Different Methods of Multiple-Choice Test: Implications and Design for Further Research

Annie W.Y. Ng and Alan H.S. Chan

***Abstract*—There are a variety of multiple-choice test methods nowadays viz. conventional multiple-choice, liberal multiple-choice, elimination testing, confidence marking, probability testing, and order-of-preference scheme. However, there are no findings identifying the best test method for use. Though the multiple-choice test is a task of human decision making of selecting the correct answer amongst several alternatives, the application of signal detection theory for quantitative analysis of one's performance in the test has not been used. Therefore, a study plan on multiple-choice test is proposed here to fill the gap in the literature. This paper would provide a clearer picture on the multiple-choice test methods and thus facilitate people to conduct effective assessments in various subject areas.**

***Index Terms*— multiple-choice test, signal detection theory, sensitivity, response criterion**

I. Introduction

Today multiple-choice test is the most common and widely used assessment tool for the measurement of knowledge, ability and complex learning outcomes [1,2]. A multiple-choice item usually consists of a stem which presents a problem situation, and several alternatives which provide possible solutions to the problem. The stem may be a question or an incomplete statement. The alternatives include the correct answer and several plausible answers called distracters.

There are various kinds of research studies to determine the optimal number of alternatives for multiple-choice items [3], position of the correct answer(s) in multiple-choice items [4,5], multiple-choice item-writing guidelines [6,7,8], and male-female differences in multiple-choice tests [9,10,11].

In this paper, a review on multiple-choice test methods is presented. Previous studies on the comparison of different multiple-choice test methods are examined. The application of signal detection theory for quantitative analysis of one's performance in multiple-choice items is also recommended. A future study plan on multiple-choice methodology based on the review is then proposed. This paper would allow people know more about multiple-choice test methods and thus facilitate them to conduct more effective assessment tests.

Manuscript received December 30, 2008. The work described in this paper was supported by a grant from City University of Hong Kong, Hong Kong, China [Project No. 6980067].

Annie W.Y. Ng is with the City university of Hong Kong, Hong Kong, China (e-mail: annieng@cityu.edu.hk). Alan H.S. Chan is with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China (phone: 852-27888439; fax: 852-27888423; e-mail: alan.chan@cityu.edu.hk).

II. Multiple-choice test methods

A conventional multiple-choice test is one of the most widely used assessment methods. When faced with a question in a conventional multiple-choice test, a candidate must evaluate each option and choose the most appropriate one. The candidate's test score is usually calculated by tallying the number of correct responses in the test, and is used as a measurement of the candidate's knowledge of the materials and contents covered by the test. However, the number of correctly answered questions may be composed of two numbers: the number of questions where the candidate actually knows the answer, and the number of questions where the candidate correctly guesses the answer. Sometimes, the candidate knows only part of the answer or is uncertain of the answer. But this partial knowledge is not captured or taken into consideration in the conventional multiple-choice test method.

The limitations of conventional multiple-choice test method have been noted and alternative methods for administering multiple-choice tests that increase the complexity of responding and scoring have been introduced. There are a variety of multiple-choice test methods nowadays to discourage scoring by guessing and give examinees an opportunity to reflect the partial knowledge actually possessed by candidates. Appendix 1 summarizes the multiple-choice test methods commonly used in previous studies from 1990 to 2008. Details of the instruction and response modes and scoring rules are also given. *Liberal multiple-choice test* allows candidates to select more than one answer to a question if they are uncertain of the correct one. The term 'liberal' is used to denote the extra dimension of choices. A method which turns out to be a variation of the liberal multiple-choice test is *elimination testing* [12,13]. This requires candidates to select the answers which they believe are wrong, rather than selecting those they believe are or may be right.

