CAA SCORING STRATEGIES FOR PARTIAL CREDIT AND CONFIDENCE LEVELS

Michael McCabe and David Barrett

CAA Scoring Strategies for Partial Credit and Confidence Levels

CAA - It's a Mean Uneducated Guesser's Scoring (MUGS) Game!

Michael McCabe and David Barrett
michael.mccabe@port.ac.uk david.barrett@port.ac.uk

Department of Mathematics, Faculty of Technology, University of Portsmouth
Buckingham Building, Lion Terrace, Portsmouth, Hants PO1 2DT

Abstract

Rigorous mathematics is preferable to computer simulations for analysing the scoring distributions obtained by guesser's in objective questions. The results obtained are more accurate, efficient and reliable. As an example, we show that scoring in objective Multiple Response Questions (MRQ) and objective Multiple Choice Questions (MCQ) is governed by the hypergeometric distribution. Hence, we consider how partial credit and confidence levels can be accounted for by adopting suitable scoring schemes for MCQ and MRQ objective questions. The Mean Uneducated Guesser's Score may be specified as zero (MUGS = 0) or some other value in advance, rather than being obtained by trial and error. The importance of mathematics in the design of CAA scoring systems is highlighted.

Mathematics in CAA

Objective testing and CAA is widely used for mathematics, but the use of mathematics in designing and analysing questions is less frequent. While most other subject areas exploit the common benefits of CAA, mathematics has a special contribution to make in the understanding of objective testing and its effective use. It is unfortunate that the avoidance of mathematics can lead to inappropriate use of CAA or unnecessary use of computers.

Item analysis (McAlpine, 2002) is a well-established application of mathematics (statistics) used for measuring the effectiveness of individual test questions (items).

Usually in the UK we restrict ourselves to considering the facility, a measure of the difficulty of a question, and the discrimination, a measure of how well the score on one question correlates with the test as a whole. The facility and discrimination of questions are calculated after a test has been marked and are easily included in reports by CAA software such as Question Mark Perception. Item Response Theory is more complicated, is widely used in the US and can include a third parameter, the probability of minimal ability candidates getting a question correct, C.

$$C = \frac{\text{Mean Uneducated Guesser's Score}}{\text{Maximum Score}} = \frac{\text{MUGS}}{\text{MS}}$$

For example, in a standard 4-choice MCQ with 1 mark for the correct choice and 0 marks for an incorrect choice, C = 0.25/1 = 0.25.

The aim of this paper is to consider what light mathematics can throw upon the scoring of objective questions. By analysing the scoring of common objective question types mathematically, some useful formulae and results are derived. These results provide scoring insight and help in identifying fair scoring strategies, which allow for partial credit and differing confidence levels.

Mean Uneducated Guesser's Score

This article was motivated by a paper at the 6th international CAA conference (Mackenzie and O'Hare , 2002). The authors developed an empirical Marking Simulator to assist test designers in question and test scoring. They used a Monte Carlo technique to determine (what they call) the base level guess factor or Mean Uneducated Guesser's Score (MUGS in our terminology) together with score distributions and percentage pass rates for different question types. Their computer simulations are successful in deriving some practical tables of results, but they:

- lack precision
- are inefficient and inelegant
- lack insight into the underlying mathematics

The probability of a fair coin coming down heads is not derived efficiently by tossing it 5000 times, the minimum number of trials used in their simulations! The authors argue that the large variety of question types used in their TRIADS system makes resort to simulations necessary. Nevertheless, all the tables that they provide could be generated analytically from formulae.

Most of the simulations yield a non-zero MUGS value. A value of 0 makes logical sense, but it does lead into the contentious issue of negative marking, for which Ryle (1996) offers a robust defence.

Three other papers at the same conference provoked further related questions.

1. Harper (CAAC 2002) considers post-processing marks to allow for guessing. How can scoring be set up for common objective question types to avoid post-processing, yet still allow for guessing?

2. Davies (CAAC 2002) considers the effect of using of explicitly specified confidence levels by students in MCQs. How can scoring be set up in such a way as to allow implicit specification of confidence levels by students?

3. McGuire et al.(CAAC 2002) consider issues of partial credit. How can scoring of standard objective question types address issues of partial credit?

