Multiple choice for honours-level students? A statistical evaluation

Principal contact:
Dave W Farthing
School of Computing
University of Glamorgan
PONTYPRIDD
CF37 1DL
UK
Email: dwfarthi@glam.ac.uk

Duncan McPhee
School of Computing
University of Glamorgan
PONTYPRIDD
CF37 1DL
UK

Abstract

Most computer-aided assessments use non-subjective techniques, such as multiple-choice questions (MCQs). However, when preparing summative assessments for honours-level degree students in their final year, MCQs are rarely considered suitable; by and large we tend to use traditional essay-type questions which are not quite so amenable to computer assistance.

We present here a technique for evaluating MCQs statistically. This evaluation can be used for:

- identifying how an MCQ can be improved; and
- proving that an MCQ test is a valid technique for assessing honours-level students.

We also present a new form of MCQ, called the Permutational Multiple Choice Question (PMCQ) or "Two-stem" MCQ. These questions have been trialled with honours-level students. Statistical analysis of the trials indicates that these PMCQs can be as good a predictor of overall performance as essay-type questions. The statistical analysis also supports practical suggestions for improving PMCQs.

We conclude that statistical analysis is useful, and good computer aided assessment systems should assist the examiner in analysing test results statistically.

Introduction

Traditional multiple-choice questions offer many benefits over traditional essay-type questions:

- consistency, reliability and efficiency in marking;
- broad coverage of syllabus; and *of course*
- the possibility of automating marking.

However, they are rarely used for final year examinations in honours-level degree courses, because many claim that (Popham, 1981):

- they may be answered simply through guessing;
- they assess only trivial recognition of facts, rather than high-level thinking, such as evaluation and synthesis;
- they offer a choice, rather than ask the candidate to construct the answer.

This paper explains a statistical approach that can be used to verify the validity of multiple choice questions. The approach can easily be adapted to analyse other non-subjective forms of assessment.

This paper also describes how this statistical approach has been applied to trials of a new type of multiple-choice question: the Permutational Multiple-Choice Question (PMCQ).

Statistical analysis

If we are to use MCQs for summative assessment of final year degree students, we must be confident that it is a valid and reliable form of assessment, and that it is a reasonable substitute for traditional assessment. In the first part of this paper, we describe a statistical approach to verify whether a given MCQ test is satisfactory, and if not, how it can be improved. This approach has three parts:

- Ensuring that a test performs satisfactorily by analysing *the entire test.*
- If this reveals a problem, *each question* must be analysed individually.
- If we can't identify why a given question is performing badly, we may need to analyse *each option* (putative answer) in that question.

Performing this analysis using a spreadsheet is fairly straightforward once the data has been entered. However, computer-aided assessment opens up the possibility of having this done automatically for us.

Statistical analysis of the entire test

There are two measures of an entire test: facility and discrimination.

Facility - how easy or difficult the test is

The simplest analysis of a test is to establish how easy or difficult it is in overall terms. The word we use to describe this is *facility* and is simply the average mark of all candidates. When we set a norm referenced test we expect the average mark to be in the region of 50%.

Discrimination - how well the test identifies the stronger candidates

A more demanding statistic is *discrimination:* how well a test correctly identifies the stronger candidates from the weaker ones. The immediate problem this raises is how we decide who is stronger and who is weaker. We have to compare the test results with other assessments (examinations, courseworks etc.) in the subject under consideration.

A common way of measuring discrimination is to correlate the mark from the MCQs with the mark from the other assessment. A correlation coefficient is a figure in the range minus one to plus one. A result of plus one means the two sets of results go hand in hand; all those who did well in the MCQ test did well in the other assessment and *vice versa.* Provided the sample size were sufficiently large, it would be reasonable to assume that the two assessments were assessing the same thing, and that one form could be substituted for the other.

> Note: In practice, it is highly unlikely that such a good fit will be achieved. However, a correlation coefficient of 0.7 or higher should be achievable. We consider a correlation coefficient of about 0.8 to be quite satisfactory.

A correlation of zero means the two sets of results are unrelated; of those who did well in the MCQ test, some fared badly in the other assessment, some did well, and others were middling. Either the two assessments are assessing different qualities, or one of them appears to be giving random results.

> Note: This can also happen when an MCQ test is so difficult or ambiguous that everyone has to guess the answers. Some candidates may get several correct answers by chance (unless we take precautions against guessing), but the lucky guessers won't necessarily be the stronger candidates.