To take into account the confidence levels that examinees have in their answers, *confidence marking* has been proposed. Candidates have to assign a confidence level to their best answer to each question. Being able to properly judge the confidence of one's answers is an important part of being knowledgeable in some professions like medical [14]. *Probability testing* is a complicated version of confidence marking [2]. Examinees are instructed to allocate 100 points among all given options so as to reflect their subjective probability of being correct. *Complete ordering* is a special version of probability testing [2]. All alternatives have to be ranked according to their degree of plausibility. In *partial ordering* candidates are instructed to select one answer, if sure.

If unsure, they are instructed to rank all feasible responses in order.

Permutational multiple-choice question is one in which each question consists of a two-part stem and N alternatives. The two parts of the stem must ask about closely related issues. The alternatives include two correct answers and N-2 distracters. Examinees must match up each stem with the appropriate key in order to answer the question correctly.

The above multiple-choice test methods are used in different levels of extent for assessment purpose. The response mode of the permutational multiple-choice question is quite different from others. Conventional multiple-choice, liberal multiple-choice, and confidence marking are three typical one-stem multiple-choice methods. The liberal multiple-choice test works in the way of elimination testing. However, it is preferable to the elimination testing as it encourages examinees to think positively rather than negatively [12]. Confidence marking, probability testing, and order-of-preference scheme are corresponding to each other, and confidence marking is the simplest version.

III. Comparison of Multiple-Choice Test Methods

There are a variety of multiple-choice test methods nowadays. Comparison of alternative multiple-choice test methods with respect to conventional one has been reported. Order-of-preference scheme is a variation of confidence marking. Both order-of-preference scheme and confidence marking are better than conventional multiple-choice [15, 17]. Alnabhan [15] found that partial ordering produced higher reliabilities than the traditional method, and produced the highest validity coefficients. Swartz [17] revealed that confidence marking offered advantages over traditional form in terms of measurement accuracy.

Elimination testing is equivalent to liberal multiple-choice. Both liberal and elimination methods are better than conventional one [13, 16]. Liberal multiple-choice was found to reward partial knowledge more generously and punish misinformed examinees more severely than conventional one [16]. Bradbard et al. [13] found that in elimination procedure, the chance of guessing is reduced and partial knowledge is measured.

A comparison of conventional multiple-choice, elimination testing, confidence marking, probability testing, and order-of-preference scheme was made by Ben-Simon et al. [2]. But none of the methods emerged as the best.

To sum up, the current findings could not help us identify which multiple-choice method is the best to use. A comprehensive evaluation on the typical multiple-choice methods (i.e. conventional, liberal, confidence marking) has not been seen in a single study.

IV. Signal Detection Theory

Signal detection theory is applicable to the problem as one of discriminating between two types of stimulus, a signal and noise [18]. Noise is an element which may influence subjects' detection of signal and contributes to the uncertainty of decision. The correct acceptance of a stimulus as a signal is referred to as a *hit*. The incorrect acceptance of a non-signal is called a *false alarm*, whereas the incorrect rejection of a true signal is called a *miss*. The individual's actual ability to discriminate true signals from non-signals is called *sensitivity*. The setting of the individual's accept/reject criterion level is referred to *response criterion*.

The signal detection theory framework has been applied to practical situations like sonar target detection, industrial inspection tasks, medical diagnosis, eyewitness testimony, and air traffic control [18]. McGuinness [19] applied signal detection theory for quantitative analysis of situational awareness with true/false probes. DeCarlo [20] found that signal detection theory offers a basis for understanding rater behaviour with respect to the scoring of construct responses.

In the multiple-choice methods, human make decisions of selecting the correct answer amongst several alternatives. The application of signal detection theory for quantitative analysis of one's performance can help us understand the behaviour and strategy in the process of decision making. The analysis results would let instructors and trainers know what accounts for a candidate's hit rate. Is the candidate as good as any other at discriminating the correct answer from distracters? Does the candidate over accept distracter as correct? It seems that research should be conducted to investigate the application of signal detection theory for quantitative analysis of human performance in multiple-choice tasks.

