MULTIPLE RESPONSE QUESTIONS – ALLOWING FOR CHANCE IN AUTHENTIC ASSESSMENTS

Mhairi McAlpine and Ian Hesketh

Multiple Response Questions – Allowing for Chance in Authentic Assessments

Mhairi McAlpine, Robert Clark Centre for Technological Education, St Andrews Building, University of Glasgow, 11 Eldon Street, G3 6NH
m.mcalpine@elec.gla.ac.uk

Ian Hesketh, TOIA Project Manager, University of Strathclyde, 155 George Street, Glasgow, G1 1RD. ian.hesketh@strath.ac.uk

Keywords: MRQs, chance factors, guessing, analysis, question weighting

Abstract

The advent of computer-assisted assessment has led to a revolution in the types of objective questions that students are being asked. Whilst once a fixed response question generally meant a traditional multiple-choice question, now there are an ever-increasing number of variations. These include assertion-reason, hotspot, permutational MCQs, ranking numerical, fixed text input, multiple-response questions (MRQs) and matrix questions (Whittington, 1998, Farthing 1999). Each can also be extended further with the appropriate use of emerging technologies such as the inclusion of multimedia elements and parameter randomisation. By integrating these technologies (normally by browser plug-ins) more 'authentic' assessments are possible.

As these methods have become more popular and easier to implement into computer-based tests, the limitations of traditional methods of analysis are becoming evident (see MacKenzie and O'Hare, 2002). Just as assessment itself is used in a variety of ways (Butterfield, 1996) so are the related statistics (McAlpine, 2002) with the most common use being to measure the quality of the question and to exclude 'bad questions' according to its non-conformance to a predetermined facility value range. However, as the complexity of question construction increases the use of a pure facility value loses utility, as there may be an inadvertently high chance factor that has not been taken into account. This may also adversely affect the capacity of the test to discriminate between weaker and stronger students if mark/score allocation is highly interdependent as is the case with MRQs, hence lowering test quality.

Traditional MCQs typically have a stem, one correct answer and then three, sometimes four, distracters – although alternative mark schemes have been suggested and developed (Bush, 1999, 2001). The probability of guessing the correct answer is thus one chance in every four (or five). There are a variety of ways of correcting for this factor, but it is widely acknowledged and accounted for (eg Muller, 1932; Rowley and Traub, 1977; Frary, 1988, Chevaliver, 1998) – although Burton (1999, 2001), regards this as a source of unreliability and as such should be minimised, however there are others that regard the development of intelligent guessing from a restricted range desirable, and a contributor to validity. Negative marking is not generally implemented in the UK multiple-choice assessments, however its use is acknowledged. Carneston et. al. (no date) discusses several methods of applying negative marking , including to the items individually and to the test as a whole. Although he acknowledges that its use is contentious, he suggests that the high guessing factor associated with certain multiple-choice questions can justify its use to aid overall test discrimination.

The authors have attempted to identify a range of issues relating to the use of MRQs with a view to proposing an authentic approach to their analysis. A review of 65 formative and summative tests from a range of disciplines (dating from 1997 to 2002) was undertaken to elucidate common approaches in the use of MRQs. .

By identifying the variety of approaches to scoring MRQs the issue of analysis became clearer. By looking at MRQs it is suggested that in order for the CAA community to continue innovative development that allows a greater degree of authenticity, there should also be clarity in the use of item statistics. By review of item analysis methods the community should be in a position to determine that innovation does not outstrip the available means of determining quality and fitness for purpose. The use of the term in this context is not intended as an exposition of the complexities of determining the nature of authenticity in performance assessment

Whilst the purpose of this paper is to discuss the issues relating to the analysis of questions and tests, not to argue about the relative merits of different forms of assessment, the paper does assume and support the position that whilst the test items considered have predetermined answers and are automatically evaluated by a computer thus falling within the broad realm of objective testing they are capable of evaluating complex cognitive skills. The capacity of objective test formats to assess the full range of cognitive levels and skills is covered in other texts such as Haladyna (1997) and Bull and Hesketh (2003)

Methodology

The examination research data was made available to the authors on the understanding that source and assessment content of individual examinations remained confidential. The files comprised of archived examinations dating from 1997 to 2002 from a UK university. The outcomes of these exams in terms of individual student performance were not made available. The authoring shell used has purposefully not been identified to enable more generic issues to be identified.