Constrained and Unconstrained MCQs

Consider the humble MCQ. The conventional MCQ is constrained, where one and only one answer can be selected. Given one (correct) key and three (incorrect) distracters, a student who can eliminate two incorrect answers (or knows that the correct answer is one of two) may select an incorrect answer, gaining no credit for partial knowledge. A scoring scheme of 3 for the correct answer and –1 for an incorrect answer, despite having MUGS = 0, would award such a student negative marks!

Yet this simple MCQ can be delivered in one of three modes:
Constrained
number of selections allowed = number of correct answers = 1

Partially constrained
number of choices > max number of selections allowed > number of correct answers

Unconstrained
maximum number of selections allowed = number of choices

However many selections are allowed, MUGS = 0 still holds for this scoring scheme when it is applied cumulatively. The difference is that the "stakes" have been changed. The inclusion of the correct answer in two selections only scores 3-1 = 2 compared with the selection of a single correct choice which scores 3.

If this MCQ becomes unconstrained, a student may select more than one answer. When cumulative scoring of 3 for a correct answer and –1 for an incorrect answer is adopted, any score between –3 and +3 can be obtained. A student able to eliminate two incorrect answers with confidence would be awarded 2 marks, a student able to eliminate one incorrect answer with confidence would be awarded 1 mark and so on.

The process of eliminating two answers is familiar enough to contestants in "Who Wants to be a Millionaire" when they go 50:50. The following table 1 shows how the scoring operates for different modes of MCQ delivery

Mode	max number of selections made	max score	min score	MUGS
constrained	1	3 (¼)	-1 (¾)	0
partially constrained	2	2 (½)	-2 (½)	0
partially constrained	3	1 (¾)	-3 (¼)	0
unconstrained	4	0 (1)	0	0

Table 1 Scores (and Probabilities) for Different MCQ Delivery Modes

The corresponding probabilities are shown in brackets. In an unconstrained question a guesser can decide how many selections to make: gambling on getting low scores with higher probabilities than high scores and vice versa. Imagine a class of male and female guessers! The males might like to gamble, but the females play it safe. While
MUGS = 0 for both groups, the spread of results for the male group would be greater.

An important aspect of intelligently constructed scoring systems is that they are carefully explained to students. When students recognise that the aim is to increase the fairness of objective test scoring, support for alternative systems follows.

Constrained and Unconstrained MRQs

This approach becomes more interesting when it is applied to other question types.
A multiple response question or MRQ can be regarded as the generalisation of an MCQ with a wider range of scoring patterns possible. While Mackenzie and O'Hare do not include unconstrained MCQs among their tables, the vast majority of the questions, which they do consider, are MRQs of different types. For example, Table 2 shows typical results from their computer simulation of a constrained MRQ where 2 correct answers are to be selected from 5 possible choices :

Number of correct answers = 2 Total number of options = 5
Constrained delivery Negative scores resolved to zero
Number of iterations = 5000 Score for correct answer = 100
Residual BLGF = (((Q% - BLGF)/(100-BLGF)*100)

[BLGF = Base level Guess Factor = MUGS = Mean Uneducated Guesser's Score]

		Negative scores on incorrect options						
Parameter		0	-10	-20	-30	-40	-50	-60
BLGF		40	34	28	21	17	9	10
% passing at 40%		70	71	10	9	11	9	10
% passing at 40% of (100-BLGF)		10	10	10	9	11	9	10
40% pass mark equiv at 100-BLGF 45			64	60	57	53	50	46
Residual BLGF assuming Qscore =0 at BLGF		20	15	12	9	11	9	10
Score Node List	1	0	0	0	0	0	0	0
	2	50	40	30	20	10	100	
100								
	3	100	100	100	100	100		
% of candidates scoring on each node								
	1	30	29	30	30	30	91	90
	2	60	61	59	61	59	9	10
	3	10	10	10	9	11		

Table 2 Scoring Simulations for an MRQ 2/5 with Constrained Selection (extracted from Mackenzie and O'Hare, 2002, where it is incorrectly captioned)

This work raised a variety of questions, for which there were no immediate answers:

1. Could the tables of Mackenzie and O'Hare be generated algebraically without resort to Monte Carlo methods?

2. Is there a scoring scheme for MRQ questions which yields an expected mark of zero for guessers (MUGS=0), analogous to the MRQ.

