International Journal of Engineering Education Vol. 29, No. 4, pp. 974–985, 2013
Printed in Great Britain

0949-149X/91 $3.00+0.00

A Scoring Method Based on Simple Probability Theory that Considers Partial Knowledge and Omission of Answers in Multiple-Choice Testing*

DESMOND ADAIR
School of Engineering, Nazarbayev University, Astana, 010000, Kazakhstan. E-mail: dadair@nu.edu.kz

MARTIN JAEGER
School of Engineering, University of Tasmania, TAS 7001, Australia, E-mail: mjaeger@utas.edu.au

A scheme for multiple-choice testing has been developed, based on simple probability theory, that takes into consideration partial knowledge and omissions of responses. It is tested against other methods which also take partial knowledge into account, namely, elimination testing for multiple-choice tests, as well as the conventional dichotomous method of scoring. The scheme is thought of as novel, in that it considers partial knowledge and omissions using a developed method based on probability theory. The results obtained were found to be quite similar to the 'more-complicated-to-use' elimination testing and lower results were found when compared with the conventional dichotomous method. Although the work here was undertaken to enhance testing within engineering education, there is no reason why this approach cannot be applied to any number of areas within science and arts subjects.

Keywords: scoring; multiple-choice assessment; computer-aided assessment

1. Introduction

Testing using multiple-choice questions is becoming increasingly popular in higher education because it can be used effectively to assess the breadth of knowledge in large cohorts of students [1] and is viewed favourably by both instructors and students [2]. Very often, especially in the first year of higher education, large cohorts of students have to be examined, a broad array of topics have to be included in the examination and students expect relatively fast feedback on their efforts. These can be achieved relatively easily by using tests with multiple-choice questions [3–6]. Multiple-choice tests can also allow the level of difficulty of a test to be more easily controlled [7, 8].

However, in spite of the apparent benefits mentioned above, multiple-choice testing should be used with caution. First there is the accusation that they typically promote shallow (factual recall) rather than deep learning (higher order skills) [5, 9] and fail to assess students' critical and communication skills or do not allow the capacity to develop an argument [10]. Other criticisms are of a decreased validity of tests due to guessing and failure to credit partial knowledge [11] and take into account the effect of omission of answers on the response sheet [12–14]. These concerns are addressed in this work.

Guessing during a test containing multiple-choice questions has long been a concern of examiners and much early work to correct for this has been carried out [15–17], although conversely others argue that guessing should not be a significant concern for test writers because examinees with a moderate level of engagement in the course material will rarely engage in truly random guessing [18, 19].

Although examinees may not be able to identify the correct answer to a multiple-choice question, they can quite often identify some of the options as being incorrect, and this is known as partial knowledge. It can be deduced then that an examinee's knowledge concerning a multiple-choice question falls into one of the following categories: full knowledge, partial knowledge, absence of knowledge, partial misconception and full misconception, and therefore any attempt to measure knowledge dichotomously would be unsatisfactory [20]. Also considered here is the tendency for examinees to omit answers. The tendency to omit items appears to be closely related to cognitive ability, which can be expected simply because higher ability examinees will be able to answer more items of a given question than lower ability examinees [21]. Along with cognitive ability, there may a higher tendency to omit answers by females as opposed to males. Today this tendency still exists [22, 23], although it is gradually becoming less pronounced [24], and is probably due to males being more willing to be higher risk takers than females. To have a method that can take into account omitted answers will help close the difference between male and female results, but not entirely.

To take both partial knowledge and omission of responses into account, a method based on simple probability has been developed here. The method is novel in that it considers partial knowledge and

* Accepted 14 March 2013.

omission of answers, using a theory based on simple probability that produces continuous functions that are suitable for multiple-choice assessment. It should be recognized that the application of the approach used here can be applied to many areas and levels of education.

The method is best used with an integrated computer-based test and item-analysis system, which has also been developed here, to reduce the tasks of grading and item analysis following testing. Computer-based tests offer several advantages over traditional paper-and-pencil testing in that there is a reduction in cost due to data entry, a much improved rate of disclosure, ease of conversion into databases and reduced risks due to human error [25, 26]. There is also some evidence that they are easier to manipulate to reduce cheating [27].

2. Research objectives

The specific research objectives associated with this work are:

1. The development of a scoring method based on simple probability theory, which considers partial knowledge and omission of answers in multiple-choice testing.
2. The development of formative assessment software and of a computerized formative assessment system.
3. The determination of the effectiveness of probabilistic scoring against the dichotomous method of assessment.
4. The determination of the effectiveness of probabilistic scoring against existing elimination methods.

3. Probabilistic scoring method

3.1 Partial knowledge

An assumption is first made that the examinees possess partial knowledge as the ability to eliminate some, but not all, of the wrong answers [28]. This definition is extended here to include partial and full misconception, which is the ability to eliminate some of the answers, one of which may be the right answer. Let the random variable X_{ni} denote examinee n's response to the multiple-choice item i, where X_{ni} can be either correct or incorrect, and must be present. Also, let θ be the continuous variable representing knowledge (which is either full, partial or none existent) that examinee n possesses regarding item i. For convenience, θ_{ni} will lie in the range –1 to +1, with important markers within the range being –1, indicating a complete misconception, 0 indicating no knowledge and 1 indicating full knowledge.

For $0 \leq \theta \leq 1$ and allowing for m alternatives for each multiple-choice item, it can be deduced that an examinee sees $m + (1 - m)\theta$ possible answers for item i so therefore the probability of choosing the correct or incorrect answer for a positive or zero amount of knowledge would be

$$p(X_{niC}|\theta) = \frac{1}{m + (1 - m)\theta}$$
$$p(X_{niI}|\theta) = \frac{(m - 1) + (1 - m)\theta}{m + (1 - m)\theta} \qquad (1)$$

where C and I in the subscripts stand for correct and incorrect responses respectively.

