Int. J. Engng Ed. Vol. 21, No. 6, pp. 1189–1194, 2005
Printed in Great Britain.

0949-149X/91 $3.00+0.00

Algorithms for Converting Raw Scores of Multiple-Choice Question Tests to Conventional Percentage Marks*

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

Multiple-choice question (MCQ) tests are not used widely in engineering subjects as a summative assessment methodology, largely because of the poor compatibility between the MCQ scores and conventional percentage marks. This paper develops algorithms for converting raw scores of MCQ tests to conventional marks based on probability theory. The algorithms are independent of class size and historical data and can be easily implemented in a spreadsheet programme by using a conversion table. The converted marks are compatible with the conventional marking scheme and can therefore be used standalone or as assessment units of a course. The algorithm for four-choice questions has been applied for a course with a satisfactory outcome. The issues concerned with the applications of the algorithms are discussed.

INTRODUCTION

WELL-DESIGNED multiple-choice question (MCQ) tests are an efficient means for the assessment of knowledge, analytical ability, language proficiency and numerical skills involving a large number of examinees. They are suitable for selection processes where the relative competence of the examinees in a large sample size is assessed. They are especially suitable for knowledge-based subjects which are well defined and do not change rapidly with time. MCQ tests are therefore used extensively in entrance examinations and aptitude tests at primary and secondary level but less so in higher education, except in some largely knowledge-based disciplines such as medicine.

MCQ tests are not used frequently in engineering subjects, due mainly to two reasons. First, the materials are not entirely suited for MCQ questions. Engineering courses generally involve acquiring knowledge, describing processes and phenomena, solving problems, developing experimental methodologies, creating designs, deriving mathematical formulas, analysing data and performing calculations. It is difficult, if not impossible, to devise MCQs for some of the components. It is often difficult to split the problems into small independent elements. Second, scaling algorithms to convert the scores accrued from MCQ tests into marks compatible with the normal marking scheme are not readily available. In engineering subjects, the conventional percentage marking scheme is usually adopted. It measures the 'absolute' competence of a student instead of the relative competence in a group. The students get full marks for questions answered correctly, zero marks for wrong answers and partial marks for partially correct answers. In the UK, the pass mark is set as 40 and the minimum marks for the first- and second-class degrees are usually set as 70 and 60. Setting an arbitrary or 'floating' pass mark for each course according to the overall performance of the students examined is neither customary nor desirable from the quality assurance point of view.

It is well recognised that raw scores accrued from MCQ tests should not be used directly [1]. This can be easily demonstrated by two examples. In a test with two-choice, true-or-false type questions, a student can get around half of the full marks by guessing the answers. In a test with four-choice questions, a student may know the answers for only 20% of the questions and guess the answers correctly for one quarter of the rest of the questions. In both cases, the student would pass the test.

To make full use of the benefits and minimise the drawbacks of MCQ tests, it is necessary to adopt a scientifically sound scaling scheme in those subjects where the conventional percentage marks are used across all the courses and averaged to give the overall marks for degree classifications. The scaling schemes currently available, such as that used in the TOEFL tests [2], often require a large size of participants, a large and well established databank of questions, and sometimes the statistics of past tests. These conditions are difficult to be met in engineering subjects. Most engineering courses have relatively small classes. Since many courses are highly specialised and the teaching staff rarely have many years to build up the courses, it is unrealistic to have a

* Accepted 21 April 2005.

large databank of questions. The contents of the courses often change quickly and the historical databank is not always useful. Even for the same contents, the cohorts of one year can be very different from another year's in terms of competence. Mapping their results to the same distribution or setting up the same pass rate may not be a good practice. Furthermore, the algorithms adopted are often too complex for the examiners to understand and use, particularly so when they are not regular setters of MCQ tests.

This paper develops algorithms for scaling the raw scores of MCQ tests based on probability theory. The aim is to convert the raw scores into standard marks compatible with the conventional percentage marking scheme and independent of class size and historical data, so that MCQ tests can be used either standalone or as assessment units of a course. The paper also provides a conversion table and illustrates the procedure to apply the algorithms using spreadsheet programmes. The issues concerned with the applications of the algorithms are discussed with the help of an example. In this paper, the raw scores of MCQ tests prior to conversion and the percentage marks after conversion are simply termed scores and marks, respectively.

ALGORITHMS

Let us first consider four-choice questions. Each question has one correct answer and three wrong answers. To a particular student, the answers can also be classified into firm answers and uncertain answers. If an answer is definitely known to the student to be either correct or wrong, then it is a firm answer. Otherwise, it is an uncertain answer. From the student's point of view, 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.