3. How should the scoring be adjusted to achieve a non-zero mean uneducated guesser's score (MUGS = C)?

4. Is such a scoring scheme fair, with similar benefits for partial credit and expression of confidence?

and lastly the rather more mundane question

5. Is there a sensible notation which can be used for specifying objective questions?

e.g. MRQ(C2/5,0,0) might be used to define a constrained multiple response question with two choices correct out of 5, a mean uneducated guesser's score of zero and all negative scores resolved to zero.

Considering the first question mathematically, the percentage of candidates scoring on each of the three nodes in Table 2 is governed by the probabilities

$$\text{P(2 correct)} = \frac{1}{\binom{5}{2}} = 0.1 \quad \text{P(1 correct)} = \frac{\binom{2}{1}\binom{3}{1}}{\binom{5}{2}} = 0.6$$

$$\text{P(0 correct)} = \frac{\binom{3}{2}}{\binom{5}{2}} = 0.3$$

Hence the expected percentage of candidates scoring on each node should be 10, 60 and 30 repectively. The Monte Carlo approach, for which the results are shown in Table 2, yields only scattered approximations to these figures. In general (see Appendix 1) the probability of getting x correct answers is given by:

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, \ldots r$$

where N = number of choices
r = number of correct choices
n = number of selections made

In appendix 1 we show that the number of correct answers selected in a multiple response question (MRQ) follows a hypergeometric distribution with a well-defined probability function, mean and variance. Armed with these results the Mackenzie and O'Hare tables could be generated rigorously without resort to Monte Carlo methods, but we can go further in analysing the question scoring.

MRQ Scoring Systems

An MRQ may use rigid or cumulative scoring. Rigid scoring only awards marks for a fully correct answer. Cumulative scoring gives partial credit for one or more correct answers. How can partial credit be awarded for answers in view of the probabilities of guessing?

If we require MUGS = 0 then we show in Appendix 1 that the score for a correct answer *a* must be related to the score *b* for an incorrect answer by the equation

$$a = -\left(\frac{N - r}{r}\right) b$$

For example, in a MRQ(2/5,0) N = 5, r = 2. If we set a = 3 then b = -2.
The result is what you might expect. The score for correct choices is equal to the number of incorrect choices and the score for incorrect choices is equal to the negative of the number of correct choices. If a student selects all the choices in an unconstrained MRQ, then the total score will, not surprisingly be zero. It is noted that this result is independent of whether the MRQ is constrained, partially constrained or unconstrained. To recap:

A cumulatively scored multiple choice question, whether constrained, partially constrained or unconstrained, with N choices and r correct options has a Mean Uneducated Guesser's Score of zero (MUGS=0) when the incorrect answers are scored at $b = \left(\frac{-r}{N-r}\right)$ of the correct answer score *a*.

The familiar result for an MCQ is simply the special case of r = 1, e.g. for a 4 choice MCQ *a* = 3 and b = -1. For a constrained MRQ(C2/5,0) *a* = 3 and b = -2. The resulting table of scores and probabilities is shown in Table 3

Number of correct choices	Probability P	Cumulative score S	PS
0	0.3	-4	-1.2
1	0.6	1	0.6
2	0.1	6	0.6
			□PS = MUGS = 0

Table 3
Scores and Probabilities for a Constrained MRQ (2/5) with MUGS = 0 Scoring

Suppose we now examine the scoring pattern for an unconstrained MRQ(U2/5), i.e. relaxing the condition for constrained selection, but still adopting MUGS = 0 scoring of *a* = 3 and b = -2.

Number of selections	2 (or maximum) correct	1 correct	0 (or minimum) correct	MUGS
0	-	-	0 (1)	0
1	3 (0.4)	-	-2 (0.6)	0
2	6 (0.1)	1 (0.6)	-4 (0.3)	0
3	4 (0.3)	-1(0.6)	-6(0.1)	0
4	2 (0.6)	-	-3(0.4)	0
5	0 (1)	-	-	0

Table 4
Scores (and Probabilities) for an Unconstrained MCQ (2/5) with MUGS = 0 Scoring

In Table 4 above the probabilities are shown in brackets. It can be seen that the scoring pattern for 2 selections is necessarily the same as for a constrained MRQ.

