MALAYSIAN ONLINE JOURNAL OF EDUCATIONAL MANAGEMENT (MOJEM)

JULY 2013, VOLUME 1, ISSUE 2, 19 - 32
E-ISSN NO: 2289 – 4489

Multiple-Choice Question Examination Process: A Descriptive Analysis

Tsze Sun Li

Corresponding Author:

Business Communication Unit, College of Economics and Political Science, Sultan Qaboos University, P.O. Box 20, Muscat 123, Sultanate of Oman.
Phone: (968) 24141837
sunny20@squ.edu.om

ABSTRACT

Multiple-Choice Question (MCQ) is a form of assessment in which respondents are asked to choose the best possible answer(s) out of a list of options, including the True / False questions that provide only two options. Despite the common use of MCQ assessment method in education, there are some teachers who may not fully understand its examination process. The main purpose of this paper is to analyse the MCQ examination process to provide a clearer picture for the teachers in their utilization and for the educational administration for their decision making purposes on the MCQ assessment method. The Analysis begins with examination of a MCQ (True/False) exam paper, and follows by discussing the MCQ strengths and weaknesses, and ends with a proposal to the MCQ Model for improving the examination method. In examining the process, issues that involved were ; items analysis to inform teaching, concept of quality MCQ exams, correlation between teaching and learning, as well as assessment validity and reliability.

Keyword: **MCQ Exam, True-false Items, Difficulty Index, Discrimination Index, Validity and Reliability, Teaching and Learning.**

INTRODUCTION

MCQ (Multiple-Choice Question) is a form of assessment in which respondents are asked to choose the best possible answer(s) out of a list of options, including the True / False questions that provide only two options. The word "question" (of MCQ) is a misnomer because many of the items are not written as questions, but incomplete statements, analogies, etc. The structure of MCQ usually comprises a stem (beginning part of the item that presents it as a problem to test the respondent) and a set of options, which are the possible answers for the respondents to choose from. The correct answer is called the key and the incorrect answers are called distracters. (Kohe,1995) Often but not always, MCQ has only one correct answer.

This testing format is frequently used in schools, universities, and professional bodies, like medical certification in Australia. This is particularly so in the United States (Phelps, 1996) and in medical and science examinations in many countries. For example, since Kelly, the first person to use MCQ on a large scale assessment, developed the Kansas Silent Reading Test in the MCQ format in 1915, (Mathews, 2006) American colleges and universities usually test students with MCQ--some professors even test their students with MCQ every lesson or after finishing a chapter of the textbook. More recently, the Royal College of Radiology uses MCQ (Single Best Answer Questions) as the only written test of knowledge for the Final Fellowship Examination in Clinical Oncology. (Tan, 2008)

It seems that MCQ has become part and parcel of the examination system everywhere. Nevertheless, not every teacher has the skill to write MCQ exam papers that can contribute to real assessment / learning. (Burton, 2005) Some teachers may not fully understand the examination process involving MCQ. Hence, the purpose of this article is for an educational administrator to examine the MCQ examination process, beginning from an analysis of a True / False MCQ exam paper, then discussing the MCQ strengths and weaknesses, and ending with a MCQ model for improving MCQ examinations. In examining the process, issues involve items analysis to inform teaching, concept of quality MCQ exams, correlation between teaching and learning, as well as assessment validity and reliability. This investigation should deserve concern in higher education, particularly in universities where most of the teachers have not received training in education or higher education teaching because at the end of the day they are held accountable to the stakeholders.

RESEARCH QUESTIONS

This paper aims to address the following questions:

1. What are the items and topics that most students did not understand?
2. What is the discrimination power of the exam paper and items?
3. Is there any significant difference between the topics taught and topics not taught?
4. How valid is the exam?
5. How reliable is the exam?

METHODOLOGY

For the purpose of this study a quantitative research method was used to analyse the process of MCQ examination. This is a descriptive research that describes the characteristics of the existing conditions without analysing relationships among variables. It is primarily concerned with the present status quo (Best & Kahn, 2006). MS Excel spread sheet was used to enter the data—"1" for correct answer, and 0 for wrong answer for each item. Formulas including descriptive statistics and t-tests were used to yield results relevant to the research questions. Besides, the procedure to analyse the level of item difficulty and discrimination power was followed.

