Centre for Education Research and Policy

A FURTHER TRIAL OF A BOOKMARK APPROACH TO GRADING OBJECTIVE TESTS

Diana Fowles

SUMMARY

A second trial of the Bookmark method has been conducted, in the context of a GCSE Science module, on the basis of a one parameter model analysis of the objective test data. Both trials have shown variation in outcomes from different judges which might be explained by a lack of a shared understanding of what "67% probability of success" means at each grade, something that might be improved with practice and training. For each grade the boundary marks selected in the awards on statistical grounds were closer to matching 67% than were the judges' Bookmark selections.

The information on item and test performance from the Rasch analyses might be more readily and effectively put to use than in the production for the Bookmark method of the ordered booklet and the judges' selection of the Bookmark item. A variation on the Bookmark task is suggested in which the "67%" item is identified for every possible boundary mark (this is the item on which candidates with that total mark have closest to 67% probability of success according to the model) and offered to the judges (in the form of a table) for their selection of the best one to represent performance at the particular judgemental grade boundary.

1. INTRODUCTION

1.1 An earlier paper received by the Research Committee in June 2004 (min.9) described the Bookmark method and a trial undertaken in the context of GCSE General Studies (Fowles, 2004). The main findings were that there was such variation in the Bookmark position judgements between judges in the first stage of the procedure that it was questionable whether they could legitimately have proceeded to a second stage, of discussion, reconsideration and negotiation of agreed decisions. A second stage may also feature in the Angoff method (Angoff, 1971) and concerns associated with the Angoff method can also surface with the Bookmark method. There are parallels in that the task required of judges in both methods is facilitated by experience, it makes similar though lesser demands on them in terms of appreciating the probabilities of success of different candidate groups on individual items, and might benefit from appropriate feedback to inform the respective second stages. The reliance on a single Bookmark judgement per grade from each participant means that the judgements made must be sound and supported by appropriate training on tests which have been screened as suitable for the method.

www.cerp.org.uk

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales
(company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX

1.2 In a previous discussion in response to a paper on an IRT enhancement to the Angoff method (Fearnley (2003)), the Research Committee agreed that IRT methodology, and hence the analysis underlying the Bookmark method, should use only the Rasch one parameter model rather than a two parameter IRT model. This was on the grounds that candidates' objective test scores are based on simple aggregates of item scores, not on the IRT-modelled candidate ability parameters. The two measures (aggregate test scores and candidate ability) are directly related, with a monotonic relationship, under the one parameter model but can be different under two and three parameter models according to individual candidates' particular patterns of correct and incorrect responses.

1.3 A further trial has been carried out using a GCSE Science module and a Rasch one parameter analysis of performance on individual items. The current GCSE modular specifications in Science, Biology, Physics and Chemistry provide three assessment occasions each year for up to 12 objective test modules at both Foundation and Higher level, and it is intended that a wide range of modules will continue into the new specifications (first awards 2007). Each module test requires extensive analyses on each occasion to produce statistically determined boundary marks (Meyer, 2004), to which the Bookmark method might contribute, if not as a time-saving alternative then, at least, giving a judgmental input. Although informing the awards in other contexts e.g. GCE Chemistry and GCE General Studies, AQA has not employed the Angoff procedure in GCSE Science, partly because of the large number of items on which examiner judgment would be sought. The simplicity of a single Bookmark decision for each test therefore commends its use in this context.

2. THE GCSE SCIENCE MODULE TEST USED IN THE TRIAL

The trial was based on Foundation and Higher tier candidates' responses to one of the General Science Double Award Modular module tests in the November 2004 series. The test in each tier has ten questions, each made up of either two or four items, giving a total of 36 items with one mark per item. Five of the questions (comprising 18 items, 18 marks) appear in both the Foundation and Higher tier tests. The Module 5 test on Metals was selected for the trial and was made up of items of three types:

(a) Four-option multiple choice items, each based on a common stem or theme which links the items into a single question, for example:

. *"This question is about some reactions of iron and aluminium"*
followed by four multiple choice items.

Comment: the four items are not entirely independent of the other three items which make up the question because of the common stem or theme. In one case for example the items all refer to a diagram of a blast furnace.

(b) Matching four words from a list to four positions, either:

cells in a table;
parts of a diagram; or
spaces in equations

An example is given in Appendix A.