Equations (1) were deduced by first assuming θ to be positive and there are m possible responses per question. If, for the moment, the correct response is disregarded, this leaves $m - 1$ wrong answers. If an examinee then crosses out some of these wrong responses in proportion to his/her amount of knowledge then $\theta(m - 1)$ of the wrong responses will have been eliminated and $(1 - \theta)(m - 1)$ wrong responses remain. On including the right response an examinee would see $1 + (1 - \theta)(m - 1)$ or $m + 1(1 - m)\theta$ possible responses and so with equal probability assigned to each response, the probability of choosing the correct answer is one in $m + (1 - m)\theta$.

For $-1 \leq \theta < 0$, and again allowing for m alternatives for each multiple-choice item, the absolute amount of knowledge is used and it can be deduced that an examinee can see $m + (1 - m)|\theta|$ possible answers for item i and therefore the probability of choosing the correct or incorrect answer for a negative amount of knowledge would be

$$p(X_{niC}|\theta) = \frac{1 - |\theta|}{m + (1 - m)|\theta|}$$
$$p(X_{niI}|\theta) = \frac{(m - 1) + (2 - m)|\theta|}{m + (1 - m)|\theta|} \qquad (2)$$

Equations (2) were deduced, this time with θ as a negative and disregarding, for the moment, one of the wrong responses. This leaves $m - 1$ responses, one of which is correct, and the examinee now crosses out some of the answers in proportion to his/her absolute value of his/her amount of knowledge. This leads to $|\theta|(m - 1)$ of the responses being eliminated and leaves $(1 - |\theta|)(m - 1)$ responses. At this stage the probability that the right response is still available is $1 - |\theta|$. After uncovering the wrong response the examinee now sees $1 + (1 - |\theta|)(m - 1)$ or $m + (1 - m)|\theta|$ responses in all. If the correct answer is still available the probability that the examinee picks it is one in $m + (1 - m)|\theta|$.

3.2 Partial knowledge with omissions

The above theory is now added to by considering the case where some of the examinee responses are omissions. The probability of omissions can be considered as a function of the probability of omission conditional on the amount of knowledge an examinee has coupled with the prior probability of the amount of knowledge. This can be written as

$$p(X_{niO}) = \sum_n p(X_{niO}|\theta_n)p(\theta_n) \tag{3}$$

where the subscript O refers to omissions.

As θ is a continuous random variable, Equation 3 can be re-written as

$$p(X_{niO}) = \int_{-1}^{1} p(X_{niO}|\theta)p(\theta)d\theta \tag{4}$$

Assuming that the examiner has no prior information about the examinee's amount of knowledge then $p(\theta_n)$ is non-informative and is equal for all n, giving $p(\theta_n) = 0.5$.

An assumption is made that an examinee will always omit the response if their amount of knowledge is, say, between, $-l$ and l, and never have omissions otherwise, which means that the unconditional probability of omission is l. The unconditional probability of omission with the proportion of the questions omitted can now be calculated, thus $p(X_{niO}|\theta) = 1$ when $\theta \in (-l, l)$, and 0 otherwise

$$p(X_{niO}|\theta) \approx I(-P_O, P_O) \tag{5}$$

where P_o is the proportion of omitted responses.

Since the amount of knowledge θ is a continuous variable, Bayes' theorem can be written as

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{6}$$

where

$$p(X) = \int p(X|\theta)p(\theta)\ \mathrm{d}\theta$$

From Equation 6, the distribution of knowledge given an omission becomes

$$p(\theta|X_{niO}) = \frac{I(-P_O, P_O)}{2P_O} = U(-P_O, P_O) \tag{7}$$

The expected value and variance of the amount of knowledge are

$$E[\theta|X_{niO}] = 0 \tag{8a}$$

$$Var[\theta|X_{NiO}] = \frac{P_O^2}{3} \tag{8b}$$

The conditional probabilities for the correct and incorrect responses are still as given in Equations 1 and 2, except that the equations only apply to knowledge greater in the range $l < \theta < -l$.

The equations used to calculate the expected values of knowledge and the variances of knowledge for correct and incorrect responses while accounting for partial knowledge and omissions of responses are quite long and not given here, except by way of demonstration, the expected value of knowledge given a correct response is shown in the Appendix.

The test score S is the average amount of knowledge θ for all the questions. To keep a test score in the range $0 < S < 1$, the average amount of knowledge is rescaled by dividing by two and adding one half.

It can be shown that the expected value of the test score is

$$E(S|\varphi) = \frac{1 + P_C E(\theta|X_{niC}) + P_I E(\theta|X_{niI}) + P_O E(\theta|X_{niO})}{2} \tag{9}$$

where

$$\varphi = \{N, P_C, P_I, P_O\}$$

and the subscripts C and I stand for correct and incorrect responses respectively. N, P_C, P_I are the total number of questions in the multiple-choice test, the proportion of correct responses and the proportion of incorrect responses. This means that the expected value of the test score is the rescaled sum of the expected amount of knowledge conditional on each type of response, times the proportion of questions that have that response.

With the assumption that the amount of knowledge associated with each question in the test is independent of the amount of knowledge associated with the other questions (a reasonable assumption if no two questions cover closely related topics), then

$$Var(S|\varphi) = \frac{P_C Var(\theta|X_{niC}) + P_I Var(\theta|X_{niI}) + P_O Var(\theta|X_{niO})}{4N} \tag{10}$$

From Equation 10 it can be seen that the variance decreases as the number of questions increases, so the addition of more questions determines an examinee's score more accurately.