The MCQ Exam Paper

The exam paper to be analysed is a final summative assessment for a non-native English Public Speaking course in a university where English is used as the medium of instruction. There are 40 true-false (T/F) items, weighting 20% of the course. Thirty-two students took the exam in two hours.

The exam paper comprises five topics: (1) Public Speaking in general, (2) Outlining, (3) Method of Delivery, (4) Use of Words, (5) Use of Visual Aids, and (6) Persuasion. There are 7, 5, 5, 15, 5, and 3 T/F items respectively. (See Appendix) All of the topics are taught except Topic 5.

RESULTS

Question 1: Which are the items and topics most students did not understand?

Table 1 below shows the difficulty index of all the 40 items with the 10 most difficult items highlighted. The index ranges from 0.19 to 1. For the 10 most difficult items, at least 40% of students answered incorrectly. For the most difficult item (Item 27), 81% could not answer correctly.

The discriminative power of this question, like the least difficult items (Items 1, 2, 5 and 20) with all correct answers, is extremely weak, and most likely to be discarded in a good test. (Discrimination will be illustrated later.)

Table 1
Difficulty Index

Item No.	Correct Answers (n=32)	Difficulty Index (p)	Descending Order of Difficulties
I27	6	0.19	1
I12	7	0.22	2
I28	9	0.28	3
I19	10	0.31	4
I30	11	0.34	5
I3	13	0.41	6
I8	14	0.42	7
I10	18	0.56	8
I21	18	0.56	8
I29	19	0.59	9
I18	20	0.63	10
I24	20	0.63	10
I9	21	0.66	11
I40	22	0.69	12
I34	23	0.72	13
I35	23	0.72	13
I7	24	0.75	14
I15	24	0.75	14
I16	24	0.75	14
I22	24	0.75	14
I4	25	0.78	15
I13	25	0.78	15
I26	25	0.78	15
I37	25	0.78	15
I6	26	0.81	16
I14	26	0.81	16
I33	26	0.81	16
I38	26	0.81	16
I11	27	0.84	17
I17	28	0.88	18
I31	28	0.88	18
I32	28	0.88	18
I25	30	0.94	19
I36	30	0.94	19
I39	30	0.94	19

I23	31	0.97	20
I1	32	1	21
I2	32	1	21
I5	32	1	21
I20	32	1	21

In addition, for the topics, students did not understand 80% of items in Topic 5 (Use of Visual Aids). This takes up 40% of the 10 most difficult items. For each of the three Topics—Outlining, Method of Delivery, and Use of Words, students did not understand 20%, whose aggregate comprises 50% of the 10 most difficult items, where Topic 1 Public Speaking in General takes up the rest of 10%. None of the items in the topic Persuasion falls into the top 11 most difficult items; its average difficulty level is 0.81.

Question 2: What is the discrimination power of the exam paper and items?

If a test and an item measure the same understandings, it would be expected that those having high test scores would have a high probability of being able to answer the item, and those having low test scores would have a low probability of being able to answer the item. Hence, a good item should discriminate between those with high and low scores--the higher an item's discrimination index "D", the better discrimination power it has.

For this analysis, the "D" is calculated as follows:

GA correct answers – GB correct answers

N largest group where:

Di= Discrimination index of item i

GA= Number of correct answers to item i among the 25% of those with highest test scores

GB= Number of correct answers to item i among the 25% of those with lowest test scores

N largest group= Number of persons in the largest group (GA or GB)

It is found that the average discrimination power for the whole test is 0.20. The following table shows the discrimination index by items.

Table 2

D-values by Items

I1*	12	I3	I4	I5	I6	I7	I8	I9	I10
0	0	0.38	0.38	0	0.38	0.38	0.13	-0.13	0.38
I11	112	I13	I14	I15	I16	I17	I18	I19	I20
0.4	0.3	0	0.4	0	-0.13	0.25	0	0.38	0
I21	122	I23	I24	I25	I26	I27	I28	I29	I30
0.13	0.25	0.125	0.38	0.25	0.38	0.13	0.75	0.25	-0.38
I31	132	I33	I34	I35	I36	I37	I38	I39	I40
0.38	0.38	0.5	0.38	0.38	0.25	0.38	-0.13	0	-0.25

* I for Item (i.e. I1 for Item 1, I2 for Item 2, etc.)

Ebel and Frisbie (1991) suggested the following rule of thumb for the quality of an item, and provided recommendations for the exam paper writer.