Comment: the four items that make up the question are not independent. In particular, any candidate placing three of the four words in the correct position should, by elimination, correctly identify the position of the fourth and gain all four marks.

(c) Identifying two correct statements from five. An example is also given in Appendix A.

Comment: this is a variation on the true/false type of objective question (where the probability of choosing a correct statement by guessing is 40% rather than 50%).

The frequency of each question type is given in Table 1 with the associated numbers of items. The relative proportions of item marks from each type of question in the common items mirror the proportions in the Higher tier test while the Foundation tier test contains most of the items from the matching words category.

In this trial the 18 items that appear only in the Foundation Tier have been re-labelled as F1 - F18, those only in the Higher Tier as H1 - H18, and those common to both tiers as C1 - C18, for simplicity.

Table 1 Question types in the Module 5 test

	Foundation Tier only (F1 – F18)		Higher Tier only (H1 – H18)		Common to both tiers (C1 – C18)		Total	
Question type	No. of qns	No. of items	No. of qns	No. of items	No. of qns	No. of items	No. of qns	No. of items
1. multiple choice	0	0	3	12	3	12	6	24
2. matching words	4	16	1	4	1	4	6	24
3. identifying statements	1	2	1	2	1	2	3	6
totals	5	18	5	18	5	18	15	54

3. THE METHOD

3.1 The trial applied the Rasch one parameter model to sample test item data for about 5000 Higher tier and 5000 Foundation tier candidates[1]. Table B1 in the Appendix gives the conventional facility values and the Rasch item difficulties for the 18 tier-specific items in each test: Table B2 gives the same information for the 18 items common to the two tier tests.

3.2 Applying the method to objective test items means that the technically unsatisfactory issue which was met by Bradshaw and Schagen (2003), of how to deal with short answer items awarded more than one mark, is again avoided. However, as noted above, the Science items in this module are not ideal as they do not meet the condition

[1] The samples of 5000 were random samples drawn from the complete Module 5 entry. As noted in the earlier report (RC/259), the candidate data for the large entry GCSE Science module tests are initially held in a few hundred batches of scanned OMR forms, each batch containing mixtures of different centres and tests. Programming development would be needed to extract data for the individual tests in a useable form on a regular basis; special arrangements were made for this trial.

of independence, particularly the matching words item type which dominates the Foundation tier test (Table 1). This will be further discussed later (5.1).

3.3 Two other issues which can present technical difficulties are, firstly, omitted questions and secondly, candidates failing to reach the end of the test. Both issues were avoided here because less than 1% of the responses on any item fell into either category. It is very likely that these two issues would be avoided in *any* GCE or GCSE objective test, because the preparation of candidates for such tests will tend to ensure that they know it is in their interests to respond to every question, even if their answers are only randomly selected.

3.4 In the GCSE context, awarders are more familiar with the 'just in' rather than the 'just out' candidate, i.e. the candidate who has gained just sufficient marks to be awarded the grade under consideration. The instructions given to the judges, as in the earlier trial, were to think of a borderline GCSE candidate and to identify the last item in the ordered booklet that he or she would be likely to answer correctly, with correctness defined as having at least a 2 in 3 chance of success. (The Angoff task involves a similar requirement to think of the 'just in' candidate (and estimate how many of a group of 100 would succeed).) This item is recorded as the Bookmark, and used in conjunction with the mathematical modelling of the items to determine where the boundary mark should be set.

3.5 A single booklet was produced, containing both Foundation and Higher tier test items, with the items appearing in a single order of difficulty determined by the parameter values, in particular those of the linking or common items. Two problems were met. The first was that there were some items with virtually identical parameters. These could not readily be identified as ties in the booklet, but this was not felt to be a hindrance to the judges, nor to the outcomes, since tied items selected as the Bookmark would give the same boundary mark. The second difficulty was that four of the 18 common items had markedly different parameter values in the two tests, showing that Foundation and Higher candidates responded differently to them. These were omitted from the ordered booklet because they did not allow a single item order to emerge for the ordered booklet. They were brought back into the proceedings at the later stage of identifying the separate test boundary marks, through the separate modelling of the responses on each tier.[2]

3.6 The judges were asked to follow the same order of decision making as would be followed in an awarding meeting. Thus, when identifying the points in the booklet that they considered marked the transition for each of the judgemental grades, they started with Grade C followed by Grades A and F. The judges were invited to comment on the task that they had completed.