For 0 and 5 selections the score is necessarily zero. For other numbers of selections an uneducated random guesser could "play the odds" and gamble at high or low stakes just as for an MCQ. The beauty of using a MUGS = 0 scoring system is that a student knowing that the correct two answers were among a set of three would get a guaranteed score of 4 out of 6 and gain partial credit for selecting those three choices. A candidate able to eliminate one choice would confidently select the other four and be guaranteed 2 marks.

The formula for the score variance is useful for determining how many students will exceed a given score, e.g. a pass mark, by guessing and enables the Mackenzie and O'Hare tables to be generated analytically.

Suppose, more generally, that a Mean Uneduated Guesser's Score MUGS = C = 100c is required, where c is the MUGS score for the question expressed as a percentage. Following the analysis of Appendix 1, the scoring system must now be adapted so that a and b satisfy the equation

$$\mathrm{n}\left(\frac{\mathrm{r}}{\mathrm{N}}\right)\mathrm{a} \quad + \quad \mathrm{n}\left(1-\frac{\mathrm{r}}{\mathrm{N}}\right)b = 100c$$

In this way, negative scoring could be eliminated at the expense of a mean uneducated guesser's score greater than zero, e.g. by setting $a > 0$ and $b = 0$. The variance of scores is determined from the same formula as given in Appendix 1 and pass rates can be calculated accordingly.

It is noted that an MRQ (2/5) is assumed to be of the form: "Which two of the following …?". A question of the type: "Which of the following …?" simply breaks down into 5 separate MCQ(1/2) questions.

CAA question authors should seriously consider the use of unconstrained MCQs and MRQs with intelligent scoring to get the benefits of partial credit and expression of confidence levels. As a further illustration of partial credit consider:

MRQ(C4/8)
probability of 4 correct by guessing = 1/70
probability of 3 correct by guessing = 16/70 $\approx$1/4

MRQ(C5/10)
probability of 5 correct by guessing = 1/152
probability of 4 correct by guessing = 25/152 $\approx$1/6

Rigid scoring would mark down those who did not get all the correct selections. The chances of getting all-but-one of the correct selections by guessing is quite small and some partial credit would seem appropriate.

Future Directions

The tables of Mackenzie and O'Hare (2002) could be generated in a spreadsheet using mathematical formulae and thereby avoid the inaccuracies introduced by Monte Carlo methods. Extensions to the tables can be made, e.g. by allowing the specification of MUGS = 0 or MUGS > 0 to give the appropriate scoring system. These would be useful in providing guidance on "pre-processing" of marks by use of intelligent scoring systems, rather than the more common post-processing.

Many other objective question types exist. The TRIADS CAA software (Mackenzie, 1999) has around 50 different question types, ranging from matching and ranking to extended matching items and drag and drop question types. For example, suppose a ranking question expects the answer 12345 and a student answers 23145. A natural way of awarding partial credit would be to consider the correlation coefficient as a measure of the answer correctness.

This raises the common problem of non-integer scores, which can easily arise in MRQs with the MUGS=0 scoring schemes. If multiplying factors are used to make scores integer, the problem of variation in question scores occurs. If software allows questions to be scored more flexibly, e.g. with non-integer values, such problems could easily be resolved. For example, it should be possible for a question author to specify the MUGS value and the maximum score for an MRQ question, without having to specify the scores for each choice.

Finally, we have considered the distribution of scores arising from a single MCQ or MRQ only. If several MRQ questions of the same type are delivered the total scores will follow a multivariate hypergeometric distribution, which is necessarily more complicated. If a mixture of question types are used the situation becomes more complicated still. It may be at this stage that

computer simulations rather than a mathematical analysis becomes prefererable.

Conclusions

Unconstrained or partially constrained MCQs are rarely used, yet they have the advantages of offering partial credit and an expression of confidence in the correct answer. In these circumstances negative marking is more acceptable and post-processing of marks to allow for guessing is not required.

MRQs may be delivered in constrained, partially constrained or unconstrained mode. Their scoring can be rigid (all or nothing) or flexible (cumulative). MRQs and MCQs can be designed to give a pre-specified Mean Uneducated Guesser's Score by setting appropriate scores for the correct and incorrect choices. The benefits of partial credit and confidence levels can therefore be achieved for both MCQs and MRQs. The score distribution resulting from guessing can be used to provide base pass rates

A more general conclusion is that considerable effort has been expended by mathematicians in developing and implementing CAA of mathematics and less effort has been put into the mathematics of CAA itself. The work described here illustrates one application of mathematics to scoring. While the use of computers may ultimately be necessary in calculating results, it is believed that rigorous mathematics rather than computer simulations should be used whenever possible.

