ALLOWING FOR GUESSING AND FOR THE EXPECTATIONS FROM THE LEARNING OUTCOMES IN COMPUTER-BASED ASSESSMENTS

Ray Harper

Allowing for Guessing and for the Expectations from the Learning Outcomes in Computer-Based Assessments

Ray Harper

Department of Sport, Exercise & Biomedical Sciences
University of Luton
Park Square
Bedfordshire
LU1 3JU

ray.harper@luton.ac.uk

Abstract

Computer-based assessments usually generate a percentage mark. It is not self-evident how this relates to the final percentage mark or final grade for the work since this depends on (i) its relationship to the "baseline" mark expected for someone who only guesses, (ii) to the "expectations" for the piece of work in relation to the learning objectives and (iii) the grading scheme employed. For some question types it is possible to allow for guessing within the marking scheme for the question using negative marking but in general it is preferable to correct for guessing within a post-test grading scheme that allows for guessing. The relationship between the assessment learning objectives and essays where choice is available and topics can be avoided compared with computer-based assessments where no choice is available and topics cannot be avoided is considered. It is concluded that commonly maximum performance should not be set at a mark of 100% but that an allowance should be made for the maximum expected performance based on the learning objectives. The use of formulae in a spreadsheet to convert the marks into grades based on a statistical allowance for guessing or additionally allowing for the maximum expected mark is demonstrated. A spreadsheet pro forma containing all of the formulae for adjusting marks and determining grades can be obtained by selecting “Grading” from the menu at http://students.luton.ac.uk/biology/webol/.

Introduction

Multiple choice assessments whether in paper form or as computer-based assessments have been used in a variety of institutions for many years. Farthing & McPhee (1999) have discussed the use of statistical analysis for the post-test analysis of assessments based on multiple-choice questions (MCQ). These authors have suggested using a permutational MCQ style with a low probability of guessing correctly in order to remove the contribution from guessing for final year assessments that may replace traditional essay-based assessments. However this approach is not suitable for the majority of assessments where good practice and the nature of the topics being assessed leads to the use of a variety of question styles.

Some authors refer to a "standard" guessing correction technique or formula (Brown, Bull & Pendlebury, 1997; Farthing & McPhee, 1999). This approach is fully described in some texts such as Schofield (1972). The formula is only suitable for assessments containing a single question style (e.g. multiple-choice) and structure (number of options). As such it can only be applied to a very restricted set of assessments. The alternative to post-test correction for guessing is to correct within the question marking scheme by setting negative marks for incorrect selections. Pritchett (1999) has discussed some of the disadvantages of this approach.

Brown et al (1997) have implied that is not worth correcting for guessing since, for example, rank order would not be altered and that intelligent guessing may be a skill to encourage. It is unlikely that being good at guessing is a learning objective in most situations where MCQ tests are used and so it is unlikely to form part of the assessment criteria. It may well be that for large groups (>100) a grade can be determined from the rank order but for other groups where there may be statistical variation from cohort to cohort and for grading directly from the mark then not correcting for guessing is not an acceptable approach. At the simplest level it can be seen that for an assessment based on true/false questions guessing alone will on average give 50%. If the pass mark were set at the traditional 40% then all would pass regardless of their ability. MCQ tests based on selecting 1 from 5 options give 20% for guessing the answer to all of the questions. A mark of 40% only represents 25% (100*20/80) of the useful range from 20 to 100% and it would still normally not be acceptable to set this as the pass mark. It appears that some form of correction for guessing is normally required. The correction must reflect the underlying statistical nature of MCQ tests and allow for the inclusion of questions without any choice component. This paper presents an approach to post-test correction for guessing that can be used in spreadsheets to automate the correction and generate a grade. The approach also allows for setting marks that reflect the difficulty of the question. Some features of relating marks to difficulty are considered. Even when using rank order to generate the grading it may be of benefit to initially apply the post-test guessing correction described below so that there is a clear knowledge of where the pass boundary lies.

Assessments may give a choice of questions, e.g. 3 from 6, such as in most essay-based examinations. Giving choice allows the selective discounting of

poor performance. This presumably reflects the learning objectives and their related assessment criteria that do not require understanding of all the topics. MCQ tests normally require all questions to be attempted. Should this type of assessment be used when the learning objectives are similar to those for assessments where choice is given then the mark will be depressed relative to what would be achieved if questions could be avoided. This paper provides an approach for combining (i) correcting for guessing with (ii) correcting for the expectations in relation to the learning objectives and assessment criteria.

Individual Questions

There are two main issues for marking and grading that are associated with individual questions: (a) relating marks to difficulty and (b) deciding how to address guessing.