3.7 The Bookmark positions provided by each judge in each version were then translated into boundary marks to compare with those established following the statistically-based

[2] An alternative approach to the parameter estimation would be to include all items and all candidates in a single Rasch analysis of the 54 items in the two tests combined, which would place all the items on a single item difficulty scale. (The item characteristic curves for the two tier groups separately would show differential item functionning for at least the four items referred to above). The tests are treated separately here as they are in the examination, other than to be brought together via the common items in the production of the single booklet.

procedure. This Bookmark process makes use of the 67% probability of success associated with the judgement and the formula in Appendix C to identify the ability of the borderline candidate and hence the candidate's likely total score, which stands as the grade boundary mark.

4. BOOKMARK RESULTS

4.1 Eight judges completed the task, and provided the following comments.

- *Grade F is the most difficult boundary (to judge) …many candidates at this level simply guess answers … they tend to be patchy in their knowledge …they may answer some of the early questions incorrectly.*
- *Even at Grade C candidates' knowledge can be patchy and their responses can depend on whether the questions are on sections of the syllabus they have revised thoroughly*
- *Examination papers with two tiers always show Foundation tier candidates correctly answering higher level questions at the expense of the easier, and vice versa for the Higher tier candidates.*
- *The order of the questions (in the Bookmark booklet) is not always as you would expect.*
- *We had to work on the assumption that the items are 'stand alone'*
- *Like any system that attempts to be hierarchical, there will be significant differences of opinion and I felt that some of the later questions involved fairly low level recall, whereas some of the earlier ones involved more processing. For instance, the graph question needs processing at a higher level than recall that sodium floats on water . Even within one question with four items to match, they are not in increasing order of difficulty in this booklet.*
- *It is difficult to conceptualise the borderline candidate's probability of success of 67%.*

4.2 Table 2 brings the Bookmark positions selected by the judges together and gives the translation of each into a test boundary mark. 'Range(1)' includes all judgments while 'range(2)' discards the lowest and highest judgments, borrowing from the Angoff procedure where removing the extreme values is integral to the method. Even without the extreme values, the judges differ by up to six marks in the location of the boundary marks. The 'average' row gives the average boundary marks over the eight judges, which are the same whether or not the extreme values are discarded. They differ from the actual, statistically driven values from the November 2004 awards by one or two marks, other than for the grade A boundary where the judges' Bookmark outcome is four marks less demanding (out of 54) than the statistically derived boundary.

4.3 Also given for information in Table 2, in the shaded cells, are Bookmark positions in brackets: those in the 'average' row are the positions that yield the judges' average boundary marks on translation, while those in the 'actual' row are those that translate into the actual boundary marks. For Grade C two 'actual' Bookmark positions are needed to cater for the actual boundary marks on each tier. In other words, the data analysis has given Rasch item difficulty values which are such that a single Bookmark judgement cannot simultaneously translate into the actual boundary marks for this grade on the two tiers. The implications of this aspect of the analysis are discussed further in 5.2.2.

4.4 The variety in outcomes in Table 2 suggests that Bookmark method judgements could not be taken for use in boundary setting without further training and practice. The method hinges on the notion that judges can identify an item on which a borderline candidate would have a 67% probability of success, as opposed to say 50% or 80%. Two of the judges specifically commented on their difficulty with this requirement, reporting that they did not feel comfortable that they had met it adequately.

Table 2 Bookmark positions and corresponding test boundary marks

	Grade C			Grade A		Grade F	
	B’mark position	F’ndation boundary	Higher boundary	B’mark position	Higher boundary	B’mark position	F’ndation boundary
Judge 1	26	27	20	37	25	10	15
Judge 2	33	30	23	44	31	17	18
Judge 3	27	27	20	48	33	7	11
Judge 4	25	26	19	39	26	5	11
Judge 5	33	30	23	43	31	10	15
Judge 6	31	29	22	39	26	9	12
Judge 7	33	30	23	41	30	9	11
Judge 8	35	30	24	43	31	12	16
range(1)	26 - 35	26 - 30	19 - 24	37 - 48	25 - 33	5 - 17	11 - 18
range(2)	27 - 33	27 - 29	20 - 23	39 - 43	26 - 31	7 – 12	12 – 16
average	(31)	29	22	(40)	29	(9)	14
actual	(29)(35)	28	24	(48)	33	(10)	15

4.5 The ‘67%’ rule is an area where training and feedback might be able to influence judgments beneficially, but first it is useful to gauge how far each judgment was from meeting the 67% requirement. This can be achieved by looking at the trial data in reverse. Working back from the actual test boundary mark, and for the time being at least accepting it as the ‘true’ boundary mark, can help answer the question ‘what probability of success for a borderline candidate is identified in each item selection?’. The results are given in Table 3. It shows much variety, both between judges and between grades. It is striking that the further the average value is from 67%, the greater is the difference between the actual boundary marks and the average Bookmark boundary marks in Table 2. Where the average is below 67% the actual boundary mark is lower, while where it is above 67% the actual boundary mark is higher.