A range of disciplines were represented within the test files, but in line with work carried out by the CAA Centre (Bull and McKenna, 2000) there was a strong leaning towards sciences and computing disciplines. Following the removal of duplicate questions (those used in different modules within the same subject area drawn from a pool of questions) a total of 637 multiple response questions from 65 exam files were reviewed to determine how they had been constructed. Information on the number of options, keys, maximum number of selections allowed, scoring variation and any use of negative marking was collected for analysis. Overall test data was collected to allow determination of proportion of marks allocated to MRQs within the test (further work is planned on the determination of overall chance factors within tests comprising of a broad range of advanced objective test items).

Before moving to more theoretical considerations over the performance of these question types some related descriptive statistics are provided in Figure 1.

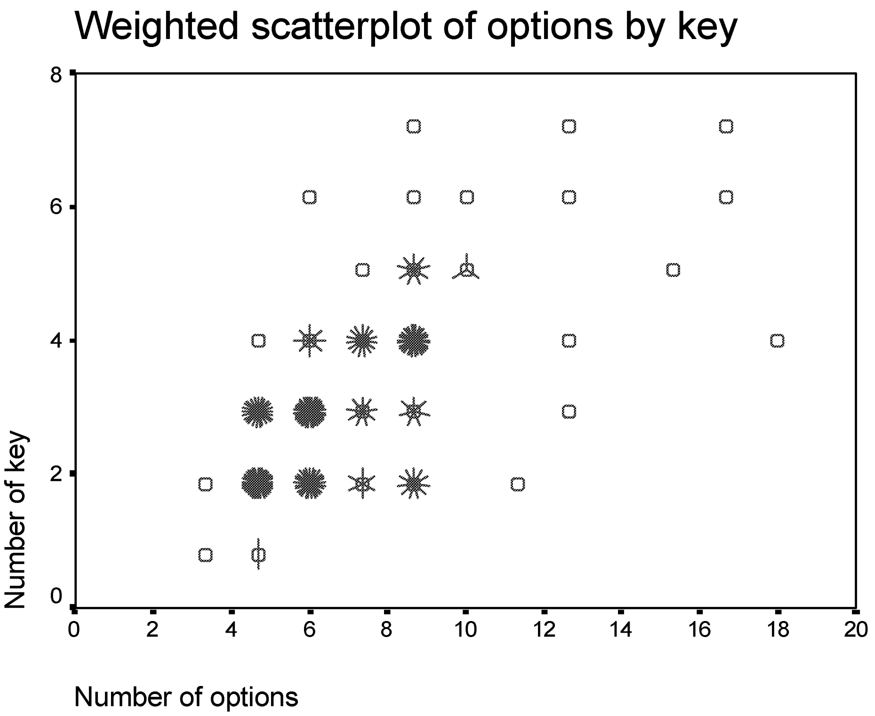
Figure 1: Descriptive Statistics

Total Number of Tests	65
Total Number of Questions	3094
Total Number MRQs	637
MRQs represent n% of tot questions presented	20.59
Overall Average of MRQ marks per test	31.62
Average Chance Factor	48%
Number of MRQs where chance factor 25% or below	26 (4.24%)
Number of MRQs where chance factor >26% -<50%	151 (23.70%)
Number of MRQs where chance factor 50% or higher	459 (72.06%)

From these figures a number of analyses were carried out to further distinguish statistical factors affecting the performance of the questions and tests.

The first analysis to be carried out was to run a weighted scatterplot of keys against options. The results of this can be seen in Figure 2 below.

Figure 2: Weighted Scatterplot of options by key



The full combination of keys and responses are also provided in Figure 3. The table shows that a total of 37 different combinations were used. Some of the more unusual combinations are on the periphery such as students in one instance being allowed to select 6 from 6 or asked to select 2 from 11.

