Int. J. Engng Ed. Vol. 22, No. 6, pp. 1281–1286, 2006
Printed in Great Britain.

0949-149X/91 $3.00+0.00

How to Design and Interpret a Multiple-Choice-Question Test: A Probabilistic Approach*

Y. ZHAO
Department of Engineering, The University of Liverpool, Brownlow Hill, Liverpool L69 3GH, UK.
E-mail: Y.Y.Zhao@liv.ac.uk

Multiple choice question (MCQ) tests involve an element of guesswork, which affects the reliability and interpretation of test scores. This article studies the probability of obtaining a certain score by pure guesswork and introduces a conversion scheme which converts raw test scores into standard percentage marks. The probabilistic analysis shows that the optimum number of choices of answers for MCQ questions is four, and for a four-choice question test, increasing from 8 questions to 18 and 48 questions reduces the probability of obtaining a converted mark above 40 by pure guesswork from about 5% to below 1% and 0.01%, respectively.

Keywords: MCQ; conversion scheme; probability; score; mark.

INTRODUCTION

MULTIPLE CHOICE QUESTION (MCQ) tests are a widely used assessment methodology. They are objective, easy to mark and quick to obtain results. They are most suitable for the assessment of knowledge, analytical ability, language proficiency and numerical skills involving a large number of candidates. Properly designed MCQ tests are an invaluable assessment tool for knowledge or fact-based subjects.

Any MCQ test invites some guesswork and therefore has an element of uncertainty. The uncertainty is manifested in two aspects. One is the reliability of test scores. A well designed test should have a narrow range of fluctuation in the test scores as a result of random guesses. Another is the reliability of interpretation of test scores. The test scores need to be converted into marks that are a true measure of the performance of the students. In other words, the proportions of scores accrued from guesswork should be deducted. These two aspects of uncertainty must be taken into account to make MCQ tests a scientifically sound assessment methodology.

It is well recognised that raw scores of MCQ tests should not be used directly [1], unless the only purpose is to select a certain number of candidates according to their relative competence. To gauge the true level of knowledge of the candidates in a subject through MCQ tests, it is necessary to convert the raw scores into more meaningful marks or grades. In well established tests involving a large number of participants, such as TOEFL, complex scaling schemes are often used. In these tests, there normally exists a large databank of questions and answers that have been tested to be reliable. The scaling schemes are developed on the basis of extensive research on the statistics of the past tests [2]. In most other cases where the number of candidates is relatively small and/or no historical data is available for comparison, no universally applicable scaling schemes are readily available. The test setters often adopt arbitrary scaling schemes based on their experiences in their specialised subjects. Very recently, Zhao developed algorithms for converting MCQ raw scores to conventional percentage marks based on probability theory [3]. The algorithms are independent of class size and historical data and can be easily implemented by using a conversion table. The converted marks are compatible with the standard marking scheme which is regarded as a true measure of the students' knowledge and competence.

The basis for the conversion algorithms developed by Zhao [3], however, is that the MCQ tests consist of sufficiently large numbers of questions. Otherwise, the part played by guesswork may be too great to generate reliable test scores and any conversions of these test scores are inherently erroneous. Considering an extreme case where a test is composed of a single true-false question, a student can make a random choice and gets either a correct or wrong answer. There is a 50% chance that the answer is correct and the student obtains a full score, and thus a full mark, for the test. This is obviously unacceptable. To reduce the effect of the part played by guesswork, the number of questions must be increased. Intuitively, the more questions the better accuracy. In practice, however, there is

* Accepted 4 November 2005.

an upper limit of number of questions that a MCQ test can accommodate. An issue arises as regards how many questions are necessary to reduce the effect of guesswork to an acceptable level.

This article is to address the uncertainty issues as a whole from a probabilistic approach. Firstly, the conversion scheme developed by Zhao [3] is summarised and explained. Secondly, the probabilities of a student obtaining a certain score by pure guesswork in MCQ tests with different numbers of choices and different numbers of questions are analysed and the number of choices and the number of questions are recommended. Thirdly, the outcome of an application of the conversion scheme to a module is evaluated. The article is not intended to address the subject-related issues, such as whether MCQ tests are suitable for the assessment of certain knowledge and skills and whether the answers to a question are appropriate choices or not. Instead, the analysis is largely on the structure of MCQ tests and is based on the assumption that all the individual questions are properly designed, i.e. all the provided choices of answers look equally feasible to a layman. In this paper, the raw percentage scores of MCQ tests prior to conversion and the percentage marks after conversion are simply termed scores and marks, respectively.