V. Proposed Study

Based on the above literature review, it is noted that the current findings cannot identify which multiple-choice method is the best to use. The three typical multiple-choice methods have not been compared and examined in a single study. In a multiple-choice assessment, subjects are required to discriminate the correct answer amongst several alternatives. The application of signal detection theory for quantitative analysis of one's performance in multiple-choice test has never been examined. To fill the gap in the literature, there is a need to conduct a research study which more thoroughly investigates the conventional multiple-choice test, liberal multiple-choice test and confidence marking considering reliability, signal detection theory, and subjective preference. The proposed study will be conducted in a quiz in an Ergonomics course.

A. Subjects

Students who are registered for a course in Ergonomics in the City University of Hong Kong will serve as subjects.

B. Instruments

A test will be administered using three different multiple-choice methods (conventional, liberal and confidence marking). The test will have 20 multiple-choice items with five options per item. Quiz paper, answer sheet and feedback questionnaire will be designed and developed for this study.

(i) Quiz paper. The response mode and scoring rule for each test method is given and illustrated with examples at

the beginning of the quiz paper. Then the quiz paper shows 20 multiple-choice questions that are created based on lecture materials on human information processing and icon usability. Each question has five alternatives which include the correct answer and four distracters. The position of correct answer in each question will be randomized. The sequence of questions on the topics of human information processing and icon usability will also be randomized.

(ii) Answer sheet. A one-page answer sheet will be designed for subjects to answer every multiple-choice item in conventional, liberal, and confidence marking methods effectively.

(iii) Feedback questionnaire. To collect students' feedback about the multiple-choice test methods, a one-page feedback questionnaire is developed. Students are asked whether they have any previous experience with the multiple-choice methods. Then students assess whether the methods really examine their knowledge and evaluate the design, complexity, overall ease of use, and measurement fairness of each method on nine-point Likert scales. The general preference for each method is also rated.

C. Procedure

After the completion of the lectures on human information processing and icon usability, students are notified that a multiple-choice quiz will be conducted two weeks later. Students are allowed to take these two weeks for revision. The quiz will be limited to 30 minutes. The response mode and scoring rule of each multiple-choice test method are briefed with examples at the beginning of the quiz. Students will also receive a complete description on the multiple-choice test methods a week before the quiz. For each question, students are asked to pick one choice only in a conventional method. In liberal method, students are asked to select more than one choice if they are uncertain of the correct one. In confidence marking method, students assign a confidence level within 0% (no confidence) to 100% (confidence) to their best choice. Immediately following the quiz, all respondents completed a feedback questionnaire regarding the experimental methods.

D. Data analysis

In the conventional multiple-choice method, a correct answer is scored 1, and an incorrect answer is scored 0. In the liberal multiple-choice method, a correct score is scored 1, and an incorrect answer is scored -1/3. In confidence marking, the confidence level is the mark awarded if a student's choice is correct, while the corresponding negative mark is awarded for an incorrect choice. The liberal multiple-choice method and confidence marking can yield negative scores as they penalize incorrect answers. To allow meaningful comparisons amongst the multiple-choice test methods, all scores will be subjected to linear transformations prior to further analyses.

The distribution of subjects' performance on each multiple-choice method will be evaluated. Cronbach's coefficient alpha will be used to estimate the quiz reliability of each testing method. To identify which multiple-choice method would let candidates achieve better performance in the quiz, the differences amongst subjects' performance on the three multiple-choice methods will be evaluated with analysis of variance.

Signal detection theory is applied for quantitative analysis of one's performance in multiple-choice items. This proposed research is the first study explored on this matter. In a multiple-choice method, the hit rate and false alarm are found for each student first and the measures of sensitivity and response criterion for each student are then evaluated. Box plots of students' sensitivity measures in the three multiple-choice methods are created to investigate whether one candidate is as good as any other at discriminating correct answer versus distracters.