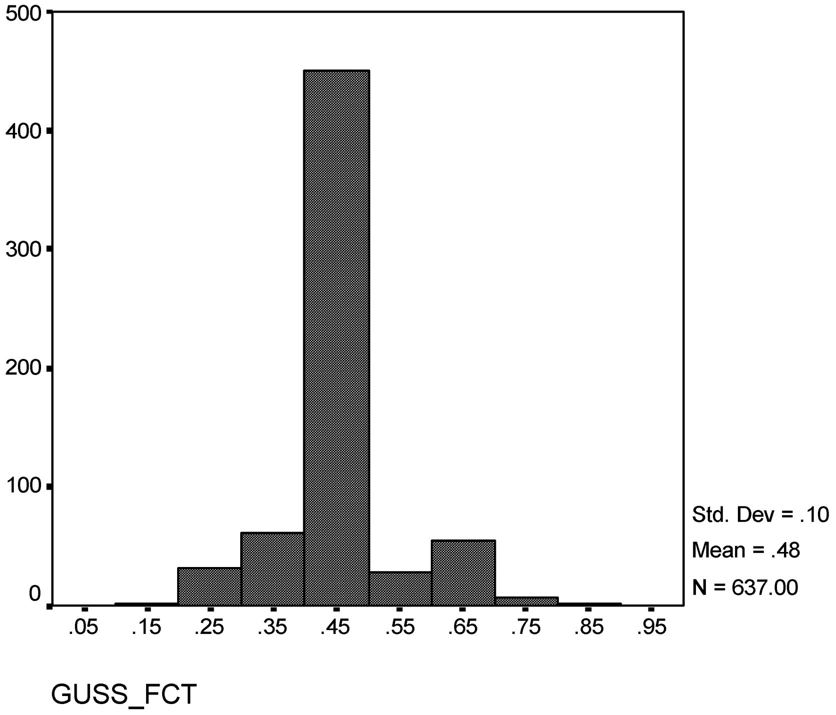
Figure 3: Table showing full response combinations used

Options\Keys	1	2	3	4	5	6	7
1							
2							
3	1	1					
4	3	56	3				
5		62	32	1			
6		42	264	12		1	
7		9	14	22	1		
8		17	9	44	8	1	
9		1	1	5	5		1
10					5	1	
11		2					
12				2			
13			2			1	1
14							
15					2		
16						2	1
17							
18				1			

From the figure 3 you can see that the most popular combination of options and keys was 3 from 6. This happened to be the default setting on the authoring shell and by far the most often used combination with almost 40% of the questions having this combination. The majority of items also clustered between 2-5 keys, and 5-9 response options perhaps suggesting that question setters are unhappy to force their students reading through a great deal of superfluous response options. The institutional guidelines for computer based tests also suggested that no more than 60 questions per hour should be set, thus speed of response would also be a factor in the question setters approach. Although there were no minimum criteria, this may have given the impression that it was desirable (or possible) to allow only one minute per question where objective test formats were used.

The data was then reviewed to identify chance factors within the questions. The results of this can be seen in Figure 4.

Figure 4: Histogram of question chance factors



The histogram in Figure 4 shows the spread of chance factors, grouped into intervals of 0.1. The figures on the X-axis indicate the average "chance factors" within each of the deciles. As can be seen clearly, the most common chance factors were in a range of 0.4 – 0.5.

Figure 5: Question Chance factor spread by decile

Chance Factor	% of questions
0.00 – 0.39	14.9
0.40	9.6
0.43	2.2
0.44	1.0
0.46	0.2
0.50	57.8
0.51 – 1.00	14.3