4.6 Grade A has the highest average value for the borderline candidate’s probability of success (85%) in Table 3 and the largest difference in Table 2, where the actual boundary mark (33) exceeds the Bookmark boundary mark (29) by four marks. For this grade the entries for all bar one judge exceed 80%. This means that instead of identifying a Bookmark item where the borderline candidate has a 2 in 3 chance of success, the judges have chosen an item with a much greater likelihood of success, of 85% on average. In their Bookmark positions they have underestimated how far into the ordered booklet the “67%” grade A candidate will be successful, which has translated into a lower, less demanding boundary mark than the one determined statistically. (Note however that the statistically determined mark might equally be described as having overestimated how the borderline grade A candidate would perform.)

Similarly (and with the same caveat that the statistically determined mark overestimated the borderline candidate's performance), the lowest average value for the borderline candidate's probability of success in Table 3 (61%) means that the Bookmark selections there have overestimated how far into the ordered booklet the "67%" Foundation tier grade C candidate will be successful, resulting in a higher boundary mark.

4.7 The probabilities of success associated with the actual boundary marks are also given in Table 3, and are much closer to 67% than the judges' average value in each case.

Table 3 Probabilities of success (p%) for borderline candidates identified from the judges' Bookmark positions

	Grade C		**Grade A**	**Grade F**
	F'ndation p%	Higher p%	Higher p%	F'ndation p%
Judge 1	71%	82%	96%	66%
Judge 2	51%	70%	82%	55%
Judge 3	70%	82%	64%	80%
Judge 4	74%	84%	94%	81%
Judge 5	55%	70%	82%	66%
Judge 6	64%	77%	94%	78%
Judge 7	51%	70%	87%	78%
Judge 8	54%	69%	82%	64%
average	**61%**	**76%**	**85%**	**71%**
actual	**67%**	**69%**	**64%**	**66%**

5. DISCUSSION

5.1 Estimating the probability of success

5.1.1 As already noted, and in common with the Angoff method, the Bookmark method hinges on the notion that judges can estimate a borderline candidate's probability of success on any given item. The Bookmark task is to identify the particular item where that probability is at a pre-defined level, usually 67%. In this trial the science judges identified items with probabilities of success reported for the actual borderline candidates in Table 3 (the actual boundaries being the statistically derived boundaries of course) that ranged from 61% to 85% on average. Identification of success probabilities can be expected to improve as the result of (a) experience, (b) training to give a greater appreciation of success rates, and (c) feedback from item analyses. As noted, judges indicated some uncertainty with estimating probabilities of success, as has been observed also in respect of the range of values attributed in the Angoff procedure.

Even with straightforward items the task facing the judges is a difficult one. With the Science module tests the estimation is made more complicated by the nature of the items (examples in Appendix A), with inter-related items grouped under a single question. There were some surprises in the order of difficulty of items in the ordered booklet and this inter-relationship may have been partly responsible. The analysis underpinning the method assumes item independence, and is weakened when the assumption is not met.

5.1.2 Grade boundary setting each year is required to carry forward the standards of the previous year(s), and is supported by inspection of archive scripts. Bramley (2005), writing in the context of National Curriculum testing, notes that judges are given archive scripts to help fix the correct standards in their minds; they are then asked to make a judgement about how pupils at the relevant level 'would' perform rather than how they 'should' perform, as part of a standard *maintaining* process rather than a standard *setting* process. He sees potential for confusion as to the nature of the judgmental task (standard setting or standard maintenance?). In the case of GCSE objective test papers, candidates' OTQ answer sheets are largely unsuitable to fulfil an archive function. The judges could, however, be supported by information at item level from previous tests, indicating the items that most closely met 67% success rates for candidates on the previously established boundary marks. In this way the Bookmark method could be more established as a judgmental standard maintaining method.