The variation of the equations used to calculate the expected values of knowledge and the variances of knowledge for correct and incorrect responses while accounting for partial knowledge and omissions of responses with the number of alternatives

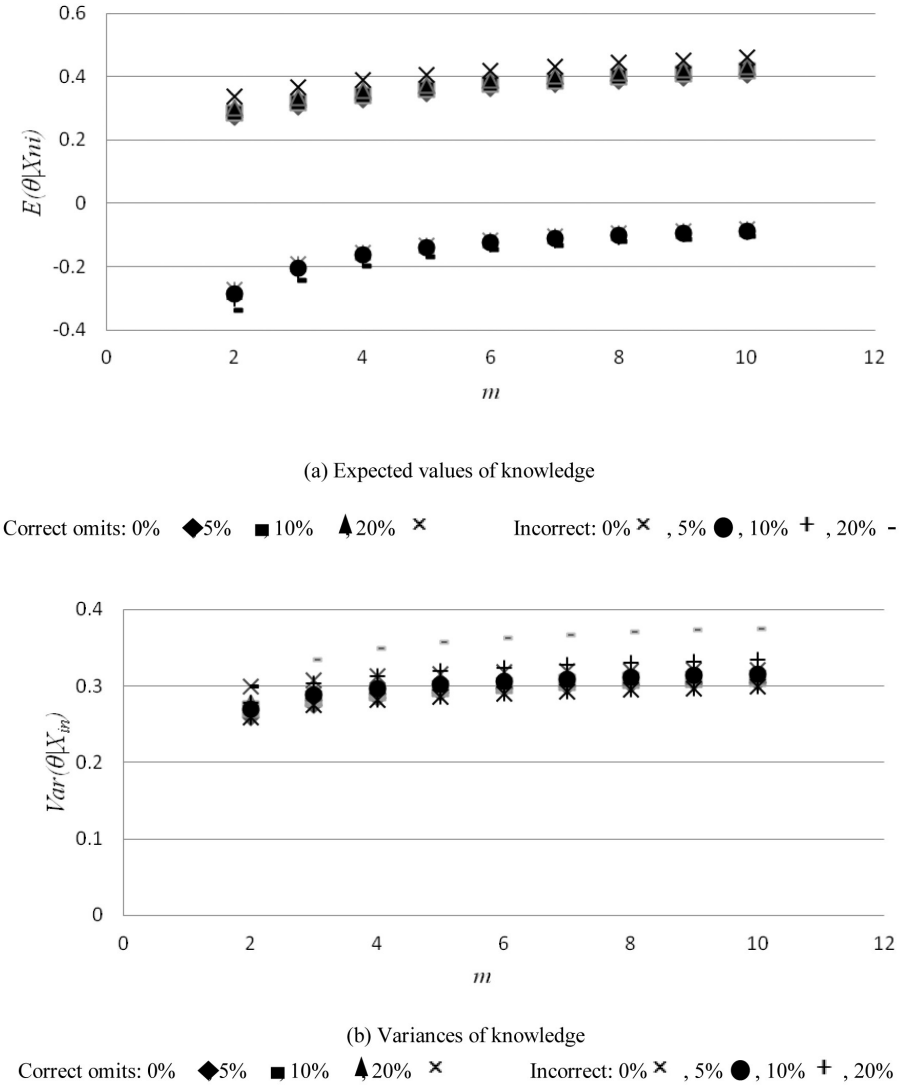

Fig. 1. Variation of expected values of knowledge and variances of knowledge with number of alternatives per multiple-choice item and percentages of omissions.

per multiple-choice item, m, and the percentages of omissions is shown on Fig. 1.

These values, together with Equation 9, are used to calculate the test score for the probabilistic scoring method. Confidence intervals for the score can also be deduced using these values together with Equations 9 and 10.

4. Methods and procedures

Because of the ease of use and efficiency, the multiple-choice format for testing is popular, but the inherent weakness of the conventional dichotomous method of scoring limits informative feedback to improve the teaching process and facilitate students' continuous learning. In the cycle of teaching and learning this weakness must be addressed so that the assessment supports teaching and learning development rather than becoming the 'final event'.

4.1 Design of a multiple-choice item

In this work, each multiple-choice item is composed of a correct answer and several distractors. The design of distractors is very important [29], for example, parallel grammar was adopted to avoid giving clues, options addressed the same content, common examinee errors were incorporated into distractors, true statements that do not correctly answer the question were used, and the distractors were all reasonable choices. A good database was maintained by the instructors who have taught the course more than once, and note was taken of questions with a large proportion of correct (or incorrect) responses as these have little value in discrimination [30].

Table 1. Personal characteristics

Characteristic	
Average age	19.5 years
Percentage female	43%
Major	Mechanical Eng. 76%, Civil Eng. 24%
Preferred assessment style(s)	
• Final written examination	21%
• MCQ topical tests	46%
• Assignments	23%
• Orals	10%

4.2 Subjects

The subjects were 63 first year university students in the first semester of an engineering course with personal characteristics as summarized in Table 1.

4.3 Formative assessment software

A computer-aided assessment package was written using the Java programming language. The Graphical User Interface (GUI) was designed consisting of a series of JFrames on which were placed panels, buttons, text fields, labels, checkboxes, images and animations as appropriate. The package was fully interactive and designed to be user friendly [31]. The GUI was networked to a central server where data could be deposited, stored and retrieved for further analysis. A database was constructed and placed on the server, consisting of 300 multiple-choice questions, which fully covered the teaching material. Each question was designated a weighting factor according to its degree of difficulty.

In order to compare the current probabilistic scoring method with other methods the examinees were asked to complete their tests by way of the GUI using the elimination testing method [32]. In this way full data for analyses and comparisons of the probabilistic scoring method and the conventional dichotomous method of scoring were also implicitly gathered. The elimination method takes only partial knowledge into account, by allowing an examinee to choose as many incorrect options as they can identify. One point is awarded for each incorrect choice identified, but k points are deducted (where $k = m - 1$) if the correct option is identified as incorrect. Then elimination testing can be classified as: completely correct score (+3), partially correct score (+2, +1), no-understanding score (0), partially incorrect score (partial misconception) (–1, –2), completely incorrect score (full misconception) (–3) [33]. It has however been claimed that examinees find the test instructions associated with the foregoing classification complicated and confusing and so a more direct and therefore easier to understand classification has been adopted for this work [6]. The classification is: full knowledge (4), partial knowledge (3, 2, 1), absence of knowledge (0), partial misconception (–1, –2) and full misconception (–3).