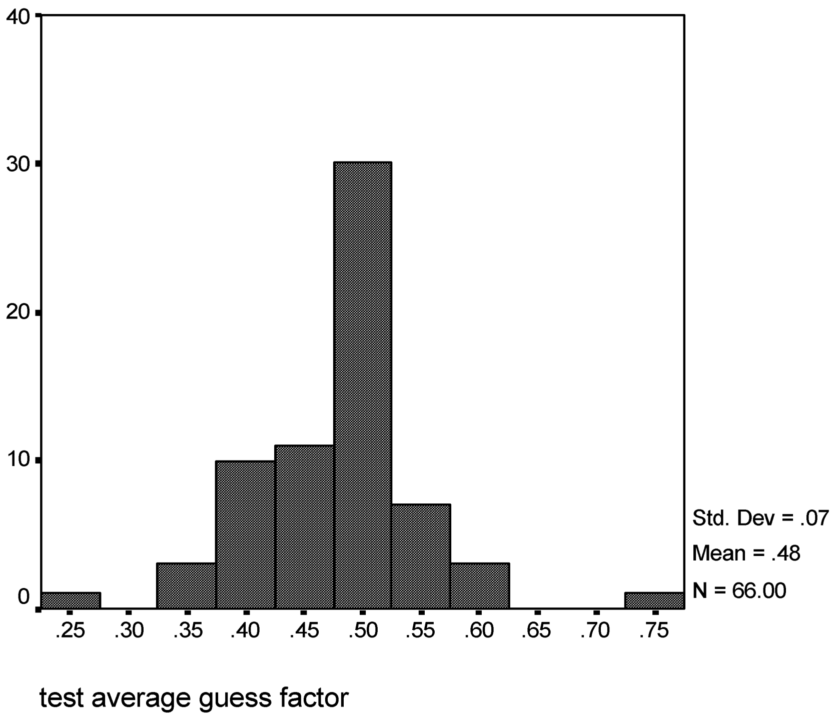
Breaking this data down further, it can be seen that this is primarily due to 2 major data points, the first (smaller) one at 0.4, and the much larger one at 0.5. Very few of the questions had a lower chance factor than a standard MCQ (keys = 1; options =4; chance factor = 0.25) and in over 50% of the questions the chance factor was equivalent to that of a true/false question. Indeed in a significant number of cases, the chance factor was even greater

In Figure 6, which looks at the tests overall, the MRQ element of the tests ranged from an average chance factor of 0.25 to 0.75, although it can be seen that these cases were outliers with the majority of the data falling between 0.34 and 0.60. Only in one test was the average chance factor comparable to test comprising of standard MCQs, and in over a quarter of cases was higher than in a true/false test.

Figure 6: Test average chance factor tabulated

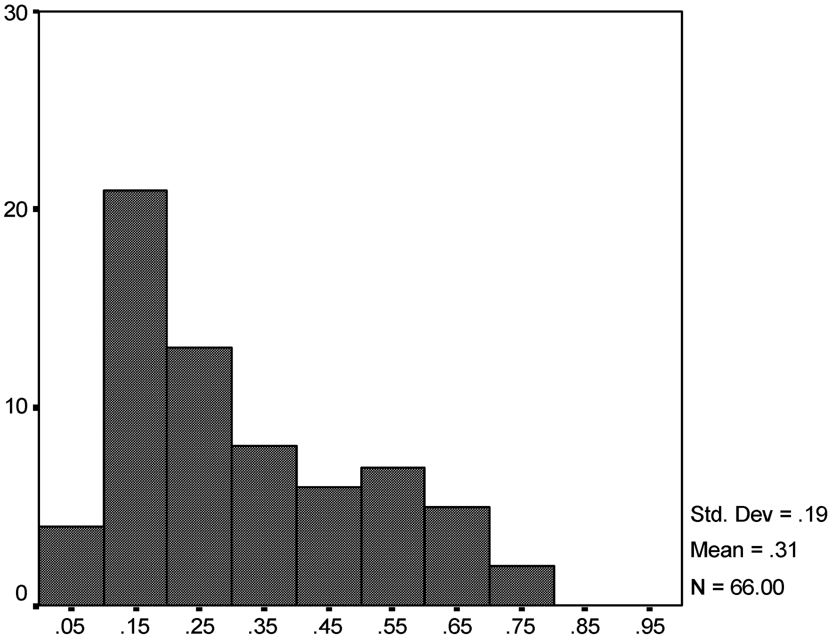
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.25	1	1.5	1.5	1.5
	.34	1	1.5	1.5	3.0
	.35	2	3.0	3.0	6.1
	.38	3	4.5	4.5	10.6
	.39	2	3.0	3.0	13.6
	.40	2	3.0	3.0	16.7
	.41	1	1.5	1.5	18.2
	.42	2	3.0	3.0	21.2
	.43	3	4.5	4.5	25.8
	.44	1	1.5	1.5	27.3
	.45	1	1.5	1.5	28.8
	.46	3	4.5	4.5	33.3
	.47	3	4.5	4.5	37.9
	.48	3	4.5	4.5	42.4
	.49	4	6.1	6.1	48.5
	.50	17	25.8	25.8	74.2
	.51	1	1.5	1.5	75.8
	.52	5	7.6	7.6	83.3
	.53	3	4.5	4.5	87.9
	.54	3	4.5	4.5	92.4
	.56	1	1.5	1.5	93.9
	.60	3	4.5	4.5	98.5
	.75	1	1.5	1.5	100.0
	Total	66	100.0	100.0	