Let us now consider the scores that the student will probably obtain for these types of questions. For a Type A question, the student can choose the correct answer without hesitation and gets a full score. For a Type B question, the student can still pick out the correct answer by elimination and gets a full score because he knows all the wrong answers. For the other types of questions, the student cannot pick out the correct answers without resorting to some guesswork. For a Type C question, knowing that two choices are wrong answers, the student is likely to choose one from the two uncertain answers by guessing. The chance of the correct answer being chosen is 1/2. For a Type D question, the student knows one wrong answer only, so the chance of the correct answer being chosen is 1/3. For a Type E question, the student knows none of the four answers, so the chance of the correct answer being chosen is 1/4. Provided the number of questions is sufficiently large, the average scores of the Type C, D and E questions are equal to the corresponding chances of the correct answers being chosen. To sum up, the average scores (in percentage) that the student probably obtains from the five types of questions are:

$$s_A = 100\ s_B = 100\ s_C = 50\ s_D = 33.3\ s_E = 25 \quad (1)$$

where s is the average score for a type of question and the subscripts designate the corresponding types of question.

Let us then consider the marks that the student should be awarded for these five types of question. For a Type A question, the student knows the correct answer, so a full mark should be awarded. For Type B, C and D questions, the student knows three, two and one answers out of the four answers, so 3/4, 1/2 and 1/4 of the full marks should be awarded, respectively. For a Type E question, the student knows none of the answers, so no marks should be awarded. To sum up, the marks for the five types of questions, in percentage, should be:

$$m_A = 100\ m_B = 75\ m_C = 50\ m_D = 25\ m_E = 0 \quad (2)$$

where m is mark and the subscripts designate the corresponding types of question.

Both the total score and total mark that the student obtains in a test depends on the frequencies of appearance of the five types of questions. Assuming that the student knows a fraction, f ($0 \leq f \leq 1$), of the correct and wrong answers of the questions in the test, the probable frequency (in fraction) of Type A questions is f and the total frequency of the other four types of question is $1–f$. The relative frequencies of Type B, C, D and E questions are the probabilities of three, two, one or none firm answers being drawn from a databank of wrong answers for the questions. If this databank is reasonably large, then the problem becomes a simple probability exercise. In a case where three wrong answers are drawn randomly from a very large databank of wrong answers with f fraction of firm answers, the probabilities of three, two, one and none firm answers being drawn are f^3, $3f^2(1–f)$, $3f(1–f)^2$ and $(1–f)^3$, respectively [3]. The probable frequencies of appearance of the five types of question are therefore:

$$p_A = f\ p_B = f^3(1-f)\ p_C = 3f^2(1-f)^2$$
$$p_D = 3f(1-f)^3\ p_E = (1-f)^4 \quad (3)$$

The probable scores, deserved marks and probable frequencies of appearance of the five types of question are summarised in Table 1.

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

Type	Number of firm answers: Correct answer	Number of firm answers: Wrong answers	Score	Mark	Frequency of appearance
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	$(1-f)^4$

All the five types of question are expected to appear in a test, if a reasonably large number of questions are constructed by drawing one correct answer and three wrong answers randomly from the databank. Their probable frequencies of appearance would be close to those listed in Table 1. The total score that the student is likely to obtain in the test, S, is:

$$\begin{aligned} S &= s_A p_A + s_B p_B + s_C p_C + s_D p_D + s_E p_E \\ &= 25(1 + 4f - f^4) \end{aligned} \tag{4}$$

The total mark that the student should be awarded, M, is:

$$\begin{aligned} M &= m_A p_A + m_B p_B + m_C p_C + m_D p_D + m_E p_E \\ &= 25f(7 - 3f) \end{aligned} \tag{5}$$

The relationship between the test score and the test mark can therefore be established by combining equations (4) and (5) and by eliminating f. Although it is possible to obtain an explicit expression correlating the score and mark, the expression is too complex to be useful. It is more practical to determine the relationship by numerical means, such as by a spreadsheet programme or MATLAB.

Following the same approach as described above, a relationship can be established between the score and mark of any test that is composed of questions with an arbitrary number of choices of answers. The explicit or implicit conversion algorithms for MCQ tests with questions of two, three, four or five choices of answers are:

$$\begin{aligned} &M = S - \sqrt{50(100 - S)} && N = 2 \\ &\begin{cases} S = 33.3(1 + 3f - f^3) \\ M = 33.3f(5 - 2f) \end{cases} && N = 3 \\ &\begin{cases} S = 25(1 + 4f - f^4) \\ M = 25f(7 - 3f) \end{cases} && N = 4 \\ &\begin{cases} S = 20(1 + 5f - f^5) \\ M = 20f(9 - 4f) \end{cases} && N = 5 \end{aligned} \tag{6}$$

where N designates the number of choices of answers.

