Permutational Multiple-Choice Questions: An Objective and Efficient Alternative to Essay-Type Examination Questions

Dave W. Farthing
University of Glamorgan
Pontypridd
CF37 1DL, UK
+44 (0)1443 480480
dwfarthi@glam.ac.uk

Dave M. Jones
University of Glamorgan
Pontypridd
CF37 1DL, UK
+44 (0)1443 480480
dmjones@glam.ac.uk

Duncan McPhee
University of Glamorgan
Pontypridd
CF37 1DL, UK
+44 (0)1443 480480
dmcphee@glam.ac.uk

1. ABSTRACT

Despite the attractive characteristics of multiple-choice questions – efficient to mark, not subjective, etc. – they are rarely considered a suitable substitute for traditional essay-type questions. This is especially true for final year honours degree examinations.

This paper introduces a new form of assessment: the Permutational Multiple-Choice Question (PMCQ). Results of trials in final year degree examinations indicate that these questions are as good as essay-type questions at discriminating among candidates. They also offer many benefits:

- **consistency and reliability in marking**
  - **reduced need for cross checking among assessment teams, or between franchised institutions,**
  - **objective and reproducible results;**
- **efficiency in marking**
  - **quicker to mark,**
  - **can be automated;**
- **broad coverage of syllabus.**

Unlike traditional multiple-choice questions, PMCQs are not susceptible to candidates guessing the correct answer. Candidates who guessed 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).

ITiCSE '98 Dublin, Ireland

1.1 Keywords

Student assessment, multiple-choice questions

2. INTRODUCTION

The advantages offered by multiple-choice questions (MCQs) – and other non-subjective assessments – are useful to educational institutions under pressure to reduce unit costs. Compared with traditional essay-type (or free response) questions, MCQs offer speedier marking, reduced need for cross-marking, greater reliability (fairness), greater coverage of the syllabus, and the possibility to automate the marking process.

However, there are a number of commonly raised objections to the use of MCQs. In particular, many claim that:

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

Although MCQs are sometimes used in the earlier stages of degree courses, they are rarely used during the final year of honours-level studies. This paper describes a new approach – called the Permutational Multiple-Choice Question or PMCQs – that seeks to solve the above problems. The paper also reports on trials with final year degree examinations.

3. PERMUTATIONAL MULTIPLE-CHOICE QUESTIONS (PMCQs)

A normal MCQ has a stem and perhaps four putative answers (one key and three distracters). A PMCQ typically has a two-part stem, and six putative answers: two of which are keys and four are distracters (see Figure 1). To answer the question correctly, the candidate must match up each stem with the appropriate key.

1. Which of these *best* describes the contents of a data dictionary of: a CASE tool **b** a data warehouse **f**	
a.	Table names, column names and formats, indexes, views
b.	Data stores, processes, entities, relationships
c.	Code look-up tables, foreign keys
d.	Dimensional descriptions, aggregation rules, drill-down procedures
e.	Program and file names, access modes, security
f.	Tables, column names and formats, data sources, data importing procedures, drill-down procedures

Figure 1: An example of a PMCQ

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 putative answers must be *feasibly correct* for both parts of the stem. In the example shown in Figure 1, all of the putative answers are data dictionaries of one sort or another, most of which were discussed in lectures. The candidate must identify which best describes the data dictionary of a CASE tool and a data warehouse from his/her understanding of what they are used for.

The correct answers (b and f) have been entered in this example.

In this example, the candidate must write the answers on the question sheet. If separate answer sheets are preferred, the candidate could simply write down "1. CASE tool b Data warehouse f".

We shall now consider how PMCQs address the three main objections to MCQs mentioned above.

3.1 Guessing

A PMCQ with a two-part stem, two keys and four distracters provides 30 valid permutations of answers (6 x 5). Using random guessing, candidates can expect to average only 3% or 4% in such an examination. Indeed the chance of someone obtaining a pass mark this way is negligible (only one in 4500 guessers might gain 40% in an exam of ten PMCQs).