It is very important that examinees are told explicitly about the scoring that is being used. When negative marking is used, it is important to indicate that answering based on partial knowledge (i.e. being able to eliminate some options) is generally advisable, but random guessing is not. Therefore at the beginning of each test, the test instructions and scoring guide were available for viewing as shown in Table 2.

Figure 2 shows a typical interface for the formative assessment software. To help ensure that an examinee had carefully considered the range of answer options fully per question, the software was designed so that the examinee could not submit the solution without choosing an answer from each of the alternatives placed in the JComboBox for each of the four answer options.

4.4 Formative assessment system

An overall view of the Formative Assessment System is shown in Fig. 3. The system provides a platform for computer-based elimination testing as well as algorithms to derive the results for the probabilistic scoring and dichotomous methods.

The computer-based assessment system is linked to the Answer Records, which is a database to collect and store the complete answer record.

Table 2. Test instructions and scoring guides

Test instructions	Scoring guide per question
1. If you are sure an option is correct then Select: **Correct** 2. If are sure an option(s) is incorrect then Select: **Wrong** 3. If you are not sure of an option(s) then Select: **Not Sure** 4. If you wish to not answer the question, simply enter **Not Sure** for each option. Four entries of **Not Sure** for a question means that you have omitted the question. You are encouraged to omit answers rather than guess answers	Your score for each question will be calculated as: One mark is awarded if the option with **Correct** is the correct answer. One mark is awarded if the option with **Wrong** is the wrong answer. No mark is awarded for the option where **Not Sure** has been selected. If **Not Sure** has been selected for all of the options of a question no marks are awarded for that question. You score per question will be in the range –3 to 4.

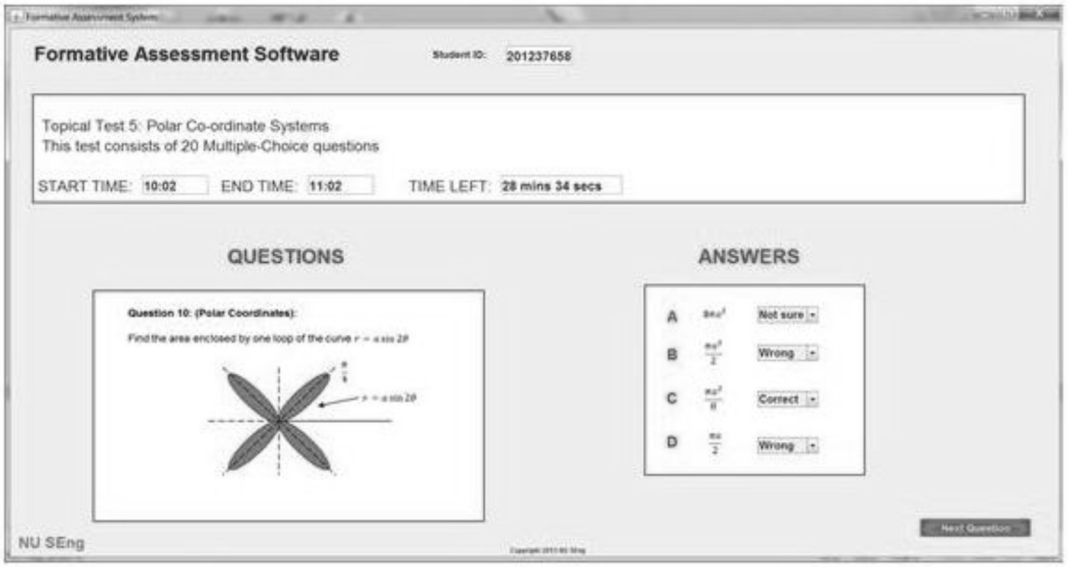

Fig. 2. GUI showing partial knowledge (a score of 2).

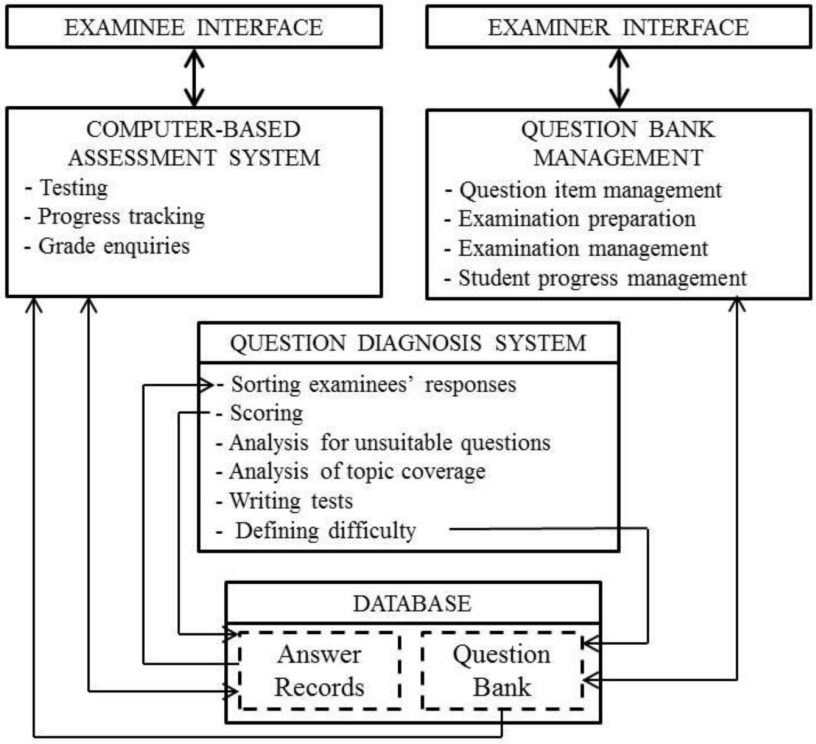

Fig. 3. Formative assessment system overview.