CONVERSION SCHEME

The conversion scheme developed by Zhao [3] is based on the analysis of the scores a student is likely to obtain and the marks the student should be awarded for different types of questions classified according to the student's knowledge of the answers. An answer is either a firm answer, which is definitely known to the student to be either correct or wrong, or an uncertain answer. All the questions can be classified according to whether the correct answer is a firm answer and, if not, the number of firm answers among the wrong answers. Taking four-choice questions with one correct answer and three wrong answers as an example, there are five types of questions altogether:

A. The correct answer is a firm answer.
B. There are three firm answers which are all wrong answers.
C. There are two firm answers which are wrong answers.
D. There is one firm answer which is a wrong answer.
E. All four answers are uncertain answers.

For a Type A question, the student can choose the correct answer without involving any guessing. The student gets a full score and deserves a full mark. For all the other types of questions, the student either has to resort to deductive reasoning to find the correct answer or eliminates the firm wrong answers and chooses an answer from the rest by guesswork. In most of these cases, the score that the student probably gets is different from the mark that the student deserves. Taking a Type D question as an example, the student knows that one answer is wrong but does not know which of the three answers left is correct. The student is likely to guess one from the three uncertain answers and the chance of the correct answer being chosen is 1/3. Given a large number of Type D questions, the student has the highest probability of obtaining 1/3 of the full score, i.e. a score of 33 expressed in percentage. However, the student does not deserve a percentage mark of 33. The student should be awarded a mark of 25 because the student only knows 1/4 of the answers. Similarly, each type of question is associated with a unique score the student is likely to get and a unique mark the student should be awarded.

A four-choice MCQ test with a large number of questions is very likely to contain all these five types of questions from a student's point of view. The frequency of appearance of each type of questions is dependent upon the student's knowledge and is a function of the fraction of firm answers among all the answers of the questions in the test, designated as f. Table 1 lists the probable score, deserved mark and probable frequency of appearance of each type of questions for a four-choice MCQ test. Similar tables can be constructed for two-, three- and five-choice MCQ tests. Summing up the frequency-weighted scores and marks of all the five types of questions gives a total score and a total mark for a test. For a MCQ test with a constant number of choices of answers for each question, there is a corresponding relationship between the total score and total mark. In other words, any percentage score can be converted into a percentage mark. Table 2 is the conversion table for MCQ tests with questions of two, three, four or five choices of answers.

It is worth pointing out that the conversion scheme has a cut-off score for each type of MCQ tests below which no marks are awarded. The cut-off scores for two-, three-, four- and five-choice MCQ tests are 50, 33, 25 and 20, respectively. Not surprisingly, these cut-off scores correspond to the scores that a student most probably obtains by pure guesswork.

NUMBER OF CHOICES AND NUMBER OF QUESTIONS

Having a degree of uncertainty is an inherent weakness of MCQ tests. It is impossible to completely eliminate the effect of guessing on the scores of MCQ tests. Even with an extremely large number of questions, there is still a chance, although very slim, for a student to obtain a good score by pure guesswork. However, it is possible to reduce the student's chance of obtaining a good score by guesswork to a predetermined level. From a probabilistic point of view, the more questions a test has, the lower the effect the guesswork has on the

Table 1. Probable score, deserved mark and probable frequency of appearance for the five types of four-choice questions [3]

Type	Number of firm answers		Score	Mark	Frequency of appearance
	Correct answer	Wrong answers			
A	1	Any	1	1	f
B	0	3	1	3/4	$f^3(1-f)$
C	0	2	1/2	1/2	$3f^2(1-f)^2$
D	0	1	1/3	1/4	$3f(1-f)^3$
E	0	0	1/4	0	$(1-f)^4$

test scores. In practice, a compromise in the number of questions is needed to achieve a right balance between accuracy and efficiency. On one hand, the number should be small enough for the students to be able to manage in a fixed period of time appropriate for the purpose of the assessment. On the other hand, it should be large enough for the probability of obtaining a certain score by pure guesswork to be below an acceptable level. For example, the criterion may be set to be that the probability of obtaining a score above the pass mark is below 5%.