According to the signal detection theory, the positioning of response criterion by a subject results in a signal to noise ratio - beta, which is useful for describing subject's strategy of determination of signal presence [18]. A ratio greater than one corresponds to a conservative strategy and leads to fewer hits and fewer false alarms, while a ratio less than one indicates a risky strategy. With a risky criterion, the subject would have a consistent tendency to wrongly accept a distracter as the correct answer. To explore which strategy (conservative or risky) would let students achieve better performance in a multiple-choice assessment, the effect of response criterion on multiple-choice performance will be studied with students't test.

The feedback questionnaire is used for obtaining the students' evaluation of and attitudes toward the testing methods. The influence of previous experience with a multiple-choice method on the corresponding multiple-choice performance will be studied. The relationship between feedback rating and multiple-choice performance is examined by using correlation analysis. The findings would help determine 'good multiple-choice method' as perceived by the students.

E. Recommendations and applications

This proposed research is the first study to explore the application of signal detection theory for quantitative analysis of one's performance in multiple-choice items. The box plot of candidates' sensitivity measures would be an effective tool to assess whether one candidate is as good as any other at discriminating correct answer versus distracters. Candidates who have a consistent tendency to wrongly accept a distracter as the correct answer would be identified by the measure of response criterion. The response criterion effect on multiple-choice performance would be identified.

In this proposed study, the reliability of each multiple-choice method would also be determined. The candidates' evaluation of and attitudes toward the testing methods would be obtained. Overall, the research findings would provide a clearer picture on the common multiple-choice methods and facilitate people to conduct more

effective assessments in various subject areas.

VI. Conclusion

Today there are a variety of multiple-choice test methods for use in assessing subjects' knowledge and decision ability: conventional multiple-choice, liberal multiple-choice, elimination testing, confidence marking, probability testing, and order-of-preference scheme. However, current findings still could not enable examiners to identify which multiple-choice test method is the best to use. The application of signal detection theory for quantitative analysis of one's performance in multiple-choice items has never been used. A study plan on multiple choice is proposed here to fill the gap in the literature. This paper would let people be familiar with multiple-choice test methods and thus facilitate people to conduct more effective assessment tests.

References

[1] N. E. Gronlund, *How to Make Achievement Tests and Assessments*, 5th ed. Allyn and Bacon, 1993.

[2] A. Ben-Simon, D. V. Budescu, and B. Nevo, "A comparative study of measures of partial knowledge in multiple-choice tests," *Applied Psychological Measurement*, vol. 21, no. 1, 1997, pp.65-88.

[3] M. C. Rodriguez, "Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research," *Educational Measurement: Issues and Practice*, 2005, pp.3-13.

[4] M. Bar-Hillel and Y. Attali, "Seek whence: answer sequences and their consequences in key-balanced multiple-choice tests," *The American Statistican*, vol. 56, no. 4, 2002, pp.299-303.

[5] Y. Attali and M. Bar-Hillel, "Guess where: the position of correct answers in multiple-choice test items as a psychometric variable," *Journal of Educational Measurements*, vol. 40, no. 2, 2003, pp.109-128.

[6] R. B. Frary, "More multiple-choice item writing do's and don'ts," *Practical Assessment, Research & Evaluation*, vol. 4, no. 11, 1995.

[7] S. Morrison and K. W. Free, "Writing multiple-choice test items that promote and measure critical thinking," *Journal of Nursing Education*, vol. 40, no. 1, 2001, pp. 17-24.

[8] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied Measurement in Education*, vol. 15, no. 3, 2002, pp.309-334.

[9] G. Ben-Shakhar and Y. Sinai, "Gender differences in multiple-choice tests: the role of differential guessing tendencies," *Journal of Educational Measurement*, vol. 28, no. 1, 1991, pp.23-35.

[10] P. Hassmén and D. P. Hunt, "Human self-assessment in multiple-choice testing," *Journal of Educational Measurement*, vol. 31, no. 2, 1994, pp.149-160.