D=	Quality	Recommendations
>0.39	Excellent	Retain
0.30 – 0.39	Good	Possibilities for improvement
0.20 – 0.29	Mediocre	Need to check / review
0.00 – 0.20	Poor	Discard or review in depth
< -0.01	Worst	Definitely discard

Quality Item Indicators (Source : Ebel and Frisbie , 1991)

According to the above rule, there are 16 items (40%) that need to be discarded or reviewed in depth, 4 mediocre items (10%), 20 good or excellent items (50%). Overall, the test paper has much room for improvement.

Question 3: Is there any difference between the topics taught and topics not taught?

Table 2 above has already shown that 80% of the items in Topic 5 (not taught) fall in the category of the 10 most difficult items, compared to a maximum of only 20% for other Topics. A "paired" t-test between the two groups shows the P-value of 9.76967E-08. (Table 3) That means it is actually 0.000000--much lower than 0.05. The difference is statistically significant.

Table 3
Comparison of Topics Taught and Not Taught

	Mean	Median	Standard Deviation	Variance
Topic 5 (Not Taught)	43.8	50	24.1	579
Other Topics (Taught)	75.3	77.2	9.0	80.4
TTEST	P=9.76967E-08 (1 tail, type 1)*			
*When one group is predicted to be lower than the other, use 1 tail. Type 1 is used for different treatments for the same group.				

Question 4: Is the exam valid?

For education, validity means what ought to be assessed. In this case, it should be all the learning about public speaking, at least all the leaning taught or covered in the textbook and nothing more or less. The important requirement is all the relevant aspects of public speaking, including concepts and skills. That is to say, if a student scores Distinction (A or A-), he or she has a good understanding and ability in public speaking. Therefore, omitting relevant aspects would pose a threat to validity.

From the above understanding, to infer the exam is valid, the exam paper must be a scientific sample of the population of all learning points taught or covered in the textbook, assuming that it is to assess only the conceptual part (knowledge) of a practical course Public Speaking.

The test comprises five chapters or topics--Public Speaking in General, Outlining, Method of Delivery, Use of Words, Use of Visual Aids, and Persuasion. There are 7, 5, 5, 15, 5, and 3 T/F items respectively. Assuming these are the conceptual areas to be assessed, a teacher needs only some professional knowledge or careful reading of the textbook to conclude that the Topics are not representative of all the conceptual knowledge in the textbook, not to mention in the discipline. The crux of the issue is not so much that the test items (learning points) are not valid, but it is not a good representative sample of what ought to be learned.

Question 5: Is the exam reliable?

Reliability refers to how well student performance is assessed. If the test as a means of measurement is reliable, similar tests following the same procedure would yield the same result. In this case, timing, disciplinary control, accuracy in phrasing the items, and the issue of validity would come into play.

1. Timing. The students were given sufficient time to complete the test, which is much more than 1 minute per item.

2. Disciplinary control. Invigilators were deployed to the test venue to keep the house in order. As there was only one set of test paper for all students, and they were sitting quite close together, there might be some chance of cheating that would affect reliability.

3. Accuracy in phrasing items. A careful screening of the test paper, wrong answers in particular, has uncovered some potential weaknesses or problems. The following are three examples.

Item 15 (8 students did not answer correctly)

Original Question	If you want to sound eloquent, you should use words that are unfamiliar to the audience. (Suggested Answer: False)
Problems	As the students are non-native English users who may not know the precise meaning of "eloquent," using this word may obscure the point that unfamiliar words should not be used in public speaking.
Revision	If you want to deliver your speech well, you should use words that are unfamiliar to the audience. (False)

Item 27 (26 students did not answer correctly)

Original Question	One of the advantages of using visual aids in a speech is that their meaning is instantly clear to the audience. (Suggested Answer: False)
Problems	Similar to Item 15. The meaning of "instantly" may be unclear, particularly to the students whose English is not good. There is no reason to penalize students whose English is not good because this is basically a conceptual rather than English test.
Revision	One of the advantages of using visual aids in a speech is that their meaning is clear to the audience at the moment the visual aids are presented. (False)

Item 30 (21 students did not answer correctly)

