paper was carried out for DRS by Tom Benton of the National Foundation for Educational Research (NFER), Slough, UK.

For further information

For further information, please contact Graham Hudson or Sid Spalding during the AEA-E Conference in Belfast, or at DRS Data Services Limited, 1 Danbury Court, Linford Wood, Milton Keynes, MK14 6LR, UK, Tel: +44 (0)1908 666088.

Annex 1 – Terminology Used in the Double-Marking Guidelines Tool

Double marking – Having a piece of marking reviewed by a peer marker. If the mark awarded by the peer reviewer is not sufficiently close (that is, within *tolerance*) it is referred to a senior marker.

Percentage double marking – The percentage of a marker's work that will be referred to a peer for double marking.

Tolerance – The acceptable level of difference between the marks awarded by two separate markers.

Penalty – If as part of the double marking process two markers provide marks for the same piece of work that are not within tolerance of each other their work is referred to a senior marker. The senior marker will then remark the work and if any of the original marks awarded are not within tolerance of the mark awarded by the senior marker either of the original markers may be given a penalty as appropriate. A record is kept of the number of penalties that have been accrued by each marker.

Penalty Cap – The number of penalties that a marker is allowed to accumulate without being stopped from marking. If a marker exceeds the penalty cap they will be prevented from marking any further.

Poor marker – A marker with a greater than average chance of accruing a penalty on any given item. The percentage double marking system is designed so as to quickly identify these markers whilst allowing most markers to continue with their work. The size of the difference between the chances of poor markers accruing a penalty and the chances of an average marker accruing a penalty may be chosen by the awarding body but previous research has indicated that a definition of a poor marker being twice as likely to accrue a penalty is reasonable.

Marking confidence – The difference between the probability of a poor marker exceeding the penalty cap and the probability of an average marker exceeding the penalty cap. For example, if there is a 75 per cent chance that a poor marker will exceed the penalty cap and a 25 per cent chance that an average marker will exceed the penalty cap then marking confidence is defined to be 50 per cent. Ideally the penalty cap should be specifically chosen to maximise the marking confidence.

Annex 2 – Quality Control for Unconstrained Items

The system devised to monitor marking quality for unconstrained answers is known as *'percentage double marking'*. This means that one marker's marks are compared with another marker's marks according to a set sampling percentage. Its scope includes:

- comparing two marking opinions in 'real time';
- where differences in marking exceed a set tolerance automated business rules are used to invoke adjudication by a senior marker;
- standard items (similar to seeded items) can be used to judge (at any point in the process) how close to the 'set standard' the marking is;
- senior markers can intervene at any point to re-sample a marker's marking and, if appropriate, re-mark work for defined periods;
- combining the benefits of seeded marking and sampling marking through double marking.

There is an automated, but configurable, quality control framework in place – which uses a number of 'caps' (or limits) to manage marking quality. For a marker who 'marks ahead' of the rest, the *'pioneer cap'* comes into play and the marker is temporarily suspended from marking that particular item. This ensures that no marker can progress too far without a double check on the marking. As soon as some of the marking is marked by another marker, he or she can resume (provided no other tolerance is exceeded).

As markers mark, the number of times that a marker exceeds a set tolerance when marking is compared with other markers is recorded. When the set tolerance is exceeded, the marker is temporarily suspended from marking that item. This limit is called a *'suspect cap'*. A senior marker has to adjudicate the marking and give the 'true mark' to enable the marker to resume marking.

When the marker's mark is adjudicated and if found to be outside the tolerance of the senior marker, they accrue a *'penalty'*. There is a configurable *'penalty cap'* that will suspend a marker if too many penalties are accrued. A senior marker has to adjudicate the marking and give the 'true mark' to enable the marker to resume marking.

These mechanisms, together with the use of some pre-marked standard items, now enable long-form answers to be checked in a well-defined manner, regularly and with real-time monitoring of marking standards.

Annex 3 – Quality Control Framework

A quality control framework has to have a starting point. In this case, the starting point is the question *'What is a poor marker?'.*

This is defined within the percentage double marking approach set by the e-Marker® system and will relate to the likelihood of a marker exceeding the tolerance for an individual question.

This approach has been chosen as exceeding the *'penalty cap'* is a clear status point that the markers' judgements have been examined by a senior marker and found wanting. It can be readily measured, counted and related to specific items in both terms of content domain and mark tariff.

For the purposes of deriving a model, a ***poor marker*** is defined, therefore, as '***twice as likely*** *to be out of tolerance with a mark awarded by a senior marker as an* ***average marker*** *for any given item'.*

In order to make this definition work, there is a need to set an *'optimal penalty cap'* for the average item being considered, so that a *poor marker* is likely to exceed it but an *average marker* is not. (For the purposes of the example given below, the optimal penalty cap was identified as $1/10^{th}$ of the number of items being marked.)