This compares favourably with normal MCQs, where candidates who guess randomly might expect to average 25%. Worse still, analysis shows that if everyone sitting an exam of ten MCQs guessed the answers, perhaps one in 5 might nevertheless gain a pass mark. Years ago, Ebel showed that normal techniques to compensate for guessing do not adequately solve this problem [1].

A potentially controversial aspect of PMCQs is that no marks should be awarded for part-correct answers. For example, if a candidate correctly identified the CASE tool data dictionary but not the data warehouse, the candidate would get zero for that question. If this "all-or-nothing" rule is not followed, the exam becomes more like a standard MCQ test. Random guessing might produce marks averaging 17% (better than normal MCQs, but not ideal).

3.2 Trivial Recognition of Facts

Over the years, much research has been conducted into devising MCQs that assess high-level thinking (for example, [2], [3] and [4]). What PMCQs offer is an opportunity to make candidates distinguish between similar concepts, and think of consequential effects of these differences. Consider the PMCQ in Figure 2.

2. Which of these *most closely* describes the *identifying characteristic* of: black-box software testing **d** white-box software testing **a**	
a.	Tests are designed with full knowledge of what is in the source code
b.	Tests are run unattended
c.	A program is re-tested after a change
d.	The application is checked to ensure the users' needs are satisfied
e.	The test results are checked for correctness
f.	The entire application is retested after one part of it is changed

Figure 2: PMCQs can seek consideration of consequences

Many candidates might identify that white-box software tests are designed with full knowledge of the source code (alternative a). They would probably be aware that black-box software tests are designed without knowing what is in the source code. However, none of the alternatives present this answer. The candidate is forced to think further. How do we design tests if we don't know what is in the source

code? Black-box tests must be designed to ensure the users' needs are satisfied . (alternative d).[1]

3.3 Construction Versus Choice

The debate on whether multiple-choice assessment can be an acceptable substitute for questions that ask candidates to construct an answer has been running a long time [5]. There are valid arguments in favour of assessments that require candidates to construct answers. On the other hand, well-formed distracters can present candidates with options they might not have otherwise considered. This can challenge woolly thinking in a way that open-ended questions do not.

The authors know of no constructed response assessments (that require critical judgement) that can be marked as quickly, efficiently and objectively as multiple-choice questions.

4. TRIALS

A trial of PMCQs in a final year undergraduate module was held in 1996, and extended in 1997. Candidates answered some mandatory PMCQs and also essay-type questions. Statistical analyses of the results are given here.

4.1 The 1996 Trial

Analysis of the 1996 examination papers proved disappointing. The correlation between the marks for the PMCQ section of the paper and the overall marks was much lower than we had hoped.

Five PMCQs were included in the examination paper (worth only 10% of the overall marks). No attempt was made to make the PMCQs and the essay-type questions "equivalent" in any way; they assessed different topics. Table 1 summarises the analysis of the 44 examination scripts.

The mean of 54% for Section D indicates the questions were of a reasonable facility. Adjusting exam questions to alter the facility is relatively easy; what is more important is ensuring the stronger candidates get higher marks and the weaker student get lower marks. This first experiment the PMCQs did not adequately discriminate between the stronger and the weaker candidates. A correlation of only 0.598 was disappointingly low. This means some weaker candidates did well with the PMCQs, and a few stronger candidates did slightly worse.

Statistically speaking, it would be very difficult to gain a high correlation between the strongly "step-wise" function of five multiple-choice questions and the much smoother function of the essay-type questions. A detailed examination of the scripts gave us grounds for encouragement that PMCQs *could* be made to work. We analysed the PMCQs with a view to improving things the following year.

N=44		Mean (facility)	Correlation (discrimination)
Section A	Essay	34%	.806
Section B	Essay	36%	.708
Section C	Essay	48%	.808
Section D	PMCQ (10% of marks)	54%	.598

Table 1: Results of 1996 trial

4.2 The 1997 Trial

In the second trial, fifteen PMCQs were included in the examination paper. These attracted 30% of the overall marks. Again, no attempt was made to make the PMCQs and the essay-type questions "equivalent" in any way. Table 2 summarises the analysis of the 54 examination scripts.