There is also a feedback loop to enable students to check their own learning process and increase their learning efficiency by means of constructive interaction. The main functions of the computer-based assessment are: to verify a student's eligibility for the test, to display test information, including times, how the test is scored and the test items. The Question Bank is normally accessed by instructors to design tests, revise tests or re-use tests. The system allows the assessment designer to edit multiple-choice questions, constructed-respond items and true/false items as well as specifying scoring modes. It can also be used for creating original questions, browsing/selecting questions and random selection of test items. For security, each instructor is allowed to specify for a test: the starting time of a test, who can take the test, the scoring mode, the time allowed and question values. Then only eligible examinees can take the test at a specified time.

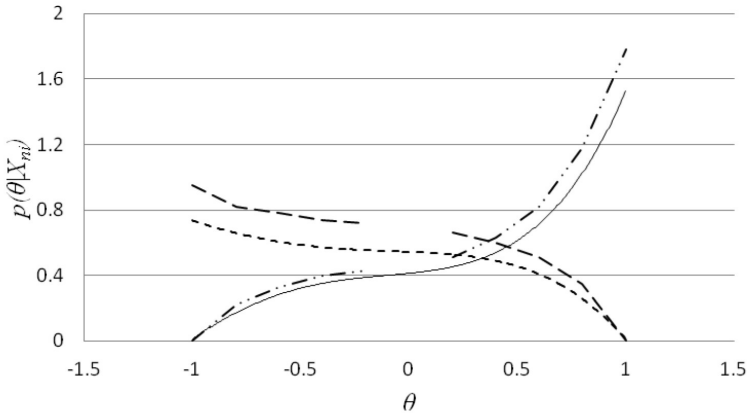

Fig. 4. Probability density functions of amount of knowledge conditional on response. ($m = 4$, 0% omits correct ——, 0% omits incorrect - - - -, 20% omits correct – – –, 20% omits incorrect —).

4.5 Procedure

Eight topical tests were administered over a 12-week semester. Each of the eight tests included 20 multiple-choice questions. The major areas covered were: differentiation, partial differentiation, integration, ordinary differential equations, polar-coordinate systems, numerical methods, matrices and vectors and probability. All multiple-choice items within the tests had three distractors and one correct answer.

Several precautions were taken to ensure that examinees understood the elimination testing procedure. First, before the initial test, lecture time was used to explain the procedure, with a demonstration given. Second, sample tests were made available through the university library, and third, examinees were given time to read and fully comprehend the instructions as given in Table 2.

5. Results and discussion

5.1 Probabilistic scoring method

The main aim of this work is the development and analysis of the use of a probabilistic scoring method for tests that use multiple-choice items. The method is based on basic rules of probability and accounts for partial knowledge and omission of answers. Fundamental to the method is the generation of distributions of the probability density of knowledge conditional on correct and incorrect responses. These distributions are shown in Fig. 4 for the alternative number for each multiple-choice item set at 4.

When m, the number of alternatives, increases it was found that the above probability density function, conditional on the correct response, shifts towards one. In the limit, guessing becomes impossible and only full knowledge allows the examinee to respond correctly. The expected knowledge given incorrect responses approaches zero and, again in the limit, only full knowledge allows for a correct response. All other amounts of knowledge, from full misconception to almost full knowledge, result in an incorrect response.

Figure 5 shows the test scores estimated from the probabilistic scoring method as a function of the correct responses when there are no response omissions. The estimated scores are found by using Equation 9 and the values derived from equations used to calculate the expected values of knowledge (for example Equation A.1) for $m = 2$, 4 and ∞.

Also shown in Fig. 5 are the confidence intervals when the number of questions equals twenty. These are calculated by assuming the number of questions in a test is large and, by using the Central Limit Theorem, that the test score S has an approximate Gaussian distribution. The 95% confidence interval can then be estimated using

$$C.I. \approx E(S) \pm 2\sqrt{Var(S)} \qquad (11)$$

Here $E(S)$ is calculated using Equation 9 and $Var(S)$ using Equation 10. It can be seen from Equation 10 that, as the value of m increases, the variance of knowledge given a correct response steadily approaches zero because it is increasingly more difficult to guess the answer.

As the number of alternative answers per question increases, the variance of knowledge given incorrect responses approaches one third because the distribution of knowledge approaches a uniform distribution from negative one to positive one.

5.2 Analyses of examinee guessing

The problem of examinee guessing, when using the

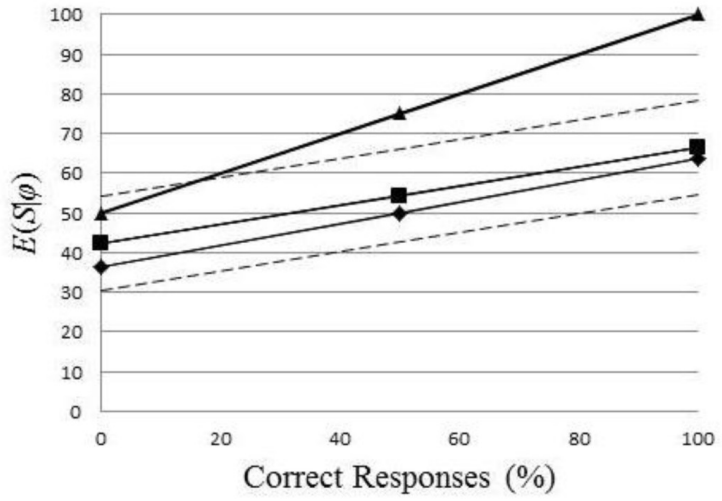

Fig. 5. Estimated score by probabilistic scoring method v. correct response. ($m = 2$, $m = 4$ ■, $m = \infty$ ▲, confidence interval limits ($N = 20$) – –).

probabilistic scoring method, was first investigated by using the results for the bottom 30% of the examinee sample ranked according to their dichotomous scores [34]. The procedure [35] is to check the performance of relatively low-ability examinees for the most difficult items, and if guessing was minimal, the performance of these examinees would be close to zero or below the chance level of 25%. The most difficult items were those with the lowest *p*-value (proportion of examinees who answer the item correctly). The focus was on the lower relatively lower ability group because they have the tendency to guess since they have only partial knowledge for most of the items [34]. The responses and dichotomous method scores of the lowest 30% of examinees to the four most difficult items are shown in Tables 3 and 4.