CONVERSION TABLE AND SPREADSHEET IMPLEMENTATION

A table for converting scores to marks for MCQ tests with questions of two, three, four or five choices of answers has been constructed using the algorithms developed above and is shown in Table 2. The scores of an MCQ test can easily be converted to marks using a spreadsheet programme such as Microsoft Excel. A typical conversion procedure can be demonstrated as follows. Let us take a test with four-choice questions as an example. Firstly, we enter the conversion table for $N = 4$ in columns A and B from row 2 to row 102, with the scores of 0 to 100 in cells A2 to A102 and their corresponding marks in cells B2 to B102. Secondly, we enter the student names in column C and their scores in column D, starting from row 2. Thirdly, we enter the following formula in cell E2: VLOOKUP(D2, $A$2:$B$102,2), which searches the value of D2 in column A and returns the corresponding value in the same row from column B. Finally, we select cell E2 and pull down the drag handle to fill in column E. The converted marks of the students are thus displayed in column E.

AN APPLICATION EXAMPLE

The conversion algorithm for four-choice questions has been applied to the course Introduction to Computing, in the Department of Engineering, the University of Liverpool. The course is to equip the first-year students in the department with the necessary computing skills for engineering applications. The course consists of four units, each of which is composed of one or two training sessions followed by a MCQ test. Tests 1, 2, 3 and 4 are on the skills of Microsoft Word, Excel basics, Excel optimisation and MATLAB, and have 20, 16, 8 and 20 questions, respectively. Each question is weighted equally, except in Tests 2 and 3, where a small number of questions weighted more than others. Test 1 is relatively easy, since almost all students have used Word extensively before taking the course. Test 2 is modest in difficulty, since most students have some prior knowledge of Excel. Test 3 is more difficult, because the techniques introduced are new to most students. Test 4 is also

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)

	Mark					Mark					Mark				
Score	(2)	(3)	(4)	(5)	Score	(2)	(3)	(4)	(5)	Score	(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

difficult, because MATLAB is a new package for almost all the students. Fig. 1 shows the histograms of the scores and marks of the four tests as well as the overall scores and marks for a class of about 170 students. Table 3 lists the basic statistics of the results. It should be noted that the overall marks are calculated by averaging the marks of the four tests.

The test results show clearly that the mean score and mark of the class decrease with increasing difficulty. The median score and mark of the class also decrease with increasing difficulty. The standard deviations of the scores and marks generally decrease with an increasing number of questions. With a small number of questions (e.g. Test 3), the scores and marks tend to be clustered. The standard deviation of the marks is always greater than that of the scores. This is because the marks spread more widely than the scores. Whereas the scores are normally between 25 and 100, the marks can be anywhere between 0 and 100. The distribution of the marks is comparable to that of the conventional tests.

SOME REMARKS ON THE APPLICATIONS OF THE ALGORITHMS

Validity of the algorithms

The conversion algorithms developed in this paper treat any scores below the cut-off scores as a zero mark. The cut-off scores are the scores a student might get by simply guessing one of the given answers for each question without knowing anything about the contents. They are 50, 33, 25 and 20 for questions with two, three, four and five choices of answers, respectively. Some people may argue that marks should be awarded progressively and any scores should be appreciated. Our experiences, however, have shown that the students who have made the efforts tend to get scores well above the cut-off scores. The cut-offs rarely pose problems. In fact, they reveal complete guesswork, which does not justify any marks.

Number of choices of answers

A student could easily obtain half of the full mark in a two-choice-question test and one-third in a three-choice-question test by guesswork. A true mark of 50 would probably result in scores of 81 and 67, respectively, as seen in Table 2. Whilst the true–false type of questions and three-choice questions are useful for formative purposes, they are not suitable for summative purposes.

Number of questions

Because of the probabilistic nature of MCQ tests, it is vital to have a sufficient number of questions in order to provide a reasonably accurate assessment of the students. With a small number of questions, it is difficult to achieve the balanced variety of difficulty necessary for distinguishing the students with different abilities. Twenty questions are found to be an acceptable assessment unit. Tests with ten questions or less (e.g. Test 3 in the example) are found to be less