Let us now consider the probability of obtaining a certain score by pure guesswork in a MCQ test consisting of N questions, each of which has m choices of answers. In each question, only one answer is correct and the other $(m-1)$ answers are incorrect. The probability of picking out a correct answer for one question is therefore $1/m$ and the probability of picking out a wrong answer is $(1-1/m)$. The probability of selecting correct answers for a specific set of n questions and selecting incorrect answers for the other $(N-n)$ questions is [4]:

$$p = \left(\frac{1}{m}\right)^n \left(1-\frac{1}{m}\right)^{N-n} \tag{1}$$

The number of possible ways of selecting n questions from the total N without regard to their order of arrangement is the combination of a set of N mutually distinguishable objects n at a time, and can be expressed by [4]:

$$C = \frac{N!}{n!(N-n)!} \tag{2}$$

The probability of selecting correct answers for any n questions and selecting incorrect answers for the other $(N-n)$ questions by pure guesswork is therefore:

$$P_n = pC = \left(\frac{1}{m}\right)^n \left(1-\frac{1}{m}\right)^{N-n} \frac{N!}{n!(N-n)!} \tag{3}$$

Table 2. Conversion table for MCQ tests with questions of two, three, four or five choices of answers, corresponding to columns indicated by (2), (3), (4) and (5) [3]

Score	Mark				Score	Mark				Score	Mark			
	(2)	(3)	(4)	(5)		(2)	(3)	(4)	(5)		(2)	(3)	(4)	(5)
≤20	0	0	0	0	47	0	22	35	43	74	38	60	69	75
21	0	0	0	2	48	0	23	36	44	75	40	61	71	76
22	0	0	0	4	49	0	25	38	46	76	41	62	72	77
23	0	0	0	5	50	0	26	39	47	77	43	64	73	78
24	0	0	0	7	51	2	28	41	48	78	45	65	74	79
25	0	0	0	9	52	3	29	42	49	79	47	66	75	80
26	0	0	2	11	53	5	31	43	51	80	48	68	76	81
27	0	0	3	12	54	6	32	45	52	81	50	69	77	82
28	0	0	5	14	55	8	33	46	53	82	52	70	78	83
29	0	0	7	16	56	9	35	47	55	83	54	72	79	84
30	0	0	9	17	57	11	36	49	56	84	56	73	80	84
31	0	0	10	19	58	12	38	50	57	85	58	74	81	85
32	0	0	12	20	59	14	39	51	58	86	60	76	83	86
33	0	0	14	22	60	15	41	53	59	87	62	77	84	87
34	0	1	15	24	61	17	42	54	61	88	64	79	85	88
35	0	3	17	25	62	18	43	55	62	89	66	80	86	89
36	0	4	18	27	63	20	45	56	63	90	68	81	87	90
37	0	6	20	28	64	22	46	58	64	91	70	83	88	91
38	0	8	21	30	65	23	48	59	65	92	72	84	89	92
39	0	9	23	31	66	25	49	60	66	93	74	86	90	92
40	0	11	25	33	67	26	50	61	67	94	77	87	91	93
41	0	12	26	34	68	28	52	62	69	95	79	89	92	94
42	0	14	28	36	69	30	53	64	70	96	82	90	93	95
43	0	16	29	37	70	31	54	65	71	97	85	92	95	96
44	0	17	31	39	71	33	56	66	72	98	88	94	96	97
45	0	19	32	40	72	35	57	67	73	99	92	96	97	98
46	0	20	34	41	73	36	58	68	74	100	100	100	100	100

Table 3. Probabilities of obtaining a mark equal to or above 40 by pure guesswork as a function of number of questions N for MCQ tests with questions of two, three, four or five choices of answers, corresponding to columns indicated by Equations (2), (3), (4) and (5)