Figure 7: Test average chance factor histogram



Whilst it was hypothesised that the chance factor of a test may be related to the proportion of MRQs in the examination, with lecturers who were aware of this issue and tried to use a lower chance factor, more reluctant to use them frequently; this was not observed (r= -0.19; p= 0.88).

Figure 8: Proportion of MRQ marks



proportion of marks which are mrqs

.

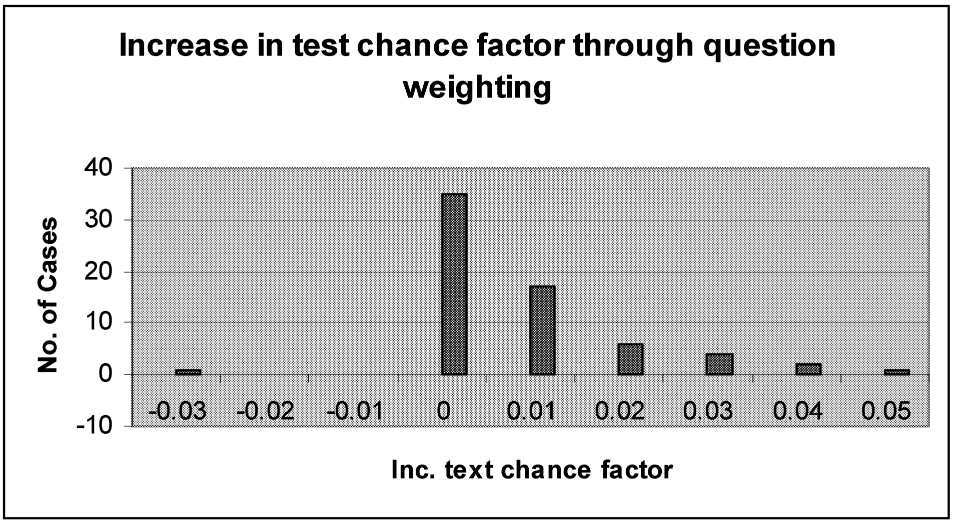
Multiple response questions naturally carry a high weighting as compared with other types of objective questions. In the vast majority of cases (>95%), one mark is given to each key and in no instances was less than one mark given per key. This leads to the question as a whole being accorded more than one mark and in some cases quite a high number of marks. This alters its relative weighting within the paper. Some lecturers will be aware of this and purposefully allocate MRQs to important parts of the syllabus to take advantage of this weighting factor, however, the high chance factor associated with many of these questions lead to an inaccurate assessment of students actual knowledge. Furthermore, where a lecturer consciously allocates a high number of marks to an important area and chooses an MRQ to assess it, from the scatterplot (Figure 2) we can see that particularly with questions with a large number of keys, there are not enough options introduced. This has the effect of inadvertently *lowering* the weighting for that question as the differentiation is so low that it will contribute comparatively little to the overall judgement of student ability.

When the question weightings within the test were taken into account, analysis would suggest that this was serving to further increase the chance factor and consequently lower the questions overall discrimination and decrease its actual weighting compared with its actual weighting compared with its intended weighting.

The question chance factors were weighted by the number of marks that that question carried within the test (to indicate how much weighting they carried in practise). A weighted chance factor was then calculated to examine what effect this had. In only one of the 65 cases did this diminish the chance factor (by 0.03). In 27 cases the question weightings did not affect the overall test chance factor – these were primarily in tests where there were a uniform number of keys and throughout the examination, but in 38 cases the test factor increased, sometimes by quite a large degree.