Table 3. Responses of relatively low-ability examinees to most difficult items

Item	p-value	Correct	Incorrect	Omits
T2, No. 10	0.481	15.40%	81.34%	3.26%
T4, No. 1	0.456	13.56%	82.21%	4.23%
T5, No. 3	0.303	11.23%	84.56%	4.21%
T5, No. 9	0.290	8.01%	90.11%	1.88%

Table 4. Performance of relatively low-ability examinees for most difficult items

Dichotomous method scores	Number of examinees	Percentage
4	1	5.3%
3	2	10.5%
2	1	5.3%
1	3	15.8%
0	12	63.1%

As can be seen from Table 3, the percentages of the relatively low-ability examinees were high for incorrect answers and the assumption is that if random guessing is involved, then the chance of an examinee being 'correct' for a four-option multiple-choice item is 25%. The percentages of correct responses can be seen to be some 25% below this figure.

Table 4 shows the scores of the examinees in the lower 30% band for these four difficult items. It can be seen that one examinee managed to get all four difficult items correct, whereas two got three answers correct, one achieved only two correct answers, three achieved one correct answer only and twelve got none of the difficult answers correct. It is noticeable that the amount of omitted answers was very low, which may indicate that this low-ability group could be considered high-risk takers. With normal elimination testing this would benefit an examinee [11]. Also, as the number of correct, incorrect and omitted answers is fundamental to the probabilistic method, high-risk takers could be a significant threat. However from Tables 3 and 4 it is evident that high-risk taking or guessing seems to be of minimal concern for this method.

A second method was used to investigate guessing, where the fit of the items to two- and three-parameter item response theory models was examined. There are three basic parameters involved in multiple-choice testing, namely, item difficulty, item discrimination and guessing. The two-parameter item response theory models take into consideration items having different item difficulty and item discrimination indices, but assumes minimal guessing. The three-parameter item response theory models

Table 5. Number of misfit items for each item response theory model

	Topical test number							
	1	2	3	4	5	6	7	8
	Number of misfit items							
2-PL logistic	2	2	2	2	2	1	2	2
2-PL normal	2	2	2	1	2	2	2	2
3-PL logistic	1	2	1	1	1	1	2	1
3-PL normal	1	1	2	1	1	1	1	1

take into account all three parameters. In theory, the three-parameter item response theory models fit data from multiple-choice test best, but data with minimal guessing also fit the two-parameter well.

Here a commercial software package [36] was used for the data analyses. The item response theory models used were the 2-Parameter Logistic (2-PL) and 3-Parameter Logistic (3-PL), and the response-function metrics set at logistic and normal. This gave a total of four possible item response theory models, 2-PL logistic, 2-PL normal, 3-PL logistic, and 3-PL normal. Since each test consisted of 20 items, χ^2 statistics was used to assess the degree of fit of the response data to the models. If the χ^2 calculated at the 0.05 level of significance is greater than the $\chi^2_{critical}$ at the associated degree of freedom, then the item did not fit the model. The number of misfit items for each of the eight tests and four models are shown in Table 5.

The number of misfit items is significantly low for the two-parameter models, which assume minimal guessing and there is a slight decrease in the number of misfit items for the three-parameter models, which take guessing into consideration. It can be seen therefore that guessing was minimal for the probabilistic scoring method since the data fit well with the 2-parameter item response theory which assumes minimal guessing.

5.3 Comparison of scoring methods

The means of the test scores for the eight tests as found by the probabilistic scoring method, the elimination testing method and the dichotomous method are shown in Table 6, with results marked out of a total of twenty. It can be seen that, except for Topical Test 5, the means (correct) found by the dichotomous method were higher than those found by both the probabilistic scoring method and the elimination method. The mean values found by the probabilistic scoring method and the elimination testing method are in general fairly similar, but the elimination method gives slightly higher values throughout. As the values of both the probabilistic scoring method and the elimination testing method are generally giving lower mean values for each test when compared with the dichotomous method, this would indicate that not all the correct answers under the dichotomous method are based on the true knowledge of the examinees. For Topical Test 5, the means for each of the test scores were low (< 55%). This could mean that the test was so difficult that the examinees could no longer rely on knowing the correct answer, which the dichotomous method demands, but were drawing more on partial knowledge to improve their scores with the probabilistic scoring method and elimination testing method.

Comparison of the probabilistic scoring method and the elimination testing method is given on Fig. 6. Here the probabilistic scoring method is compared with the elimination testing method results of the present work and results found by the Coombs elimination procedure [33].

It can be seen from Fig. 6 that the probabilistic scoring method gives results similar to both the present elimination method and the Coombs elimination procedure, for the mid-range of 'Correct Responses'. A difference in results occurs above a score of say 16 (80%), where the elimination method gives results some 20% higher compared with the probabilistic scoring method. Further testing for results above 80% is needed to see if this is just an anomaly. The other area of concern is what happens for a score of say less than 10 (50%). More data would need to be generated for low scoring tests to properly investigate this range.