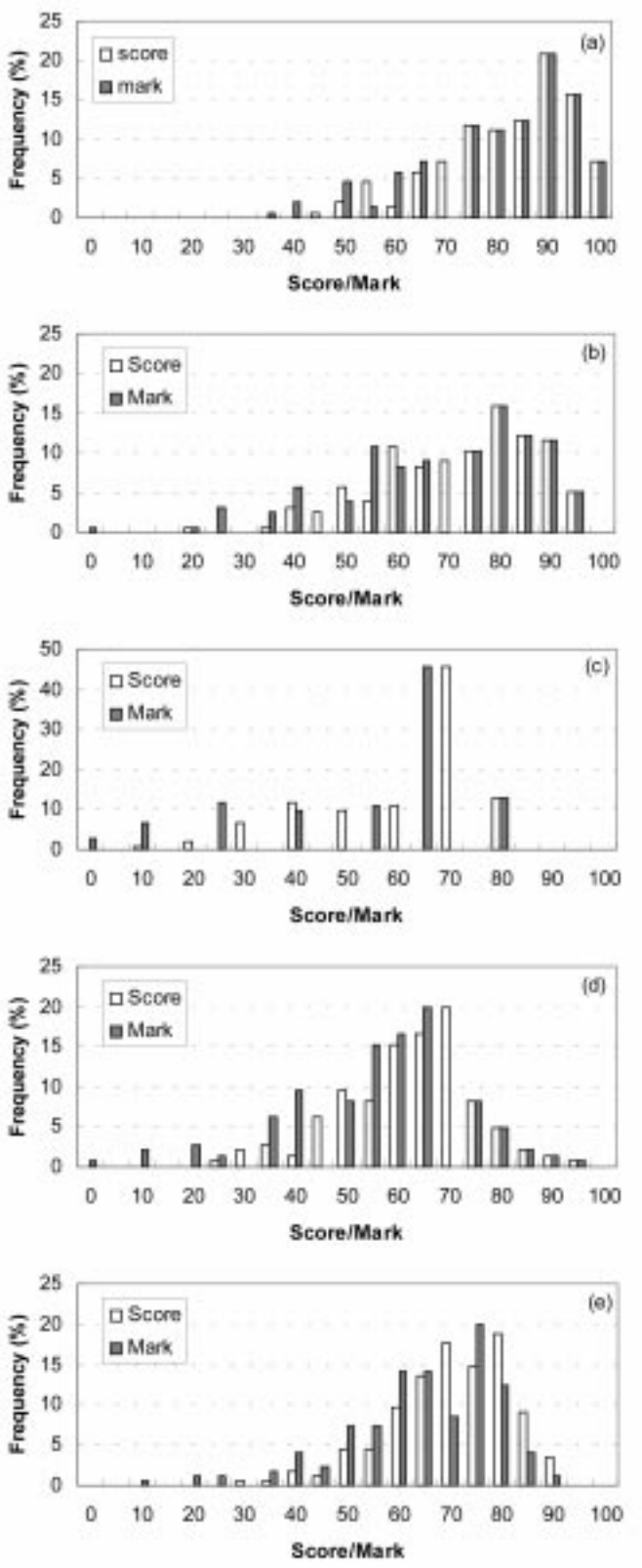

Fig. 1. Histograms of the scores and marks of (a) Test 1, (b) Test 2, (c) Test 3, (d) Test 4 and (e) the overall average of a class of 170 students.

Table 3. Statistics of the scores and marks of the individual tests and the overall average, as shown in Fig. 1

		Test 1	Test 2	Test 3	Test 4	Overall
Mean	Score	82.1	72.1	60.8	62.1	68.7
	Mark	78.1	66.4	52.7	54.4	62.6
Median	Score	85	75	70	65	69.5
	Mark	81	71	65	59	64.5
Mode	Score	90	80	70	70	68
	Mark	87	76	65	65	62
Standard deviation	Score	12.9	15.3	16.1	13.0	11.7
	Mark	14.9	18.5	21.1	16.9	14.5

reliable. The cumulative marks of a number of assessment units would give more satisfactory results.

Weightings of questions

Because the algorithms are developed for individual questions, it is acceptable to have questions with different weightings in a test. However, care should be taken to avoid very high weightings for questions that are either too easy or too difficult. Different weightings should be avoided when the number of questions is less than 20.

Calculations of overall marks

The overall marks should normally be calculated by averaging the marks of the individual assessment units. Marks obtained from converting from the overall average scores can lead to anomalies. Take an extreme example, where a student has obtained a score of 100 in one test and does not take the other three tests. The normal procedure gives the student an overall mark of 25, whereas the score-averaging approach gives a zero overall mark. The anomalies can result either from the different degrees of difficulty of the individual assessment units or from the different attitudes of the students towards the assessment units.

Compatibility and versatility

The marks converted by the algorithms are meant to be a true reflection of the students' competence, so they can be treated as the same as conventional percentage marks. The MCQ tests can be used standalone or as part of a course with several components with different assessment methodologies.

CONCLUSION

Algorithms have been developed for converting raw scores of MCQ tests to percentage marks based on the probability theory. A conversion table for questions with two to five choices of answers has been constructed and the method of implementing the conversion in spreadsheet programmes is demonstrated. The converted marks are compatible with the conventional marking scheme and are independent of class size and historical data. MCQ tests can therefore be used standalone or as assessment units in conjunction with the conventional assessment units. Provided a sufficient number of questions are used and the questions are constructed properly, the algorithms have been found to give satisfactory outcomes.

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, *TOEFL Technical Report TR-16: Using a New Statistical Model for Testlets to Score TOEFL*, Education Testing Services, Princeton, NJ (2001), pp.1–23.
3. 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.