5.2 Translation of a Bookmark position into a boundary mark

5.2.1 The Bookmark positions are converted into boundary marks as described in Appendix C. While both the current and earlier trials of the Bookmark method have shown that the judges can differ quite markedly in their Bookmark selections, choosing items as far as eleven items apart in the present trial, the differences may prove much less striking when the selections are translated into boundary marks. This is because items can have tied positions in the ordered booklet, or item parameter values that are very close. This means that the different boundary test mark suggested by items quite far apart in the booklet may not, in practice, be so different. There is an example in the second and third columns of Table 2, where the Bookmark choices for Grade C of 33 and 35 give identical Foundation tier boundary marks of 30.

5.2.2 Table 2 revealed that the Foundation and Higher tier boundary marks (statistically determined) actually set for Grade C could not both be derived from a single Bookmark (judgmental) position. The ordered booklet was drawn up from the two tests on a single item difficulty scale on the basis of the item difficulty values of the linking or common items on the two tests. This use of common items data contrasts with the procedure for deriving the actual boundary marks, which is not informed by separate statistical data for the common and the non-common halves of the two tests; AQA awards meetings are provided with statistical data at component level but not routinely at sub-component or item level. As outlined by Meyer (2004), the statistical recommendations for the module boundary marks are usually determined directly as the statistically equivalent marks to those of the reference year, that is, giving the same cumulative percentages of candidates as were obtained in the series 12 months previously, provided that the composition and size of the entry is stable. (If it is not stable, subsets of the entry, of centres with stable entries, are used instead of the whole entry.) The implication here is that the IRT analysis of the common item data could also be used to inform the statistically recommended boundary marks.

5.3 An alternative procedure for the Bookmark method

5.3.1 Although the Bookmark booklet gives the judges more than the Angoff method by way of item level feedback (because the items are ordered in terms of difficulty), it still requires them to make estimates of likely candidate performance without knowledge of how candidates have actually performed, initially at least. Fearnley (2003) suggested, in the context of the Angoff method, how item level feedback could be used to inform examiners' judgements. Noting that facility values relate to average performance of candidates rather than the performance of only those candidates near the grade boundaries, he pointed out that IRT data could give awarders information about performances on items for candidates at each total score on the question paper, that is, at all possible grade boundary marks. As a change to the Angoff method the extra information would be used as additional input and would mean redefining the task to be choosing what is judged to be the best set of item success rates, and accepting the associated total mark as the boundary mark.

5.3.2 Similarly, in the Bookmark method, it would be feasible from IRT data to give as feedback the success probabilities for all the items for the particular group of candidates with the total test mark (the candidate ability measure) for each item that might be selected as the Bookmark. The extra information would be used as additional input and would also mean redefining the task. Instead of working through the ordered booklet and selecting a single Bookmark item, the task would be to review the different sets of item success rates which the IRT analysis associates with each total test mark. The items would be listed in descending order of difficulty, and the item most closely associated with the 67% success criterion identified in each score group. The judge's task would be to review the item on the module test paper, and judge it as being more demanding, less demanding or about right as an item that candidates at the borderline in question would have a 67% chance of answering correctly. The total test score of the group carrying the 'About right' item would be identified as the boundary mark. Table 4 gives an example of how the item data might be presented in this revised method. The probabilities of success are derived from the one parameter model and will decrease from the item identified as easiest to the item identified as the most difficult overall. The task of the judge is to identify a likely boundary mark and decide if the item identified with a 67% (or close) probability of success for that total score group is the correct item and, if not, to review the alternative items in adjacent column(s). If Item H11, for example, was judged to be the correct item for Grade A, the associated boundary would be 30. Note that the judge needs the same strong grasp of what "67% probability of success" means for this task as for identifying the Bookmark item in the ordered booklet.