Table 6. Means for Topical Tests 1–8

	Probabilistic scoring method		Elimination testing method	Dichotomous method
	Score	Percentage omits	Full knowledge	Correct
Topical Test 1 (Differentiation)	12.56	4.3%	12.86	15.67
Topical Test 2 (Partial Differentiation)	11.48	3.5%	11.51	12.43
Topical Test 3 (Integration)	12.45	3.3%	13.52	16.11
Topical Test 4 (ODEs)	11.45	6.15%	11.66	11.74
Topical Test 5 (Polar-Coordinate Systems)	*11.10*	7.8%	*9.45*	9.04
Topical Test 6 (Numerical Methods)	13.02	1.3%	14.96	18.97
Topical Test 7 (Matrices & Vectors)	12.10	2.4%	12.48	14.45
Topical test 8 (Probability)	13.06	0.7%	15.67	19.03

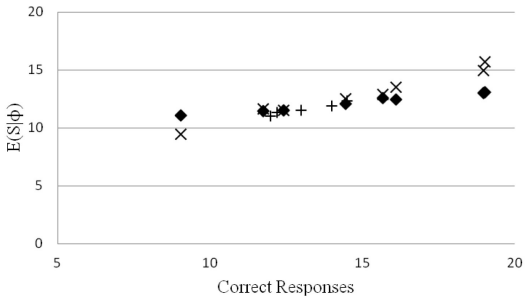

Fig. 6. Comparison of mean values (probabilistic scoring method ◆, elimination testing method **X**, Coombs elimination method +).

Repeated measures, one-way analyses of variance were conducted on the eight test scores to determine if the tests differed in difficulty. The within-subject factor was the number of times (8) the tests were administered, whereas the dependent variables for the analyses were the test scores based on the probabilistic scoring method, the elimination method and the dichotomous method. No significant differences in means for any of the three methods were found across the eight tests with typical values of the F ratio being $F(7, 434) = 1.54$, $p = 0.15$.

5.4 Reliabilities

Table 7 contains Cronbach α values, internal consistency estimates of reliability, for the eight tests using the probabilistic scoring method, the elimination scoring method and the dichotomous method. The range of values, between 0.60 and 0.85 has been proposed [37] as suitable for reliability and all the values shown in Table 7 are within these limits. For all eight tests, the reliabilities for the probabilistic scoring method were greater than those for the dichotomous method, although these differences were somewhat marginal for three of the tests.

The probabilistic and elimination scoring methods have similar reliabilities. It can be concluded therefore that there is a slight advantage in reliability of both the probabilistic and elimination methods when compared with the traditional dichotomous testing. These results are in line with [32, 38], who observed that tests using elimination testing are more reliable than the dichotomous method of testing.

Table 7.

Topical test number	Probabilistic scoring method	Elimination testing method	Dichotomous method
1	0.74	0.72	0.68
2	0.82	0.77	0.76
3	0.69	0.73	0.68
4	0.78	0.76	0.71
5	0.71	0.75	0.71
6	0.82	0.79	0.77
7	0.75	0.72	0.73
8	0.83	0.78	0.80

5.5 Limitations

While the results, in the main, are encouraging, more work needs to done for correct responses lying in the range > 50% and < 80%. It is still not clear nor proven that the probabilistic approach is suitable in these ranges. This of course will require well designed tests to direct student responses into these areas and it is envisaged that a trial period will be required to achieve this.

In addition to the above, more tests will be carried out with different cohorts hopefully to confirm the results obtained so far.

6. Conclusion

A scoring method theory based on simple probability theory that considers partial knowledge and omission of answers in multiple-choice testing has been developed, together with software and a system for formative assessment.

The results obtained by the probabilistic approach when compared with the conventional dichotomous method of scoring were similar to those obtained by elimination methods of scoring in that the scores decreased significantly. The exception was when the Topical Test became difficult.

The results were tested against elimination and dichotomous methods for multiple-choice questions with similar results to existing elimination methods obtained in the central portion of the 'Correct Response' range.

Therefore the results have shown the feasibility of adopting the probabilistic scoring method for multiple-choice testing, providing the results lie in the range of 'Correct Responses' of > 50% and < 80%. More tests will have to be designed and conducted to give results outside these ranges in order to provide more confidence in the method. However, given this proviso, the probabilistic scoring method performed as well as other elimination methods and gave similar results to these methods.

References

1. L. R. Bretts, T. J. Elder, J. Hartley and M. Trueman, Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, **34**(1), 2009, pp. 1–15.
2. M. G. Simkin and W. L. Kuechler, Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, **3**, 2005, pp. 73–79.
3. M. Bush, A multiple choice test that rewards partial knowledge, *Journal of Further and Higher Education*, **25**, 2001, pp.157–163.
4. R. L. Williams and L. Clark, College students' ratings of