(a) Relate marks to difficulty

The following discussion assumes that the questions in a single assessment are being considered and that all distracters are authentic and of equal merit. Some questions may be more demanding than others are. This can be due to one or more of the features identified below:

(i) The topic

Questions may ask for the recall of knowledge but for one question this may be on a general point but for another be on a very specific point of detail and hence may be considered more demanding.

(ii) The cognitive process

The learning objectives for the question may be related to different cognitive processes, knowledge, analysis, etc. as originally described by Bloom B. (1956). It may be considered that a question recalling knowledge is not as demanding as one requiring analysis and so forth when comparing the different cognitive activities.

(iii) The style of the question

A question style that requires a response to be typed in and does not have any options to choose from may be considered more demanding than one where the user is choosing between given options. A multiple-response question (MRQ) may be considered less demanding than equivalent MCQ's.

1a. MCQ (6 marks)	1b. MCQ (6 marks)	1c. MRQ (2 x 4 marks)
Which of the following lipids can form the bilayer that is the basis to the membrane structure of animal cells?	Which of the following lipids can form the bilayer that is the basis to the membrane structure of animal cells?	Which two of the following lipids can form the bilayer that is the basis to the membrane structure of animal cells?
phosphoglycerides# triacylglycerols cholesterol waxes	sphingomyelins# triacylglycerols cholesterol waxes	[ ] phosphoglycerides# [ ] sphingomyelins# [ ] triacylglycerols [ ] cholesterol [ ] waxes
#=Correct. Basic mark=2 x number of distracters(3)		

Table 1: Comparing two MCQ's with one MRQ

If identifying each correct element of this MRQ (Table 1c) is viewed as of equivalent difficulty to two separate MCQ's (Table 1a & 1b) then 6 marks can be assigned to each correct component. However, if it is felt that seeing all of the options together in MRQ style (Table 1c) makes the question relatively easier than two separate MCQ's then 4 marks may be assigned to each correct component. For the MRQ in Table 1c, seeing phosphoglycerides and sphingomyelins together in the MRQ may trigger the memory that they have similar structures and properties and that these properties allow bilayers to form. This makes the question simpler that the two separate MCQ's where there are no such memory triggers; so the mark can be set less than for two MCQ's.

(iv) The structure of the question
A MCQ structured with four distracters is more demanding than one with three distracters. A similar argument applies to multiple response styles of question.

Hence it is normally not appropriate to set the same mark for all questions. The marking regime used should allow the flexibility of giving different marks to the questions.

Recommendations that address these issues are:

(1) Use a consistent structure (number of distracters and correct elements) for each question style.
(2) Base marks relative to a knowledge-based MCQ that has a mark equal to the number of distracters or a multiple thereof.
(3) Only adopt a new question style after determining how it will fit into the marking system and into the grading scheme (see below).

(b) Decide how to address guessing

For some question styles, such as those requiring keyboard input without any options, guessing is not an issue. However for other styles, such as MRQ and matching, the user can always make a guess. For some question styles, such as MCQ and MRQ, the marking regime can allow for guessing by setting negative marks for the wrong options. This requires an analysis of the probabilities for the possible responses. For example, a MCQ with 3 distracters can be given 3 marks for the correct option and minus 1 for the distracters. Guessing the answers to all of the questions would on average give a mark of zero. Although it may be possible to create a default-marking scheme in which setting a negative mark for an incorrect choice is straightforward, after marks have been adjusted for the difficulty of the question it may become impossible. In the above example if the mark had been increased to 4 for difficulty the incorrect responses need setting to minus one and a third. This is impossible in decimal notation (though the small rounding error may be acceptable) and is impossible in some software question designers that require integer marks. An overall change to the marking scheme may be required so that an incorrect response can be given by an integer. In this example the marks can be scaled up by a factor of 3 so that the default mark for questions is set to 9. For this example question the marks become 12 for correct and minus 4 for incorrect. However, the

requirements for questions will vary and the approach can rapidly become impracticable.

Correcting for guessing within the marking scheme for an individual question has the advantage that users do not have to answer all of the questions. Under these conditions not attempting the question would be given a mark of zero. This allows intelligent determination of when to guess based on how many of the distracters can be discounted through knowledge. This approach also allows setting assessments that require attempts at only a proportion of the questions. This is analogous to the approach commonly used for essay style examinations (3 from 6, etc.) and is discussed below when addressing the expectations from multiple-choice assessments. Post-test approaches to correcting for guessing such as those described below require all of the questions to be attempted and this should be included in the instructions. This ensures that the users who would otherwise not guess if they did not know the answer are not disadvantaged by not gaining marks from guessing questions.