	Probability (%)					Probability (%)			
N	(2)	(3)	(4)	(5)	N	(2)	(3)	(4)	(5)
1	50	33.33	25	20	26	0.47	0.30	0.15	0.23
2	25	11.11	6.25	36.00	27	0.30	0.15	0.25	0.35
3	12.5	25.93	15.63	10.40	28	0.63	0.27	0.11	0.15
4	31.25	11.11	5.08	18.08	29	0.41	0.14	0.18	0.22
5	18.75	20.99	10.35	5.79	30	0.26	0.25	0.08	0.09
6	10.94	10.01	3.76	9.89	31	0.17	0.13	0.13	0.13
7	6.25	4.53	7.06	3.33	32	0.35	0.07	0.06	0.06
8	14.45	8.79	2.73	5.63	33	0.23	0.12	0.10	0.08
9	8.98	4.24	4.89	8.56	34	0.15	0.06	0.04	0.04
10	5.47	7.66	1.97	3.28	35	0.09	0.11	0.07	0.05
11	3.27	3.86	3.43	5.04	36	0.20	0.06	0.03	0.08
12	7.30	1.88	1.43	1.94	37	0.13	0.10	0.05	0.03
13	4.61	3.47	2.43	3.00	38	0.08	0.05	0.02	0.05
14	2.87	1.74	1.03	1.16	39	0.05	0.03	0.04	0.02
15	1.76	3.08	1.73	1.81	40	0.11	0.05	0.02	0.03
16	3.84	1.59	0.75	0.70	41	0.07	0.03	0.03	0.01
17	2.45	0.80	1.24	1.09	42	0.05	0.04	0.01	0.02
18	1.54	1.44	0.54	1.63	43	0.10	0.02	0.02	0.01
19	0.96	0.74	0.89	0.67	44	0.06	0.01	0.01	0.01
20	2.07	1.30	0.39	1.00	45	0.04	0.02	0.01	0.02
21	1.33	0.68	0.64	0.41	46	0.03	0.01	0.01	0.01
22	0.85	0.35	0.29	0.61	47	0.05	0.02	0.01	0.01
23	0.53	0.62	0.46	0.25	48	0.04	0.01	<0.01	<0.01
24	1.13	0.32	0.21	0.38	49	0.02	0.01	<0.01	<0.01
25	0.73	0.56	0.34	0.15	50	0.02	0.01	<0.01	<0.01

The probability of obtaining a percentage score equal to or higher than $100 \times n/N$ is the sum of the probabilities of selecting correct answers for n or more questions from the total N by pure guesswork and can be calculated by:

$$P_{\geq n} = \sum_{i=n}^{N} P_i = \sum_{i=n}^{N} \left(\frac{1}{m}\right)^i \left(1 - \frac{1}{m}\right)^{N-i} \frac{N!}{i!(N-i)!} \tag{4}$$

From a practical point of view, the test setters are more concerned about the probability of obtaining a mark above a critical value by pure guesswork. In this article, the critical values of 40 and 60, which are the common pass marks in most subjects including medicine and engineering, are considered. For two-, three-, four- and five-choice MCQ tests, a pass mark of 40 corresponds to scores of 75, 60, 51 and 45, respectively, as shown in Table 2. Table 3 lists the probabilities of obtaining a mark equal to or above 40 by pure guesswork in two-, three-, four- and five-choice MCQ tests as a function of the number of questions. The probabilities are calculated by Equation (4), choosing an integer n which makes $(100 \times n/N)$ equal to or above 75, 60, 51 and 45 with respect to two, three, four and five choices. Similarly, the probabilities of obtaining a mark equal to or above 60 by pure guesswork as a function of the number

Table 4. Probabilities of obtaining a mark equal to or above 60 by pure guesswork as a function of number of questions N for MCQ tests with questions of two, three, four or five choices of answers, corresponding to columns indicated by Equations (2), (3), (4) and (5)