[11] W. B. Walstad and D. Robson, "Differential item functioning and male-female differences on multiple-choice tests in economics," *Journal of Economic Education*, 1997, pp.155-171.

[12] M. Bush, "A multiple choice test that rewards partial knowledge," *Journal of Further and Higher Education*, vol. 25, no. 2, 2001, pp.157-163.

[13] D. A. Bradbard, D. F. Parker, and G. L. Stone, "An alternative multiple-choice scoring procedure in a macroeconomics course," *Decision Sciences Journal of Innovative Education*, vol. 2, no. 1, 2004, pp.11-26.

[14] A. R. Gardner-Medwin, "Confidence assessment in the teaching of basic science," *Association for Learning Technology Journal*, vol.3, 1995, pp.80-85.

[15] M. Alnabhan, "An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test," *Social Behavior and Personality*, vol. 30, no. 7, 2002, pp.645-652.

[16] S. Jennings. and M. Bush, "A comparison of conventional and liberal (free-choice) multiple-choice tests," *Practical Assessment, Research & Evaluation*, vol. 11, no. 8, 2006, pp.1-5.

[17] S. M. Swartz, "Acceptance and accuracy of multiple choice, confidence-level, and essay question formats for graduate students," *Journal of Education for Business*, 2006, pp.215-220.

[18] M. S. Sanders and E. J. McCormick, *Human Factors in Engineering and Design*, pp. 62-65, 7th ed. McGraw-Hill, 1993.

[19] B. McGuinness, "Quantitative analysis of situational awareness (QUASA): applying signal detection theory to true/false probes and self-ratings," *The 9th International Command and Control Research and Technology Symposium Copenhagen*, 2004.

[20] L. T. DeCarlo, "A model of rater behaviour in essay grading based on signal detection theory," *Journal of Educational Measurement*, vol. 42, no. 1, 2005, pp. 53-76.

[21] A. Hobson and D. Ghoshal, "Flexible scoring for multiple-choice exams," *The Physics Teacher*, vol.34, 1996, pp.284.

[22] G. S. Frandsen and M. I. Schwartzbach, "A singular choice for multiple choice," *ACM SIGCSE Bulletin*, vol. 39, no. 4, 2006, pp. 34-38.

[23] M. Pressley, E. S. Ghatala, V. Woloshyn, and J. Pirie, "Sometimes adults miss the main ideas and do not realize it: confidence in responses to short-answer and multiple-choice comprehension questions," *Reading Research Quarterly*, vol. 25, no. 3, 1990, pp.232-249.

[24] Y. Bereby-Meyer, J. Meyer, and O. M. Flascher, "Prospect theory analysis of guessing in multiple choice tests," *Journal of Behavioral Decision Making*, vol. 15, 2002, pp.313-327.

[25] P. Davies, "There's no confidence in multiple-choice testing," *Proceedings of 6th CAA Conference*, Loughborough: Loughborough University, 2002, pp.119-130.

[26] P. Brusilovsky, M. Grigoriadou, and K. Papanikolaou, "Evidential multiple choice questions," *Proceedings of 11th International Conference on User Modeling*, 2007.

[27] D. W. Farthing, D. M. Jones, and D. McPhee, "Permutational multiple-choice questions: an objective and efficient alternative to essay-type examination questions," *ACM SIGCSE Bulletin*, vol. 30, no. 3, 1998, pp.81-85.

[28] D. W. Farthing and D. McPhee, "Multiple choice for honours-level students? A statistical evaluation," *Proceedings of the 3rd Annual CAA Conference*, 1999.