Figure 9 shows the increase on the test chance factor. (Note: these are rounded to 2 significant figures, producing a slight variation on the figures given above).

Figure 9: Increase in test chance factor through question weighting



Where the lecturers are unaware of this weighting issue, it is leading to less important parts of the syllabus being afforded high intended weightings but with an associated high chance factor. However, this high weighting will, in practice, be depressed by the low discrimination that a high chance factor brings. This has important assessment implications. It would in essence unfairly and disproportionately penalise a good student who unluckily chose the wrong options, by allocating a high proportion of the marks to it.

Discussion

It appears expected that as questions become more complex in form and allow students to exhibit an element of construction within their responses, that these questions are *de facto* more 'difficult'. Whilst on a pedagogical level, assuming that students interact with the questions in the way that was originally envisaged, MRQs have the capacity to act as more than a monitor of student knowledge and may encourage and support learning in a formative environment given the increased potential for feedback provision (see Steven and Hesketh, 1999), however the high chance factor suggests that within summative testing the data shows that in many instances they may be 'easier' than MCQs. So the question must be posed: are MRQs a valid summative assessment format?

As mentioned the questions within this study were subject to a constraint on the maximum number of selections possible but with no minimum. This allowed the award of partial credit within questions.

In the absence of any proven model for using MRQs and improved test construction that accounts for the combination of chance factors within the increasing variety of question types available within modern CAA packages, authors may be forced to restrict question types used to those with statistical validity – or rely upon the application of post-test guess correction formula which may be seen as inherently unfair on students as the correction would be applied universally thus penalising students who had answered 'fairly' without guessing.

Within multiple response questions if negative marking is used, its application should balance out or reduce the chance factor to 1 in 4 or below to allow comparison with single response items. (i.e. if the ratio of keys to distracters is higher than within MCQs, negatives should be applied to distracters to penalise incorrect response where partial credit has been allowed) This would not disallow the provision of extra weighting to questions with added cognitive complexity. McCabe and Barratt (2003) have suggested a formula for the computation of an MRQ's chance factor, which would allow item writers to compute the chance level of any MRQ item that they design. If the chance factors within the items used within the test can be factored prior to test it then these should be related to an overall test chance factor to allow test authors to have an accurate pre-test indication of overall chance.

By removing the option for the award of partial credit this in essence converts the MRQ to a single response question. With increased weighting given to these single responses (sic) the chance factor of MRQs can be predetermined and reduced accordingly. For analysis purposes the chance factor of a grouped and scored response (without replacement) can be determined. This is also an approach adopted by some analysts in the determination of difficulty statistics for MRQs although by ignoring the combination of responses test takers may not give an accurate indication of the true difficulty of the question. This is also an issue of question construction and improvement by using MRQs individual key facility and frequency analysis values as well as a question of chance and combinations.

The issues are further complicated when one moves away from Fixed Item Tests (where each student is presented with the same set of questions in the same order) to either randomised or adaptive tests. Here the variation in chance factors between whole presented tests could be significant unless the fundamental issues concerning the internal chance factors of advanced question types are addressed and both theoretical and practical solutions are reached.

Recommendations

By reviewing the small (but growing) body of literature focussing upon the scoring and guess correction of non-MCQ question types, it is clear that it is time for a community based approach to identifying and resolving issues of analysis.

As the IMS QTI matures and is further developed there should be further work carried out on how the outcomes of computer based questions and tests are handled. This would include the chance factors associated with the combination of response types and the implications of using constrained or forced response approaches.

The further development of mathematical or statistical approaches to chance calculation and guess correction in advanced or emerging question formats should be tackled by cross-disciplinary research into the mathematics of CAA as McCabe and Barratt (ibid) suggest.

In order to allow CAA to retain validity more effort should be on the analysis of tests that exploit the advances made in authoring complexity. This may focus upon the use of negative marking or post-test normalisation but should be cognisant of the broader issues of test construction and psychometrics.