	Probability (%)					Probability (%)			
N	(2)	(3)	(4)	(5)	N	(2)	(3)	(4)	(5)
1	50	33.33	25	20	14	0.65	0.07	0.03	0.04
2	25	11.11	6.25	4.00	15	0.37	0.03	0.08	0.01
3	12.50	3.70	15.63	10.40	16	0.21	0.08	0.03	0.02
4	6.25	11.11	5.08	2.72	17	0.12	0.03	0.01	0.01
5	3.13	4.53	1.56	0.67	18	0.07	0.01	0.02	0.02
6	1.56	1.78	3.76	1.70	19	0.04	0.04	<0.01	<0.01
7	6.25	0.69	1.29	0.47	20	0.02	0.02	<0.01	<0.01
8	3.52	1.97	0.42	1.04	21	0.07	<0.01	<0.01	<0.01
9	1.95	0.83	1.00	0.31	22	0.04	<0.01	<0.01	<0.01
10	1.07	0.34	0.35	0.09	23	0.02	<0.01	<0.01	<0.01
11	0.59	0.14	0.12	0.20	24	0.01	<0.01	<0.01	<0.01
12	0.32	0.39	0.28	0.06	25	<0.01	<0.01	<0.01	<0.01
13	0.17	0.16	0.10	0.12					

of questions can also be calculated by Equation (4), and they are listed in Table 4.

For any type of MCQ test, whether with two, three, four or five choices, the probability of obtaining a pass mark by pure guesswork decreases with increasing number of questions as a general trend. However, more questions do not always make it more difficult to obtain a pass mark by guessing, especially when the number of questions is small. Take a MCQ test with two-choice questions as an example. It is easier to pass the test with 4 questions than 3 questions. This discrepancy is due to the discrete distribution of scores. In a two-choice test, a percentage score of 75 is needed to pass. For 3 questions, all 3 questions, i.e. a score of 100, must be guessed correctly in order to pass. For 4 questions, only 3 out of 4, i.e. a score of 75, are needed for a pass. The probability of guessing 3 correct answers out of 4 questions is higher than that of guessing 3 correct answers out of 3 questions.

The number of choices of answers in MCQ questions has a profound effect on the effectiveness of MCQ tests. From the analyses of the conversion scheme and of the chances of guessing success, four choices are demonstrated to be the optimum number. Fewer choices not only tend to need more questions to lower the chances of obtaining a pass mark by pure guesswork but also have a narrow range of pass scores. For a pass mark of 40, the range of scores that a student needs to obtain for a pass in a MCQ test is 75–100 for two-choice questions and 60–100 for three-choice questions. If the pass mark is set as 60, the ranges of scores are further reduced to 86–100 and 74–100 for two- and three-choice questions, respectively. These ranges are too narrow to differentiate the performance of the students. A very large number of questions would be needed to achieve a properly spread distribution of marks. Although two- or three-choice question tests can be used in formative assessments, they are not recommended for summative assessments. It is normally not necessary to have more than four choices. From a probabilistic point of view, more choices do not offer significant benefits, as evidenced by comparing the probabilities of guessing success rates in four- and five-choice questions in Tables 3 and 4. For most subjects, it is also practically difficult to design questions with five or more choices of answers.

Let us focus on the four-choice questions and examine the strategy for selecting the number of questions further. For a short test, 8 or more questions would be sufficient to ensure the probability of obtaining a mark above 40 by pure guesswork to below 5% and that of obtaining a mark above 60 to below 1%. 18 or more questions would reduce the probability further to below 1% for obtaining a mark above 40 and below 0.2% for obtaining a mark above 60. 48 questions would reduce the probability to below 0.01% and the part played by guesswork becomes very small.

APPLICATION AND EVALUATION

A first-year module, Introduction to Computing, in the Department of Engineering, the University of Liverpool was conducted in 5 units, each of which was assessed by a one-hour MCQ test. Each test consisted of 20, 16 or 8 equally weighted four-choice questions. For example, the following question was designed to test whether the students had mastered the skills of using the Solver tool in Microsoft Excel to find user-defined best-fitting formulas to experimental data.

Question: The pressure of a gas in a closed cylinder varies with the volume of the gas, as a piston moves inside the cylinder. The experimental values of pressure, P, and volume, V, are listed in the following table.

V (m^3)	0.0072	0.0058	0.0044	0.0031	0.0023	0.0010	0.0006
P (bar)	0.9	1.3	1.1	1.5	2.2	2.5	3.6

Which of the following formulas best fits the experimental results?