N=54		Mean (facility)	Correlation (discrimination)
Section A	Essay	34%	.806
Section B	Essay	54%	.807
Section C	PMCQ (30% of marks)	41%	.808

Table 2 : Results of 1997 trial

Differences between the three correlation coefficients are statistically insignificant. Although it would have been preferable if every section had correlated closer to unity, what we can say is that all three section marks are equally good as a predictor of the total mark. We see that every section of the exam paper discriminated equally well between the stronger and weaker candidates.

Marking the PMCQs was completed quickly, and a second check of the marks revealed very few errors.

4.3 Ignoring the All-or-Nothing Rule

In Section 2.1 above we referred to the fact that no marks should be awarded for part-correct answers to PMCQs. To establish whether this rule is very important in practice, the exam papers were re-marked ignoring this rule; candidates

[1] Since *all* software test results must be checked, alternative e cannot be the *identifying characteristic* of either black-box or white-box testing.

were given credit for answers that were partly correct. This marking scheme resulted in a slightly lower correlation: only 0.74 (instead of 0.81).

This result may indicate that weaker candidates are guessing some answers. We conclude that ignoring the all-or-nothing rule increases the effects of guessing, and thus harms the identification of stronger and weaker candidates.

5. ADVANTAGES AND DISADVANTAGES OF PMCQs

We do not present Permutational Multiple-Choice Questions as a "better" form of assessment, since there is no commonly accepted notion of what makes one assessment better than another. What we can say is that PMCQs have strengths when compared with other forms of assessment, and these have to be balanced against the known weaknesses of non-subjective assessment.

5.1 Comparing PMCQs with Essay-Type Questions

5.1.1 Advantages

PMCQs offer a reliable marking process. They are not prone to subjective biases of assessors, consequently examination scripts do not need to be checked for even-handedness. Since PMCQs demand little writing, examiners have few or no cues on race, gender or penmanship [2]. Further, PMCQs assess candidates' knowledge and understanding, not their writing speed or stamina.

The validity of an examination using PMCQs is high, because a large part of the syllabus can be assessed. Typically, an examination composed of only essay-type questions can assess only part of the syllabus. Such questions can be "broad and shallow," or "deep and narrow." That is, they can cover a lot of topics without being too taxing on any, or very detailed on certain topics (but ignore the rest).

PMCQs can be marked quickly. Typically, marking a script of 10 or 20 PMCQs might take about a minute or so. Because the marks are reliable, no further time is spent re-marking scripts. This advantage is valuable with large classes, and even more so where teams of assessors are needed, since cross-marking is unnecessary.

Statistical analysis is easier where candidates can make only limited responses. Standard pointers to remedial action may then be used.

Because PMCQs permit a large numbers of questions to be answered in a short time, it becomes impractical for candidates to memorise questions and answers.

PMCQs offer the possibility of automated marking. They are also suitable for computer based learning, including the World Wide Web and other multimedia systems (e.g. [6]).

5.1.2 Disadvantages

As has already been discussed, candidates who may not be able to suggest the correct answer perhaps will recognise it when prompted. On the other hand, it may also be true that a convincing set of distracters may make candidates question something they thought they understood, but now aren't so sure [7] [8].

The cost of preparation of PMCQs is higher than essay-type questions. This may be recouped during marking, and through reuse of test items.

Tailoring remedial action for weaker students can be more difficult with MCQs (see [2]). Also, answers to essay-type questions allow us to trace the candidate's solution process where they exhibit only partial understanding [3]; MCQs assess only the final result. The all-or-nothing rule tends to exacerbate this problem.

Candidates may treat PMCQs as trivial due simply to social conditioning, and may not give them the attention they deserve.

5.2 Comparing PMCQs with Normal MCQs

Many of the above advantages and disadvantages are shared with traditional MCQs. However, PMCQs exhibit these additional characteristics.

5.2.1 Advantages

PMCQs are more suitable for assessing high-level thinking processes, since they ask candidates to compare and contrast similar concepts, and to consider consequences of decisions.