As shown in the examples above, setting negative marks requires knowledge of the probabilities for choosing correct/incorrect for the question. This may beyond the statistical competence of many question designers, particularly if the assessment includes some of the "matching" or "permutational" questions (see Table 3) for which the process is more complex. When not correcting for guessing within the question the correction needs to be addressed when grading the assessment (see Overall Grading below). Pragmatically, the latter approach reduces the time required to set the questions since it avoids the requirement to determine or set any mark for incorrect responses. Incorrect responses can be given the default mark of zero.

Overall Grading

It is not self-evident how the % mark from a computer-based assessment can be used to generate the grade for the work whether this is expressed as a percentage, grade point or other scheme. The grading depends on the relationship of the mark to the "baseline" mark produced by making an allowance for guessing and to the "expectations" for the piece of work in relation to the assessment criteria. It can be useful to compare the assessment with other forms of assessment such as essay-based examinations.

(a) Baseline Mark

The "baseline" % is the mark for someone who knows nothing about the subject and so must always be guessing. This "baseline" % reflects the proportion of the different types of questions used and the marks allocated to the questions. The contribution from guessing can be determined statistically and for most assessments can be determined by entering summary data into a spreadsheet. Examples for the common question styles are shown in Table 2. For question styles that have more than one "maximum question mark", reflecting the difficulty of the question as discussed above, then additional entries need to be entered for each case. The match (once) question style where options in a list can only be used once requires a much more complex formula and a different formula for different numbers of options. The latter also

applies to permutational questions. The contribution from guessing can then be set to zero as shown in the first two columns of Table 3. Here the formula adjusts the marks so that the baseline value is set to zero. Hence the working range for the adjusted marks shown in the second column is from 0 to 100. Having worse than average luck when guessing can give a mark of less than zero; these should be treated as zero.

Style	Options	Select	Max'm. Question Mark	No.	Style Maximum Mark	Guess Mark
MCQ[1]	4	1	3	28	84	21
MRQ (Mark parts) [1#]	5	2	4	4	16	6.4
MRQ (All correct)[2#]	5	2	6	0	0	0
Hot Spot (No options) [1]	1	0	3	2	6	0
Hot Spot (Options) [1]	4	1	3	4	12	3
Match (Use any option)[1#]	4	1	12	0	0	0
Match 3 (Use once)[4#]	3	3	9	0	0	0
Match 4 (Use once)[4#]	4	4	12	0	0	0
Match 5 (Use once)[4#]	5	5	15	0	0	0
Numerical[1]	1	0	3	2	6	0
Permutational MCQ[3#]	5	2	6	0	0	0
Text Area Input[1]	1	0	3	0	0	0
Text Input[1]	1	0	3	0	0	0
Text Area Keywords[1]	2	0	6	0	0	0
			Totals =	40	**124**	**30.4**
					Guess % =	**24.52**

[1]Guess Mark Formula (includes no guessing) = No.*Question Mark*Select/Options
[2]Guess Mark Formula = No.*Question Mark*Select factorial /(Options*(Options-1)
the denominator continues for Select number of components (Options-2), etc.
[3]Guess Mark Formula = No.*Question Mark /(Options*(Options-1)
the denominator continues for Select number of components (Options-2), etc.
[4]Guess Mark Formula - See Appendix 1
[#]The formula assumes each selection/option is equally weighted.
Note: Where possible formulae are provided in a form for readers not statisticians.

Table 2. Summary of Questions - An Example

Mark %	25% Baseline Adjustment[#]	Set 80% as maximum*
0.0	-33.3	-41.7
20.0	-6.7	-8.3
25.0	0.0	0.0
30.0	6.7	8.3
40.0	20.0	25.0
49.0	32.0	40.0
50.0	33.3	41.7
55.0	40.0	50.0
60.0	46.7	58.3
70.0	60.0	75.0
80.0	73.3	91.7
85.0	80.0	100.0
90.0	86.7	108.3
100.0	100.0	125.0
Guess =	25	
Max'm =		80
[#] =(Mark - Guess)*100/(100 - Guess)		
* =Baseline adjusted mark *100 / Max'm		