A correlation coefficient of minus one would be very rare indeed in ordinary assessment; it would imply that candidates who did well in the MCQ test fared badly in the other assessment and *vice versa.*

Figure 1 demonstrates a test result graphically. Here the candidates are ranked by their examination results (diamonds), and the results of their MCQ test are overlaid (squares).

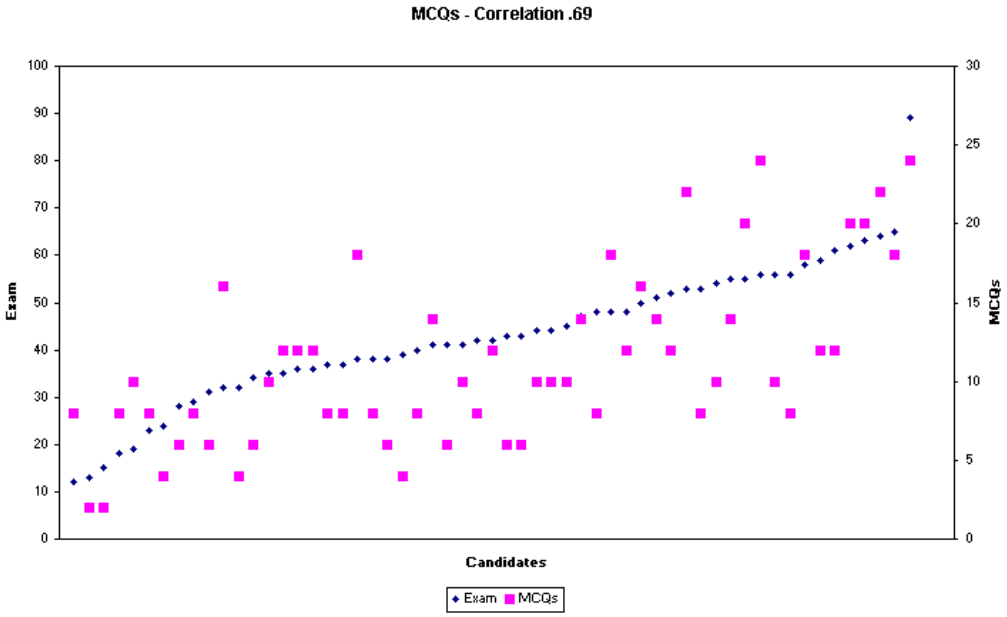

Figure 1: Results of MCQ test compared with examination results - Correlation 0.69.

Validity of statistical comparisons

Clearly some of the stronger candidates achieved a lower mark in the MCQ test than some weaker ones. We now need to ask why the two sets differ. It may be that the examination permitted candidates to be selective about which topics they revised whereas the MCQs assessed all the topics. It may be that some of the MCQs were ambiguous, confusing some stronger candidates. If so, the test probably needs to be improved.

On the other hand, perhaps we are correlating the test results with some invalid or unreliable data. For example, if an MCQ test correlates badly with a traditional exam, can we be sure the exam is returning valid results? We all know of perfectly capable people who perform badly under examination conditions. Perhaps an MCQ test with a low correlation coefficient is more accurately identifying the capable candidates. It may be that MCQs are better at identifying strong candidates impartially than formal examinations.

In general, the authors take the view that traditional examinations are a reasonably accurate indicator of stronger and weaker candidates. We aim to make our MCQ tests produce a similar profile to essay-type questions. Hence we talk about MCQs as a *substitute* for essay-type questions.

If we decide we want to improve the facility or discrimination, we need to analyse each question to identify which ones must be improved.

Statistical analysis of each question

Facility - how easy or difficult each question is
If the facility of the whole test is too low, analysis of the individual questions will help us identify which particular questions have an abnormally low facility and are in need of improvement or replacement. We can achieve an overall facility of 50% (or whatever figure is desired) in one of three ways:

1. Questions of increasing difficulty are set, so all but the very weakest candidates get the first one right, most get the second one right and so on until the final question which is answered correctly by only the very strongest candidates.
2. Every individual question has a facility of 50%, so about half the candidates select the correct answer on each question.
3. Some compromise between the previous two approaches, where *some* questions are harder (low facility) and others are easier, but on balance the overall average mark is 50%.

The first approach may seem attractive, but if a candidate's strongest topics are tested in the last few questions, he/she may have difficulty answering *any* correctly. The second approach is ideal but may be difficult to achieve, so the third approach is acceptable in practice.

Discrimination - how well the question identifies the stronger candidates
If the discrimination of the whole test is too low, analysis of the individual questions will help us identify which particular questions have an abnormally low discrimination and are in need of improvement or replacement. There are two techniques for calculating the discrimination of an individual question: correlation (again) and the Discrimination Index.