A. $P = 0.078V^{-0.516}$
B. $P = 0.082V^{-0.510}$
C. $P = 0.086V^{-0.504}$
D. $P = 0.089V^{-.0497}$

The module mark for each student was obtained by converting the raw scores of the 5 tests into standard percentage marks followed by averaging these marks. The student's mark of this module has been compared with the student's overall average mark of all the modules taken in that semester. Figure 1 shows the relationship between the marks of this module and the semester average marks for all the 168 students who took this module. It is shown that the module marks correlate reasonably well to the semester averages, with

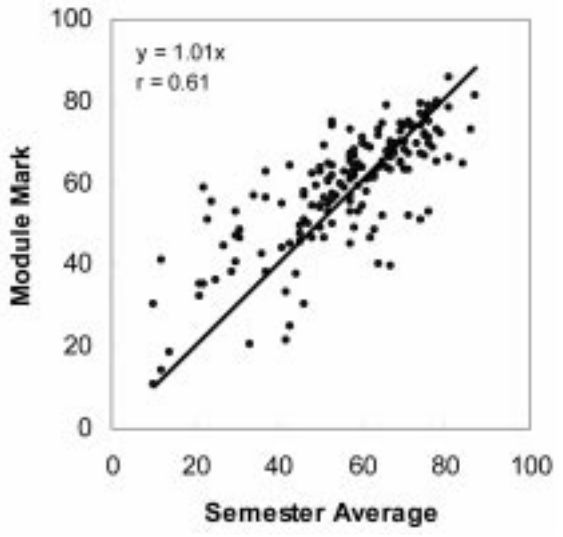

Fig. 1. Correlation between the marks of a MCQ module and the overall average marks of all the modules for a class of 168 students.

the correlation coefficient $r = 0.61$. The differences between the two are generally small. 19% of the students have a difference within 2 marks, 47% within 5 marks, 70% within 10 marks and 90% within 20 marks. The overall average of the class is 56.2 for this module and 58.9 for all the modules in the semester. It is demonstrated that the conversion scheme worked well and the MCQ tests were a reliable assessment method.

CONCLUSION

The format of a MCQ test has a dominant effect on the part played by guesswork on the test scores. The optimum number of choices of answers is four. Two- or three-choice question tests have a higher chance of obtaining a pass mark by pure guesswork and are difficult to differentiate students' performance due to a narrower range of pass scores. Five or more choices can be difficult to construct from a subject point of view and do not offer significant benefits in reducing the effect of guessing. A higher number of questions generally results in a lower probability of obtaining a pass mark by pure guesswork. For a four-choice question test, 8, 18 and 48 questions guarantee that the probabilities of obtaining a mark above 40 by pure guesswork are below 5%, 1% and 0.01%, respectively. Raw scores of MCQ tests can be converted into standard percentage marks compatible with the conventional marking scheme. The application of the conversion scheme to a module has resulted in a satisfactory outcome.

REFERENCES

1. J. C. McLachlan and S. C. Whiten, Marks, scores and grades: scaling and aggregating student assessment outcomes, *Medical Education,* **34**, 2000, pp. 788–797.
2. H. Wainer and X. Wang, *Using a New Statistical Model for Testlets to Score TOEFL*, TOEFL Technical Report TR-16, Education Testing Services, Princeton, NJ (2001) pp. 1–23.
3. Y. Y. Zhao, Algorithms for converting raw scores of multiple choice question tests to conventional percentage marks, *Int. J. Eng. Educ.,* **21**, 2005, pp. 1189–1194.
4. A. Jeffrey, *Mathematics for Engineers and Scientists,* Van Nostrand Reinhold Co. Ltd, Wokingham, UK (1985) pp. 741–755.

Yuyuan Zhao is a Senior Lecturer in the Department of Engineering at Liverpool University. He is currently teaching Computing and Metallurgical Thermodynamics. His research interests are in the area of manufacture, characterisation and modelling of particulate and porous materials. He graduated with a B.Eng. and a M.Sc. in Materials Engineering from Dalian University of Technology, China, in 1985 and 1988, respectively. He received his D.Phil. in Materials from Oxford in 1995. He was a Research Associate at Grenoble University, France, for a short period of time in 1995, and was a Research Fellow in the IRC in High Performance Materials at Birmingham University from 1995 to 1998. He is a winner of the 2005 Sir Alastair Pilkington Award for Teaching Excellence from the University of Liverpool.