student effort, student ability and teacher input as correlates of student performance on multiple-choice exams, *Educational Research*, **46**, 2004, pp. 229–239.
5. D. Nicol, E-assessment by design: using multiple-choice tests to good effect, *Journal of Further and Higher Education*, **31**, 2007, pp. 53–64.
6. A. Ben-Simon, D. V. Budescu, and N. Nevo, A comparative study of measures of partial knowledge in multiple-choice tests, *Applied Psychological Measurement*, **21**(1),1997, pp. 65–88.
7. W. A. Mehrens and I. J. Lehmann, *Measurement and Evaluation in Education and Psychology*, 3rd edn, Holt, Rinehart and Winston, New York, 1984.
8. R. Bennet, On the meanings of constructed response, in R. Bennet and W. Ward (Eds), *Construction verses Choice in Cognitive Measurement*, Lawrence Erlbaum, Hillsdale NJ, 1993, pp. 1–27.
9. K. Scouller, The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay, *Higher Education*, **35**, 1998, pp. 453–472.
10. M. Paxton, A linguistic perspective on multiple choice questioning, *Assessment and Evaluation in Higher Education*, **25**, 2000, pp. 109–119.
11. T. B. Kurz, A review of scoring algorithms for multiple-choice tests, Annual Meeting of the Southwest Educational Research Association, 21–23 January, San Antonio, TX, 1999.
12. M. J. Slakter, The effect of guessing strategy on objective test scores, *Journal of Educational Measurement*, **5,** 1968, pp. 217–221.
13. G. Ben-Shakhar and Y. Sinai, Gender differences in multiple choice tests: The role of differential guessing tendencies, *Journal of Educational Measurement*, 28, 1991, pp. 23–35.
14. D. Zhu, Gender and ethnic differences in tendencies to omit responses on multiple-choice tests and impact of omits on test scores and score ranks (Doctoral dissertation, University of Iowa, 1995). *Dissertation Abstracts International*, **56**, 1995, p. 2213.15.
15. J. Diamond and W. Evans, The correction for guessing, *Review of Educational Research*, **43**(2), 1973, pp. 181–191.
16. R. B. Frary, Formulae scoring of multiple choice tests (Correction for guessing), *Educational Measurement: Issues and Practice*, **7**(2), 1988, pp. 33–38.
17. D. Budescu and M. Bar-Hillel, To guess or not to guess: A decision-theoretic view of formula scoring, *Journal of Educational Measurement*, **30**(4), 1993, pp. 277–291.
18. R. L. Ebel, Blind guessing on objective achievement tests, *Journal of Educational Measurement*, **5**(4), 1968, pp. 321–325.
19. C. Andrà and G. Magnano, Multiple-choice math tests: should we worry about guessing? *Quanderni di Ricerca in Didattica (Mathematics)*, **21**, 2011, pp. 235–243.
20. T. P. Hutchinson, Some theories of performance in multiple choice tests, and their implications for variants of the task, *British Journal of Mathematical and Statistical Psychology*, **35**, 1982, pp. 71–89.
21. J. Grandy, Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring (GRE Board Professional Report No. 83-16P), Educational Testing Service, Princeton, NJ, 1987.
22. S. von Schrader and T. Ansley, Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000, *Applied Measurement in Education*, **19**(1), 2006, 41–65.
23. P. Everaert and N. Arthur, Constructed-response versus multiple choice: the impact on performance in combination with gender, Working Paper, Faculty of Economics and Business Administration, University of Ghent, 2012.
24. N. S. Cole, *The* ETS gender study: How females and males perform in educational settings, Educational Testing Service, Princeton, NJ, 1997.
25. S.-H. Chang, P.-C. Lin and Z. C. Lin, Measures of partial knowledge and unexpected responses in multiple-choice tests, *Educational Technology & Society*, **10**(4), 2007, pp. 95– 109.
26. A. S. Hagler, G. J. Norman and L. R. Radick, K. J. Calfas and J. F. Sallis, Compability and reliability of paper- and computer-based measures of psychosocial constructs for adolescent fruit and vegetable and dietary fat intake, *Journal of the American Dietetic Association*, **105**(11), 2005, pp. 1758–1764.
27. S. M. Bodmann and D. H. Robinson, Speed and performance differences among computer-based and paper–pencil tests, *Journal of Educational Computing Research*, **31**(1), 2004, pp. 51–60.
28. R. B. Frary, The effect of misinformation, partial information and guessing on expected multiple-choice test item scores, Applied Psychological Measurement, 4, 1980, pp. 79–90.
29. T. M. Haladyna and S. M. Downing, A taxonomy of multiple-choice item writing rules, *Applied Measurement in Education*, **2**, 1989, pp. 37–50.
30. J. Kehoe, Basic item analysis for multiple-choice tests, *Practical Assessment, Research and Evaluation*, **4**(10), 1995.
31. W. O. Galitz, *The Essential Guide to User Interface Design,* 3rd edn, Wiley Publishing, Indianapolis, IN, 2007.
32. C. H. Coombs, J. E. Miholland and F. B. Womer, The assessment of partial knowledge, *Educational and Psychological Measurement*, 16, 1956, pp. 13–37.
33. D. A. Bradbard and S. B. Green, Use of the Coombs elimination procedure in classroom tests, *Journal of Experimental Education*, **54**, 1986, pp. 68–72.
34. R. K. Hambletin, H. Swaminathan and H. J. Roger, *Fundamentals of Item Response Response Theory*, Sage, Newburg Park, CA, 1991.
35. P. K. Agble, A psychometric analysis of different scoring strategies in statistics assessment, Doctoral dissertation, Kent State University, OH, 1999.
36. BILOG-MG 3, SSI (Scientific Software International), Lincolnwood, IL, 2012.37.
37. R. L. Linn and N. E. Gronlund, *Measurement and Assessment in Teaching*, 8th edn, Prentice Hall, Upper Saddle River, NJ, 2000.
38. A. R. Hakstian and W. Kansup, A comparison of several methods of assessing partial knowledge in multi-choice tests: II. Testing procedures, *Journal of Educational Measurement*, **12**, 1975, pp. 231–239.

Appendix: Expected value of knowledge given a correct response

$$E[\theta|X_{niC}] = \frac{f(m.l')}{g(m.l')} \qquad \text{A(1)}$$

where

$$l' = \frac{1}{l} - 1 \approx \frac{1}{P_0} - 1 \qquad \text{A(2)}$$

Here

$$
\begin{aligned}
f(m,l') &= [(-2\ln(l'+1)+2\ln(ml'+1))m^2 \\
&\quad + [((-4+4\ln(ml'+1)-4\ln(l'+1))m^2-2+6m]l' \\
&\quad + [4m-1+(-3-2\ln(l'+1)+2\ln(ml'+1))m^2](l')^2 \\
g(m,l') &= (m-1)(l'+1)[2[-2\ln(ml'+1)+2\ln(l'+1)+(\ln(ml'+1)-\ln(l'+1)+1] \\
&\quad + 2[(\ln(ml'+1)-\ln(l'+1))m+2\ln(l'+1)-2\ln(ml'+1)]]
\end{aligned}
$$

Desmond Adair holds a Ph.D. in Mechanical Engineering from Imperial College of Science and Technology, University of London. He spent a number of years working as a Senior Research Engineer with NASA in Mountain View, California and NPL in Teddington, England. Professor Adair has worked for British Aerospace and the UAE Defence Forces in senior research and education positions, and is currently a Professor of Mechanical Engineering at Nazarbayev University, Republic of Kazakhstan.

Martin Jaeger holds a Ph.D. in Civil Engineering (Construction Economy and Management) from the University of Wuppertal, Germany. He has spent the last fifteen years working as a site manager, consultant, and a Senior Lecturer in Germany and the Middle East and is a Research Associate with the University of Tasmania, Australia.