Table 3. Adjusting marks for guessing and expectations

(b) Expectations

Using as a "standard" an essay based examination with a choice of 3 from 6; it can be seen that an understanding of all of the topics set is not required. The 3 topics perceived as being most demanding would not be attempted. The assessment criteria for the multiple-choice assessment may similarly not require the demonstration of an understanding of all of the topics set. With computer-based assessments it is normally not possible to allow a subset of the questions to be attempted. As discussed above, using a subset of questions is only possible if negative marking is used and this approach is usually more difficult and time-consuming than the post-test approach. Attempting all of the questions has the effect of depressing the mark relative to an assessment where the user can selectively discount their poor performance. On this basis some of their poor performance should be discounted in the grading scheme, i.e. 100% should not be set to represent correct answers to all of the questions because this is an unreasonable expectation. The marks can be corrected for the actual maximum performance expected. An example of this correction is shown in column 3 of Table 3 where the maximum expected score is set to 80%. The final grade % is in the usual 0 to 100% range. This can be converted to other grade scales in a spreadsheet using a lookup table as shown in Table 4. This argument does not apply where the learning objectives are that for the topics assessed everything should be understood. A spreadsheet pro forma containing all of the formulae for adjusting marks and determining grades can be obtained from http://students.luton.ac.uk/biology/webol/.

Name	Mark %	Grade %†	Grade Point#
Student 1	50.0	41.7	5
Student 2	70.0	75.0	14
Guess % =	25		
Maximum % =	80		

Grading Lookup Table:

(%)	Grade	Point
0	G	0
30	F	2
35	E	4
40	D-	5
43	D	6
47	D+	7
50	C-	8
53	C	9
57	C+	10
60	B-	11
63	B	12
67	B+	13
70	A-	14
80	A	15
90	A+	16

†Formula =
(Mark -Guess)*100/(100-Guess)*100/Maximum
#=VLOOKUP(C8,$A$18:$C$32,3) or
=VLOOKUP(Cell reference,Lookup range,Column)

Table 4. Pro forma for converting marks to grades

Further Considerations

In table 3 it can be seen that for this example the range of the original marks that is finally used is from 25% to 85%, i.e. 60% of the marks. This means that there must be sufficient questions to provide grading discrimination within this range. Considering 50 questions and assuming they are equally weighted then only 30 contribute to the useable range. Assuming the normal fail and first class grading boundaries of 40% and 70% respectively then only 9 questions provide discrimination through all of the third and second class categories. All the above calculations have a statistical basis and give values for users who have average luck (a normal distribution) for questions for which they are guessing the answer. It is quite possible for a lucky/unlucky user to get 2 or 3 more correct (or incorrect) from guessing than average and move up/down by a whole honour's category. The chance that the grade will not reflect ability but will reflect more or less luck than average should be kept to a minimum. This may be achieved by raising the number of questions in the useable range (i) by increasing the number of questions or (ii) by decreasing the "baseline" mark. Increasing the number of distracters in multiple-choice questions will reduce the "baseline" mark. However, there are practical limits to increasing the number of distracters since it is often difficult to generate many authentic distracters as discussed by Pritchett (1999). The same effect

can be achieved by increasing the proportion of questions where no guessing occurs or those for which guessing has little chance of selecting the correct answers such as permutational MCQ's as proposed by Farthing & McPhee (1999).

After determining the grades the pattern of grades should be considered. A normal distribution is expected with a few failures, a few high grades and most in the middle range. If this is not the case the distribution may be skewed high or low, i.e. most of the students are not centred on the middle range or the whole distribution may be shifted high or low. This should lead to an analysis of the facility and discrimination indices for the test and for individual questions as discussed by Farthing & McPhee (1999). The analysis may identify whether the test was too easy or too difficult and may identify the questions that were too easy or too difficulty. The analysis also helps to identify whether any questions were poorly designed. The results of the analysis may provide the basis for raising or lowering the maximum % discussed above in order to re-grade the assessment making allowance for any anomalies identified.

Conclusions

Through applying a consistent approach to question design and marking and using a statistical determination, the assessment mark can be adjusted by a factor that allows for the contribution from guessing. Similarly an allowance can be made for the maximum mark expected for the assessment in relation to the learning objectives. These adjustments can be made with formulae in a spreadsheet that generates the grade mark automatically.

References

Bloom B. S. (1972) *Taxonomy of Educational Objectives*: Handbook 1, Cognitive Domain. New York: Makay

Brown, G., Bull, J. & Pendlebury, M. (1997) *Assessing Student Learning in Higher Education (p.93)*, Routledge

Farthing D. & McPhee D. (1999*) 'Multiple choice for honours-level students? A statistical evaluation'*. In Proceedings of the 3rd Annual CAA Conference (pp. 105--116), Ed's Danson M & Sherratt R. Loughborough.

Pritchett, N. (1999) *'Effective Question Design'*. In Computer-Assisted Assessment in Higher Education (pp. 29--37), Ed's Brown S., Bull J. & Race P. Kogan Page Ltd. London

Schofield, H. (1972) *Assessment and Testing: An Introduction (pp. 39-40)*. George Allen & Unwin Ltd.