Table 4 Example extracts of probabilities of success for candidates by total score group in 36 item Higher tier test

	total score group									
item	**32**	**31**	**30**		**27**	**26**	**25**		**21**	**20**	**19**
H5	97%	96%	91%		86%	85%	84%		73%	69%	**65%**
H8	95%	92%	86%		84%	83%	77%		71%	**67%**	64%
H3	93%	91%	84%		80%	78%	76%		**66%**	62%	60%
H7	89%	86%	81%		79%	75%	72%		65%	61%	59%
C3	86%	84%	75%		76%	73%	69%		62%	58%	56%
H1	83%	81%	73%		73%	72%	**67%**		60%	57%	53%
H4	79%	75%	72%		72%	**67%**	62%		57%	55%	50%
C6	76%	73%	69%		**68%**	65%	61%		49%	43%	41%
H11	73%	72%	**66%**		65%	62%	58%		46%	40%	37%
H6	72%	70%	62%		62%	59%	56%		44%	39%	36%

5.3.3 For the alternative method proposed above there would actually be no need to physically produce an ordered booklet, and the term 'Bookmarking' would no longer be appropriate. Data of the kind in Table 4 would be the 'stimulus material' needed by the judges in conjunction with the question papers: their judgements on the particular items identified in the analysis would be the guide to the boundary mark. There would however be similar training requirements to those identified for the Bookmark method, to help interpret data of the kind in Table 4. There would also be the same need for judges to discuss and, so far as possible, reconcile their views.

5.3.4 As noted earlier (para 5.1.2), archive items identified from analyses of the equivalent tests in earlier year(s) could be supplied to the Bookmark judges to help by both illustrating 67% probability of success for borderline candidates and providing a link to the standards set on previous occasions.

5.3.5 To take either the Bookmark or this alternative approach forward in the context of the GCSE Science module tests, the item and test results for other modules should be examined, to establish whether or not the '67% rule' fits the majority of the decisions taken (as it does for Module 5, as evidenced by the final row of Table 3) or whether indeed some other value might be preferred. (The literature does not indicate any irrevocable theoretical reason for this value, although empirically it seems a good choice.) Also to be considered further would be the use of the common items to link the Higher and Foundation tests and here the analyses of all the modules would be of interest (5.2.2) (In this exercise with Module 5, the one parameter model analysis suggests some divergence of the two Grade C boundaries.)

5.3.6 On the assumption that a judgmental 'Bookmark' element can be added along these lines, the awards meeting would need the following inputs:

- last year's probabilities of success for candidates at the boundary marks (as in the last row in Table 3)
- the probabilities of success associated with this year's SRBs
- this year's 'Bookmark' boundary mark recommendations and their associated probabilities of success

CONCLUSIONS

Two trials of the Bookmark method have shown variation in outcomes from different judges which might be reduced with a degree of practice and training, especially on a shared understanding of what "67% probability of success" means at each grade boundary, illustrated by items from earlier tests. The Bookmark selections for grade A candidates, for example, were associated with a much higher probability of success than 67% (85% on average, Table 3), and this was reflected in a total test boundary mark that was lower than the actual boundary by four marks on average.

The differences in the judges' Bookmark selections can be reconciled by simple averaging of the resulting boundary marks, as in Table 2. Alternatively they might be helped, with training, to improve on the judgements first; as with the Angoff judgements; they could be made more secure by giving feedback on how candidates have responded to the items, as interpreted by an IRT (Rasch) analysis.

The information on item and test performance from the Rasch analyses might however be more readily put to use in a variation on the Bookmark task, in which the "67%" items would be identified for every possible boundary mark and offered to the judges (in the form of a table like Table 4) for their choice of the best ones to represent performance at the judgemental grade boundaries.

Dee Fowles
March 2005, revised May 2005

REFERENCES

Angoff, W.M. (1971) Scales, norms, and equivalent scores. In Thorndike, R.L. (ed.) *Educational measurement* (2nd edition, pp. 508-600). Washington, DC: American Council of Education

Bradshaw J. and Schagen, I. (2003) *Use of the Bookmark for Setting Standards in Reading Tests.* Paper presented at the IAEA Conference, Manchester

Bramley, T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, v6, n2, p202-223

Fearnley, A.J. (2003) *An investigation of the possible application of item response theory to provide feedback of information to awarders in the use of Angoff's method of standard setting in AQA OTQ components*. Research Committee Paper RC/240

Fowles, D.E. (2003) *Standard setting: a review of some recent approaches to setting boundary marks on the basis of examiner judgement.* Research Committee Paper RC/207

Fowles, D.E. (2004*) A trial of a bookmark approach to grading and comparisons with the Angoff method.* Research Committee Paper RC/259

Meyer, L. (2004) *November 2004 GCSE Modular Science module test awarding meeting. Technical report.* Internal AQA report

APPENDIX A

EXAMPLES OF ITEM TYPES IN GCSE SCIENCE MODULE 5 (METALS) TEST

1. Matching four words from a list to four cells in a table

This question is about the properties and uses of some elements.