Conclusions

The study of the examination data identified a number of important implications for the use of MRQs. The variety of uses in terms of key/option balance suggests that people are using MRQs in quite different ways whilst inadvertently increasing the likelihood of students gaining credit for random selection.

The use of 'matrix' type questions may offer a working solution to some of the issues identified. As Braswell and Kupin (1993) discovered in their use of grid formats for paper-based responses to objective tests their use 'virtually eliminated' guessing and backdoor approaches to question response. The matrix format is now a feature of many systems and if used as an alternative could prove effective although as Braswell only used a small number of multiple response questions further study would be required.

The default settings on authoring shells have influenced the implementation of MRQs by suggesting that the initial balance is appropriate thus care should be taken during the construction of questions and tests over the number of keys and options to ensure that there are sufficient distracters to make the item valid.

References

Braswell, J. and Kupin, J. (1993) '*Item Formats for Assessment in Mathematics*' in Bennett, R.E. and Ward, W.C. (eds) Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing and Portfolio Assessment, Lawrence Erlbaum, New Jersey.

Bull, J. and Hesketh, I. (2003) 'Computer Assisted Assessment and Higher Order Skills', LTSN.

Bull, J. and McKenna, C. (2000) '*Computer-assisted Assessment Centre (TLTP3) Update*' (keynote) in Cooper, H. and Clowes, S. (eds) Proceedings of the 4th International Computer Assisted Assessment Conference, Loughborough.

Burton (2001) '*Quantifying the effects of change in multiple choice and true/false tests: question selection and guessing of answers*', Assessment and Evaluation in Higher Education, 26:1

Bush, M. (1999) 'Alternative Marking Schemes for Online Multiple-Choice Tests', Paper presented at the 7th Annual Conference on the Teaching of Computing, Belfast. Available online at http://www.caacentre.ac.uk/dldocs/BUSHMARK.PDF

Bush, M. (2001) '*Multiple Choice Test That Reward Partial Knowledge*', Journal of Further and Higher Education, 25: 2.

Butterfield, S. (1995) *Educational Objectives and National Assessment*, Open University Press.

Carneston, Delpierre and Masters (no date) 'Designing and Managing multiple-choice questions', University of Cape Town, South Africa.

Chevalier (1988) '*A review of scoring algorithms for ability and aptitude tests*' Annual Meeting of the Southwest Psychological Association*,* New Orleans, April.

Farthing, D.W. & McPhee, D. (1999) '*Multiple choice for honours level students? A statistical evaluation'* In Danson, M. (ed) Proceedings of the Third Annual Computer Assisted Assessment Conference, Loughborough.

Frary (1988) '*Formula scoring of multiple choice tests (correction for guessing)*' , Instructional Topics in Educational Measurement*, No. 3* in Plake (ed) Educational Measurement (Issues and Practices) 7: 2

Haladyna, T. (1997) *Writing Test Items to Evaluate Higher Order Thinking*, Allyn Bacon, Needham Heights, MA.

Mackenzie, D. and O'Hare, D. (2002) 'Empirical Prediction of the Measurement Scale and Base Level Guess Factor for Advanced Computer-Based Assessments' in Danson, M. (ed) Proceedings of the 6th Annual International Computer-Assisted Assessment Conference, Loughborough.

McAlpine, M. (2002) 'Design Requirements of a Databank' Bluepaper No.3; Computer-Assisted Assessment Centre, Loughborough.

Muller, K. (1932) 'A method for correcting guessing in true-false tests and empirical evidence to support it', *Journal of Social Psychology,* Volume 3.

Rowley & Traub (1977) 'Formula scoring number right scoring and test taking strategy', *Journal of Educational Measurement,* 14: 1

Steven, C. and Hesketh, I. (1999) '*Increasing learner responsibility and support with the aid of adaptive formative assessment using QM designer software'* in Computer Assisted Assessment in Higher Education by Brown, S., Bull, J. and Race, P. (eds) Kogan Page, London.

Whittington, D. (1998) 'There's more to the Web than Multichoice' In Danson, M. (ed) *Proceedings of the Second Annual Computer Assisted Assessment Conference*, Loughborough.