Correlation
As when calculating the discrimination of the entire test, we can correlate the mark for an individual MCQ with some other set of figures, normally the test marks. The important difference here is that a correlation coefficient of around 0.4 is quite acceptable; indeed it is difficult to achieve a correlation coefficient greater than 0.5 of tests with more than 20 candidates.

To explain this, we need to consider an imaginary question that achieves the best discrimination possible: all the weaker candidates answer incorrectly, and all the stronger candidates answer it correctly. Figure 2 shows the results from an imaginary "perfect" question: the weaker half of the class answers it incorrectly (0 marks), and the stronger half answers it correctly (2 marks). However, even such a "perfect" question cannot achieve a correlation of 1.0. Indeed, if *every* question in a test had that profile, half the class would gain 0%, and the other 100%; not a typical distribution of marks!

In practice we would prefer a profile more like that shown in Figure 3; most *but not all* of the weaker candidates will get it wrong, and most *but not all* of the stronger will get it right. This profile produces a correlation coefficient of

around 0.5; this is the best we can hope for when analysing individual questions.

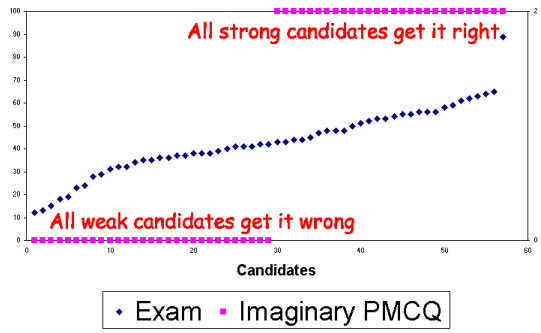

Figure 2. Imaginary "perfect" question - Correlation 0.8.

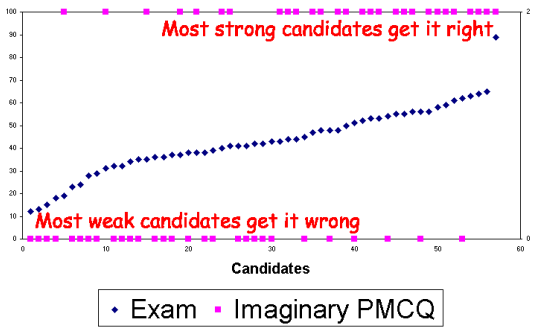

Figure 3. Imaginary "near perfect" question - Correlation 0.5.

Discrimination Index

The Discrimination Index is another technique that help us to identify questions that need to be improved (see Macintosh & Morrison, 1969 and others). Calculating the Discrimination Index is feasible without computer assistance.

1. Determine how many candidates form 27% of the total class (we'll call this *n*).
2. Select the test papers of the strongest 27% of candidates.
3. For a given question, add up how many of these answered correctly (N_S).
4. Select the test papers of the weakest 27% of candidates.
5. Add up how many of these also answered it correctly (N_W).
6. Subtract N_S - N_W and divide by *n*.

$$D = \frac{(N_S - N_W)}{n}$$

A Discrimination Index of 0.4 and above is quite acceptable.

Curing poor discrimination

There are many guidelines on preparing MCQs (e.g., Haladyna, 1995 and Farthing, 1998). They may reveal why stronger candidates found a given question difficult, or why weaker candidates found it easy. If the cause of the problem isn't immediately obvious, we will need to analyse which candidates selected which options.

Statistical analysis of each option

The question part of an MCQ is known as the *stem.* Among the putative answers or options, the correct one is called the *key* and the others are *distracters.* Frequency analysis helps us to identify the cause of an

abnormally high or low facility. But to correct a poor discrimination is more difficult.

Frequency analysis
If we expected an overall facility of 50%, we would expect every key to be selected by *about* 50% of candidates in each question. We would hope that each distracter would be selected by roughly the same number of candidates. So a question with one key and three distracters should be chosen in roughly the proportion 3:1:1:1.

If a question has an abnormally low facility, we identify which of the distracters most candidates tended to select in error. Is the question ambiguous? Could the popular distracters reasonably be considered correct? Is the key too unconvincing?

Conversely, for a question with a high facility we look for distracters rarely selected. Can we replace them with more plausible distracters? Is the key obvious because it contains a cue from the question? Can the question be rephrased to be less obvious?

Improving discrimination

Identifying distracters chosen by stronger candidates
The above process can be extended to identify why a question exhibits poor discrimination. Ambiguity, distracters that are arguably true and similar errors will confuse stronger candidates along with the weaker ones.