In order for this framework to be of practical use, the *quantity of marking that has to be carried out for an average item to enable poor marking to be identified* has to be established.

In this case, the number of items that *optimises the difference in the probabilities of the* ***poor marker*** *and the* ***average marker*** *exceeding the penalty cap* has been chosen as the measure.

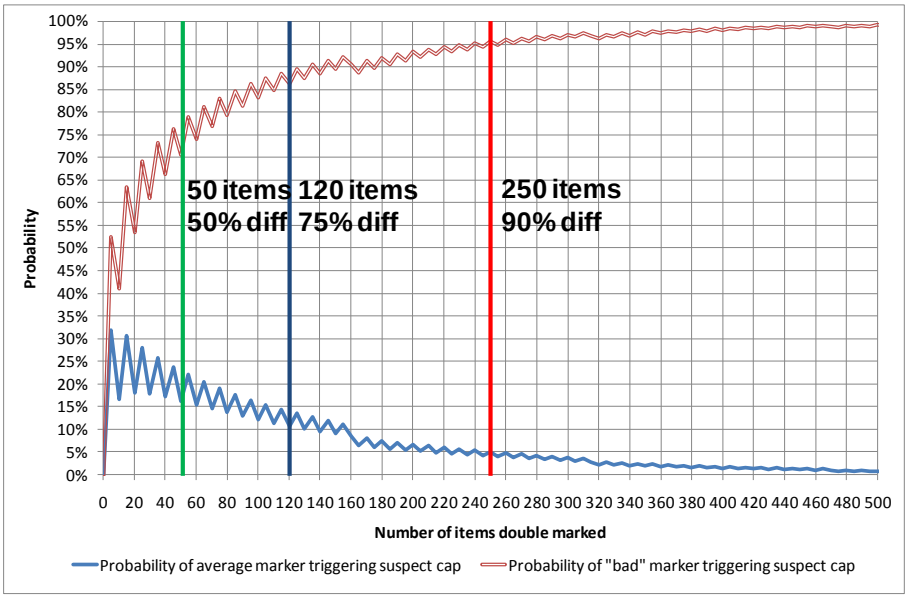
Table 2 shows how this works out for a 10 mark item within a particular examination. The upper (red) line shows the probability of a poor marker exceeding the penalty cap – an increasing trend. The lower (blue) line shows the probability of an average marker exceeding the penalty cap – a decreasing trend.

The optimal difference occurs at *250 items* in this instance – where going beyond a difference of 90% in the probabilities brings little gain (shown by the red vertical line).

However, once this plot has been drawn, other lines can be reviewed that reflect other values of differences in probability and the number of items marked where detection of poor marking will occur.

So, in this example, only *120 items* would have to be marked to determine, with a *75% probability difference*, that a marker was poor (blue vertical line). Or, if a reduced difference in probability could be accepted, then only *50 items* would have to be marked to determine, with a *50% probability difference*, that a marker was poor (green vertical line).

Table 2 Difference in probability of a poor marker exceeding the optimum penalty cap and an average marker exceeding the penalty cap for an average item



It is from this point that a framework for determining how many items require marking before poor marking can be derived. This is based upon:

- the mark tariff for the item and the consequential optimal penalty cap that is determined;
- the degree of *marking risk* that an organisation is prepared to accept – ie, what probability difference for detecting poor marking is appropriate for the type of examination, number of candidates and use of the outcomes;
- the tolerance allowed between the mark given by the marker and the mark given by the adjudicating marker (which will change as the mark tariff increases).

Given those factors, a framework can be drawn up, as shown in **Table 2**. This shows potential different 'risk probabilities' and the number of items to mark before a poor marker should be identified. The values discussed above are shown against a 10-mark item.

Table 2 Tabulating the framework

	A poor marker – 2 times as likely to exceed the penalty cap		
	Number to double mark for difference in probability of exceeding the penalty cap greater than...		
Maximum score	**90%**	**75%**	**50%**
1	1000	490	180
2	180	90	30
:	:	:	:
9	280	140	50
10	250	120	50
:	:	:	:
15	270	130	50
16	240	120	40

This means that for an allocation of 500 average 10-mark items:

- 50% would have to be double-marked for a 90% probability difference for identifying the poor marker;
- 25% would have to be double-marked for a 75% probability difference for identifying the poor marker;
- 10% would have to be double-marked for a 50% probability difference for identifying the poor marker.

The results shown in **Table 3** will vary depending upon the definition of a ***poor marker***. For example, if a poor marker were 3 or 4 times as likely to exceed the *penalty cap* then the proportion of double marking required to identify them would be smaller.