Appendix 1 Summary of multiple-choice test methods commonly used in previous studies from 1990 to 2008

Test method	**Instruction and response mode**	**Scoring rule**	**Reference**
Conventional multiple-choice test	Candidates are told to pick one of the N choices.	1 mark is awarded for a correctly chosen option while 0 mark is awarded for an incorrectly chosen one. Negative marking and normalization are used to counteract the effect of lucky guesses.	
Liberal multiple-choice test	Allow candidates to select more than one answer to a question if they are uncertain of the correct one	Hobson and Ghoshal (1996) For a 5-choice test, award 3 points for a single correct answer. Candidates who choose two answers including the correct one get 2 points; candidates who choose three answers including the correct one get only 1 point. Bush (2001); Jennings & Bush (2006) In a question with N options and one correct answer, 1 mark is awarded for a single correctly chosen option and -1/(N-1) for an incorrect one Frandsen and Schwartzbach (2006) $S_{axioms}(k,a,c) = \begin{cases} 0 & if \quad a = 0 \vee a = k \\ \log(\frac{k}{a}) & if \quad a > 0 \wedge c = 1 \\ -\frac{a}{k-a}\log(\frac{k}{a}) & if \quad a > 0 \wedge c = 0 \end{cases}$ where S_{axioms} (k,a,c): a numeric function assigns a score to a single multiple-choice question k : the number of possible options a: the number of checked options c: a Boolean indicating whether one of the checked options is the correct one	[12, 16, 21, 22]
Elimination testing	Candidates are asked to mark as many incorrect options as they can identify in a question with N options	1 point is awarded for each incorrect choice that is identified, but N-1 points are deducted if the correct option is identified as incorrect.	[2, 13]
Confidence marking	Candidates have to attach a confidence level to their best answer to each question.	Hassmén and Hunt (1994) For each question, candidates gain +10 for 'correct answer & almost guess' +27 for 'correct answer & probable guess' +37 for 'correct answer & neutral' +45 for 'correct answer & fairly certain' +50 for 'correct answer & almost certain' +5 for 'wrong answer & almost guess' -4 for 'wrong answer & probable guess' -16 for 'wrong answer & neutral' -32 for 'wrong answer & fairly certain' -60 for 'wrong answer & almost certain' Gardner-Medwin (1995) The confidence level (1, 2 or 3) is the mark awarded if their answer is correct, while 0, -2 or -6 (respectively) is awarded otherwise. Candidates are told to choose level 2 unless they are very confident (>80% chance of being right), when they should choose level 3, or rather hesitant (<67% chance of being correct), when level 1 is appropriate. Davies (2002) For each question, candidates award +5 for 'very confident & right answer' +3 for 'fairly confident & right answer' +1 for 'not confident & right answer' -2 for 'very confident & wrong answer' -1 for 'fairly confident & wrong answer' 0 for 'not confident & wrong answer'	[2, 10, 14, 23, 24, 25, 26]

Test method	Instruction and response mode	Scoring rule	Reference
Probability testing	Candidates are instructed to allocate 100 points among all the options so as to reflect their subjective probability of being correct.	The item's score is the probability assigned to the correct answer.	[2]
Order-of-preference scheme	*Complete ordering:* All N options have to be ranked according to their degree of plausibility	The item's score is N-r, where r is the rank given to the correct answer.	[2]
	Partial ordering: Candidates are instructed to select one answer, if sure. If unsure, they are instructed to rank order all feasible responses	Ben-Simon et al. (1997) The score per item is determined by (N-r) – [(m-r) / (N-r)], where N is the number of options, r is ranking awarded to the correct answer and m is the total number of options selected and ranked. Alnabhan (2002) For each multiple-choice question (with 4 alternatives), the scoring procedure was as follows. If 1 alternative was selected, award +3 for correct response included and -1 for correct response not included; If 2 alternatives were selected, award +2 for correct response included and -2 for correct response not included; If 3 alternatives were selected, award +1 for correct response included and -3 for correct response not included.	[2, 15]
Two-stem multiple-choice question / Permutational multiple-choice question (PMCQ)	A PMCQ typically has a two-part stem, and N putative answers: two of which are keys and N-2 are distracters. The two parts of the stem must ask about closely related issues.	All-or-Nothing Rule: To answer the question correctly, the candidate must match up each stem with the appropriate key.	[27, 28]