PMCQs do not require any guessing correction. The guessing correction techniques used with traditional MCQs can produce negative marks, and they are not as effective at preventing candidates from obtaining a pass mark through guessing. PMCQs never produce negative marks, and the likelihood of obtaining a pass mark through guessing is negligible.

Devising distracters for PMCQs should be easier, since the examiner has to devise only two distracters per key (compared with three distracters per key in most normal MCQs).

5.2.2 Disadvantages

Devising suitable questions may be difficult. Identifying a pair of similar concepts that may tested in a single question can sometimes be elusive.

6. CONCLUSIONS

6.1 Trial Results

Our statistical analysis shows evidence that PMCQs are a valid substitute for essay-type questions; they are as good as essay-type questions at predicting the total mark. This result

mirrors many studies made in the past. For example, Bennett, Rock and Wang made an extensive analysis of answers to MCQs and essay-type questions. They concluded that "the evidence presented offers little support for the stereotype of multiple choice and free-response (essay) formats as measuring substantially different constructs" [3].

This paper does not suggest that PMCQs should completely replace traditional essay-type questions. For instance PMCQs do not adequately assess the creative abilities of the candidates, they may lead the candidates towards an answer which they would not have provided unprompted, and assessing ambiguous concepts can be difficult using PMCQs. Further, if a module were assessed solely through PMCQs, it might have a detrimental effect on student learning strategies. However, used in conjunction with traditional questions, PMCQs can provide a reliable way of assessing students while reducing the workload of marking large numbers of examination scripts.

Our conclusions must be delimited as follows. They are delimited to the study of Computer Science. Our trials establish that PMCQs assess high-level thinking only through comparing results with essay-type questions, not through psychological means. Finally, it is possible that the statistical factors are pointing to another variable not yet identified.

6.2 The Validity of Improving Efficiency

Some might question whether improving efficiency in assessment is a justifiable goal. We argue that efficient use of funds is important, even in the assessment process. Furthermore, improvements in efficiency can lead to improvements in the *quality* of assessment. For example, assessors under time pressures are prone to make mistakes. Reducing time pressures will allow assessors more time for double-checking, and so on. Also, non-subjective approaches eliminate deliberate and accidental bias.

6.3 Future Plans

We plan to roll out our trials to larger student groups. We also intend to establish the equivalence of PMCQs and essay-type questions in other ways, such as through using psychological techniques. Finally, we also need to identify what are the attributes of PMCQ items that discriminate well, in order to improve further the assessment process.

7. ACKNOWLEDGEMENTS

We thank our colleagues Bob Large and Mike Reddy for helpful comments on earlier drafts of this paper.

8. REFERENCES

[1] Ebel, R L. Measuring educational achievement. Prentice Hall, Englewood Cliffs NJ, 1965

[2] Haladyna, T M. Developing and validating multiple-choice test items. Lawrence Erlbaum Associates, Hillsdale NJ, 1994

[3] Bennett, R E; Rock, D A & Wang, W C. Equivalence of free-response and multiple choice items. Journal of Educational Measurement, Vol.28, No.1, pp77-92, 1991

[4] Bridgeman, B & Rock, D A. Relationships among multiple-choice and open-ended analytical questions. Journal of Educational Measurement, Vol.30, No.4, pp 313-329, 1993

[5] Bennett, R E & Ward, W C (Eds). Construction versus choice in cognitive measurement. Lawrence Erlbaum Associates, Hillsdale NJ, 1993

[6] Large, A; Beheshti, J; Breuleux, A; & Renaud, A. Effect of animation in enhancing descriptive and procedural texts in a multimedia learning environment. Journal American Society for Information Science, Vol.47, No.6, pp.437-448, 1996

[7] Fenderson, B A; Damjanov, I; Robeson, M R; Veloski, J; & Rubin, E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Human Pathology, Vol.28, No.5, pp.526-532, 1997

[8] Schuwerth, L W T; Vandervleuten, C P M; & Donkers H H L M. A closer look at cueing effects in multiple-choice questions. Medical Education, 1996, Vol.30, No.1, pp.44-49, 1996