Match words from the list with the numbers 1 – 4 in the table.

carbon
copper
magnesium
sodium

Element	What we can say about the element
1	it floats on water
2	it is a non-metal that conducts electricity
3	it is mixed with aluminium to make a stronger alloy
4	it is used to make electrical cables

(This question provides four of the 54 separate items in Appendix B)

2. Identifying two correct statements from five.

This question is about the metal copper and its compounds.

Which **two** of the following statements are correct?

copper belongs to Group 1 in the periodic table
copper oxide is a base
copper oxide is soluble in water
copper oxide reacts with sulphuric acid to produce copper chloride
many copper salts are coloured

(The two correct statements are two of the 54 separate items in Appendix B)

ITEM INFORMATION APPENDIX B

Table B1 Non-common items, in the order of presentation in the ordered booklet

	Foundation tier (N = 4955)			**Higher tier** (N = 4898)		
Order	**Tier F item**	**Facility %**	**Rasch difficulty**	**Tier H item**	**Facility %**	**Rasch difficulty**
1	F4	89.1	-1.93	H2	96.5	-2.47
2	F3	79.7	-1.12	H3	95.7	-2.25
3	F14	79.4	-1.10	H1	90.4	-1.33
4	F1	79.2	-1.08	H12	90.2	-1.31
5	F10	79.1	-1.08	H4	89.8	-1.26
6	F2	78.4	-1.03	H15	70.8	0.23
7	F6	70.3	-0.54	H7	70.7	0.23
8	F12	69.7	-0.51	H6	65.4	0.53
9	F18	69.0	-0.47	H5	58.0	0.92
10	F8	66.2	-0.32	H8	55.1	1.06
11	F7	65.6	-0.29	H9	51.3	1.25
12	F11	60.8	-0.05	H11	50.9	1.27
13	F9	54.2	0.27	H10	50.7	1.28
14	F5	53.3	0.31	H18	48.4	1.39
15	F13	51.9	0.38	H17	46.4	1.49
16	F15	46.7	0.63	H14	39.8	1.83
17	F16	46.1	0.66	H13	39.7	1.84
18	F17	19.0	2.19	H16	32.5	2.23

Table B2 Common items, in presentation order

(N=9853)

		Facility %		**Rasch difficulty**	
Order	**Item ref**	**Tier F**	**Tier H**	**Tier F**	**Tier H**
1	C8	78.8	94.2	-1.79	-0.89
2	C9	75.1	94.2	-1.54	-0.89
3	C4	75.0	89.7	-1.53	-0.35
4	C5	68.8	88.6	-1.16	-0.26
5	C6	64.7	82.5	-0.93	0.17
6	C7	54.6	86.0	-0.41	-0.06
7	C18	49.4	86.8	-0.16	-0.12
8	C16	52.7	82.5	-0.31	0.17
9	C11	53.1	73.9	-0.33	0.62
10	C2	49.2	77.4	-0.14	0.45
11	C10	50.6	71.4	-0.22	0.73
12	C14	50.0	72.1	-0.19	0.70
13	C17	42.7	79.0	0.18	0.37
14	C13	40.1	74.9	0.32	0.57
15	C1	44.1	66.4	0.12	0.94
16	C15	42.9	67.2	0.17	0.92
17	C3	40.5	63.7	0.30	1.06
18	C12	33.1	67.1	0.70	0.92

APPENDIX C

CALCULATING GRADE BOUNDARY MARKS FROM BOOKMARK ITEMS

The parameters estimated by the Rasch model are used to calculate the probabilities of candidates answering items correctly using the formula:

Probability of success = 1/(1+(exp(-1.702*(theta-b)))).

where

theta is candidate ability,
b is the Rasch item difficulty parameter (Appendix B).

1. Substituting the item difficulty of the particular item selected as the Bookmark item into the formula together with 0.67 as the probability of success gives the candidate ability associated with that item. In this way the Bookmark item is uniquely associated with a candidate ability measure.

2. Substituting the Bookmark candidate ability measure identified in 1 into the formula with the item difficulty of each other item in turn gives the probability of success associated with each item for a candidate of that measure of ability.

3. The probabilities of success calculated in 2 for all the items are summed over the separate Foundation and Higher tier tests to produce a total score on the respective test. This is then the total test score associated with the candidate identified by the Bookmark item as being at the grade boundary, and hence is the grade boundary mark.