Acknowledgements

We thank LTSN Maths, Stats and OR for including a first draft of this paper as the monthly article in the April 2003 on-line Maths CAA Series at
http://ltsn.mathstore.ac.uk/articles/maths-caa-series/index.shtml
and Prof. Cliff Beevers, O.B.E., for his helpful comments.

References

Davies, P. There's No Confidence in Multiple Choice Testing, Proceeding of the 6th Annual Conference on CAA (2002)
http://www.lboro.ac.uk/service/ltd/flicaa/conf2002/pdfs/davies_p1.pdf

Harper, R. Allowing for Guessing and for the Expectations from the Learning Outcomes in Computer-Based Assessments, Proceeding of the 6th Annual Conference on CAA (2002)
http://www.lboro.ac.uk/service/ltd/flicaa/conf2002/pdfs/harper_r1.pdf

McGuire et. Al. Partial Credit in Mathematics Exams – A Comparison of Traditional and CAA Exams, Proceeding of the 6th Annual Conference on CAA (2002)
http://www.lboro.ac.uk/service/ltd/flicaa/conf2002/pdfs/mcguire_gr1.pdf

McAlpine (2002) A Summary of Methods of Item Analysis, CAA Centre Blueprint paper 2,

Mackenzie and O'Hare , Empirical Prediction of the Measurement Scale and Base Level Guess Factor for Advanced Computer-Based Assessments, Proceeding of the 6th Annual Conference on CAA (2002)
http://www.lboro.ac.uk/service/ltd/flicaa/conf2002/pdfs/Mackenzie_d1.pdf

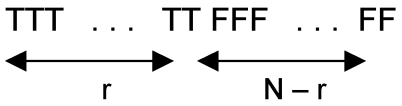
Mackenzie, D. , Recent Developments in TRIADS, Proceeding of the 3rd Annual Conference on CAA (1999)
http://www.lboro.ac.uk/service/ltd/flicaa/conf99/pdf/mckenzie.pdf

Ryle, A.P. (1996) Objective Tests: In Defence of Negative Marking, Life Sciences Educational Computing, Vol 7, No 1

Appendix 1
Application of the Hypergeometric Distribution to Multiple Choice and Multiple Response Questions

Consider a population of N possible answers to a question.

A known number, r where $0 \leq r \leq N$, of the answers are correct (T) while the remaining n-r answers are incorrect (F).



A random sample of n answers, where $1 \leq n \leq N$, are selected at random from the population without replacement. The number of correct answers in the sample is a random variable X having the hypergeometric distribution that has a probability function

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, \ldots r$$

it can be shown that

$$E[X] = r\left(\frac{n}{N}\right)$$

and

$$V[X] = n\left(\frac{r}{n}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Assume that each correct answer in the sample is awarded ‘a’ marks and each incorrect answer is awarded ‘b’ marks then the total marks for the question is a random variable T given by

$$T = aX + b(n - X)$$

$$= (a - b)X + bn$$

It follows that

$$E[T] = (a - b)\,E[X] + bn$$

$$E[T] = (a - b)\,r\left(\frac{n}{N}\right) + bn$$

$$E[\,T\,] = n\left(\frac{r}{N}\right)a \quad + \quad n\left(1-\frac{r}{N}\right)b$$

For E[T] =0 we have

$$n\left(\frac{r}{N}\right)a \quad = \quad -\, n\left(1-\frac{r}{N}\right)b$$

$$a \quad = \quad -\,\frac{n\left(1-\frac{r}{N}\right)}{n\left(\frac{r}{N}\right)}\,b$$

$$a \quad = \quad -\left(\frac{N-r}{r}\right)b$$

Note that the relationship between a and b is independent of the sample size n, i.e. whether the MRQ is constrained, partially constrained or unconstrained..

The variance of T is given by

$$V[\,T\,] \;=\; V[\;(a-b)X \;+\; bn\,]$$

$$= (a-b)^2\; V[\,X\,]$$

Hence

$$V[\,T\,] \quad = \quad (a\;-b)^2\left(\frac{n\,r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N}\right)$$