We can look for distracters that were *more* popular among the stronger candidates (those in the top 27%) than among the weaker ones. For example, perhaps 10 of the top candidates and only 2 of the bottom candidates selected distracter (b). An inspection of the wording may reveal why this should be.

Assessing higher level thinking
We also need to consider what type of intellectual process a given question is assessing. Bloom's taxonomy of cognitive educational objectives is commonly used when discussing various learning processes (Bloom, 1956). It describes six categories of intellectual process:

- *Knowledge:* remembering and recalling facts (lowest level).
- *Comprehension:* perceiving and understanding what has been learnt.
- *Application:* using knowledge in a specific manner.
- *Analysis:* separating a concept into its elements, and determining their relationships.
- *Synthesis:* combining elements into something new.
- *Evaluation:* exercise of judgement about value (highest level).

Candidates of all abilities may answer a question that assesses only knowledge, especially trivial recall of facts. A question that demands some analysis, creativity or evaluation would be more likely to discriminate well between candidates.

A new form of MCQ

Problems with traditional MCQs
Suppose candidates taking an MCQ test guessed the answers at random. If each question had one key and three are distracters, we'd expect their mark to average 25%. However, some would be lucky and gain a higher mark, and others a lower mark. If there were ten questions in the test, about *one in five* who guessed would be lucky enough to gain a mark of 40% or more just through random guessing.

There is a standard guessing correction technique that produces an average mark of 0% for candidates who guess. It also reduces the likelihood of gaining a pass mark purely by chance to only *one in fifty.* However, the technique has the disadvantage that it can result in candidates being awarded negative marks.

How PMCQs solve the problems
A PMCQ has a two-part stem and six putative answers: two of which are keys and four are distracters (see Figure 4). To answer the question correctly, the candidate must match up each stem with the appropriate key. The two parts of the stem must ask about closely related issues. Typically PMCQs ask candidates to distinguish between two similar concepts. All of the options should be feasibly correct for both parts of the stem.

Q4. Many organisations plan their computer procurement policy to secure advantages. Which of the following is a characteristic of *mixed suppliers,* and which of *preferred suppliers?*	
Mixed suppliers	e
Preferred suppliers	c
a.	Allows the customer to build on expertise with a given manufacturer's equipment
b.	Ensures consistency among components, such as common Human Computer Interaction
c.	Encourages competition, but avoids the overheads of repeated competitive tenders
d.	Payment may be staged over a long period of time
e.	Permits price competition and avoids lock-in to a single manufacturer
f.	All components are designed to work together, so reliability increases

Figure 4. An example of a PMCQ

In Figure 4, the two parts of the stem are orthogonal concepts rather than opposites, as some candidates may assume. The correct answers (e and c) have been entered for demonstration purposes.

Unlike traditional multiple-choice questions, PMCQs are not susceptible to candidates guessing the correct answer. Because there are thirty permutations of possible answers (6×5), the chance of getting both parts correct through random guessing is small. Candidates who guessed all the answers in a PMCQ test could expect a mark of only 3% (compared with 25% in a "choose one from four" test), and the likelihood of gaining a 40% pass mark in a test of ten PMCQs would be only 1:4500 (rather than approx. 1:5). These figures are at least comparable with traditional essay-type questions. Further, since there is no need for guessing correction, we avoid the possibility of producing negative marks.

Results of trials

Three trials of PMCQs in final year undergraduate examinations were held in 1996, 1997 and 1999.

In the first trial, candidates answered - on paper - five mandatory PMCQs and three essay-type questions about managing databases. The trial was extended in 1997 to include 15 PMCQs, which accounted for 30% of the exam marks.

For the third trial, we created 15 new PMCQs for a different subject; these also accounted for 30% of the exam marks.

A summary of the statistical analysis is in Table 1.

	Question type:	**PMCQ**	Essay	Essay	Essay
Trial 1 1996 N=44	Mean: (Facility)	54%	34%	36%	48%
	Correlation: (Discrimination)	.**598**	.806	.708	.808
Trial 2 1997 N=54	Mean: (Facility)	41%	34%	54%	
	Correlation: (Discrimination)	.**808**	.806	.807	
Trial 3 1999 N=57	Mean: (Facility)	38%	45%	45%	
	Correlation: (Discrimination)	.**693**	.793	.727	

Table 1. Results of three trials

Trial 1 in 1996

Analysis of the first experiment in 1996 proved a little disappointing. The correlation between the marks for the PMCQ section and the overall marks was much lower than we had hoped. A correlation coefficient of only 0.598 means some weaker candidates did well with the PMCQs, and a few stronger candidates performed less well.

Trial 2 in 1997
The second trial, in the same subject, was much more encouraging. Differences between the three correlation coefficients are statistically insignificant. Although it would have been preferable if every section had correlated closer to one, what we can say is that all three section marks are equally good as a predictor of the total mark. Every section of the exam paper discriminated equally well between the stronger and weaker candidates.

Trial 3 in 1999
In the third trial, all new PMCQs were written for a different subject. The correlation coefficients of both the PMCQ section and the second essay section were rather low. We undertook an analysis of the individual PMCQs to see how to improve things next year.

Analysis of individual PMCQs

In this section we give two examples of poor-performing questions. Statistical analysis was used to determine the cause of the problem.

Because the PMCQs were used in a final year examination for an honours-level degree, all of them *should* have assessed high order intellectual processes. When we critically re-examined each question using Bloom's taxonomy (Bloom, 1956), we found some were not so intellectually demanding. In general, those questions exhibited poor discrimination.

For example, Figure 5 shows a question that should have demanded evaluation skills (the highest intellectual process according to Bloom).

Q5. Which of these rule sets gives the *most* specific instruction, and which the *least* specific?	
Most specific	e
Least specific	f
a.	Information Engineering (James Martin's analysis and design method)
b.	BCS Code of Practice
c.	PRINCE: Projects in a controlled environment
d.	BS7799: Code of practice for information security management
e.	CRAMM: The CCTA Risk Analysis and Management Methodology
f.	BS EN ISO9000: the international standard for quality systems

Figure 5. Low-level thinking processes

We expected the candidates to evaluate each option, drawing from their own experience. Instead we found that candidates answered it in one of two ways:

- they expressed an opinion not based on experience; or
- they recalled a lecture slide that ranked them.

The first of these is a valid approach, but not easily amenable to a multiple-choice style of question. The second is a low-level thinking process. Either way, it is not surprising that this question exhibited almost no correlation with the overall marks (0.05). Stronger and weaker candidates alike selected almost all options in equal proportions. This question must be discarded or rewritten completely.

Figure 6 shows a question that required analytical thinking, but nevertheless exhibited almost no correlation with the overall marks (0.04).

Analysis of who selected which options revealed that several of the stronger candidates wrongly thought options (a) and (b) best describes *leadership.* Also, some stronger candidates wrongly thought options (b) and (d) best describes *management.* Clearly, options (a), (b) and (d) were too confusing to these candidates and need to be redrafted.

We feel that this question is valid, and can be improved. Redrafting it should be sufficient to improve discrimination next year. Although the lecture notes are quite clear on this topic, we will also look at how we can improve the delivery of the teaching material.

Q6. Which of the following *best* describes the personal role of *leadership*, and which *management*?		
Leadership	e	
Management	c	
a.	An interpersonal quality that has been scientifically identified as something a "great person" is born with	
b.	The application of comparative techniques to ensure optimal efficiency of a procedure or process	
c.	Getting things done through other people to achieve stated organisational objectives	
d.	The ability to empower others, through the use of policies and guidelines	
e.	A quality earned by the ability to encourage, motivate and inspire others	
f.	The ordering of people to do jobs they don't want to do	

Figure 6. Confusing distracters

By comparison, the question given earlier in Figure 4 performed quite well. It required candidates to evaluate the options - the highest thinking process according to Bloom - and correlated well with overall marks (0.44). Evaluation of the options has revealed how we might improve its rather low facility (30%).

Conclusions

We conclude that statistical analysis can help improve the quality of non-subjective testing. Good computer aided assessment software should assist the examiner in performing this analysis. A statistical analysis of each

question and each option could be produced automatically to highlight scope for improvement. Although data from the assessment software could be transferred into a spreadsheet for analysis, there remains significant work for the examiner in setting up the formulae.

We also suggest examiners consider using PMCQs for summative assessment of final year degree students. Care needs to be exercised in drafting such questions, but statistical analysis of trials should ensure a satisfactory result.

References

Bloom, B. S. (1956) Taxonomy of educational objectives: Handbook of cognitive domain, New York: McKay

Farthing, D. W. (1998) 'Best practice in non-subjective assessment', Monitor, (10), 24-26, ISSN 0961-3757

Haladyna, T. M. (1995) Developing and validating multiple-choice test items, Hillsdale, NJ; Hove, UK: Erlbaum

Macintosh, G. & Morrison, B. (1969) Objective testing, London: Unibooks, University of London Press

Popham, W. J. (1981) Modern educational measurements, Englewood Cliffs, NJ: Prentice Hall