Original Question	A visual aid is only as useful as the explanation that goes with it. (Suggested Answer: False)
Problems	It is rather tricky to use the word "only" in the item construction. The meaning is unclear. Is it saying "A visual aid is not as useful as the explanation that goes with it?" The meaning of "only", just like other specific determinants such as all, none, always, might, can, generally, sometimes, etc, must be used with caution because the meaning is ambiguous. In fact, the usefulness of a visual aid is not self-evident; it has to be accompanied with explanation to be useful.
Revision	A visual aid, when used with or without explanation, is better than the use of explanations without visual aids. (False)

Reliability and validity are inter-related. Not only accuracy (e.g., phrasing of the items as illustrated above) affects reliability, but also scientific sampling of the aggregate of knowledge. As a rule of thumb, the larger the sample, the more reliable it is. In this case, the 40-items treatment is problematic in two important ways. Firstly, it is too small to represent six topics, not to mention the whole course, for a university high-stake summative assessment. According to Pamphlett (2005), it takes one hundred (100) True / False items in a MCQ medical examination. Secondly, and more importantly, the items selection does not follow a scientific cluster sampling procedure. Therefore, it is not sure that the result will be the same if the exam is conducted again.

DISCUSSIONS

MCQ Strengths and Weaknesses

There are pros and cons for using MCQ for assessments. First of all, if the teachers are well trained and items are quality assured, it can be an effective and efficient assessment technique. It is generally accepted in education that MCQ exams require less time to administer for a comprehensive coverage of subject matter than exams requiring long written answers. This is particularly so if the exam involves a very large of number of students. Aided with online testing techniques and software, MCQ exam can be even more efficient.

Moreover, MCQ items in general are suitable for assessing well-defined knowledge and/or lower-order skills, as for example, who the author of the Public Speaking textbook is, by giving the names of four textbook authors. It does not make sense to rule out this kind of basic "recall questions" that assess knowledge of definitions or isolated facts, particularly for freshmen or sophomores exams. Due to the relative objectivity of MCQ, factors irrelevant to the subject such as handwriting and text organizational skills will not affect the students' performance.

Some argue that MCQ is not as good as other types of questions (e.g., essays, short written response, etc.) in testing high-order reasoning skills such as critical and creative thinking. This might be true to some extent. Nevertheless, in the academy, particularly medical education, many scholars disagree that MCQ cannot measure high-order thinking skills. They have developed test banks with items for testing problem-solving and decision-making skills in relation to diagnostics of medical problems. They claimed the MCQ questions are designed to measure the breadth and depth of knowledge of the examinees. (Tan, 2008) Although these scholars in medical schools are referring to the MCQ format with four options rather than T/F format with two options, recent studies (e.g., Mobalegh, 2012) have shown the format of multiple true-false questions (MTF) is not only preferred by examinees, but as reliable as the multiple-choice (MC) format.

However, in most universities teachers or textbook developers set the test items. If they are not well trained, MCQ can be subjective with the creator's bias. For example, the teacher / developer may favor mostly negative questions / answers. Other typical examples include writing short stems but long distracters; presenting the key longer, and more specific than the distracters; allowing the key and distracters to fall into different categories; using absolutes such as always, never, all, etc.; using vague words such as frequently, usually, and so on.

Even if training has been provided, there can still be item ambiguity to some extent. This might be due to the discrepancy of item interpretation between the developer and the examinees. Furthermore, unlike written responses, MCQ usually do not consider partial understandings of the items, nor the chance of guessing right. To prevent guessing, therefore, some MCQ exams favor negative marking for wrong answers.

All in all, despite MCQ weaknesses, MCQ (T/F) exams can be improved.

CONCLUSION

From the findings of the above analysis, the following can be concluded.

1. An items analysis of an MCQ exam can yield useful knowledge to inform the educational administrator how well the teacher has taught the subject (e.g., revealing difficult learning points, identifying potentially flawed items, etc.), thus helping the teachers in evaluating the quality of their exam questions. This confirms the findings of Chiavaroli (2011) in his Biology post-exam item analysis.

2. Some teachers may not know how to write MCQ exam paper / items that would measure students' understanding of key concepts (knowledge). Therefore, educational administrator may arrange assessment workshops to help teachers design exam items; it is particularly important to increase their awareness on avoiding unclear / ambiguous / difficult language usage, particularly in non-native English universities using English as the medium of instruction.

3. As correlation between teaching and student learning is confirmed, a MCQ exam paper should measure, as far as possible, what the teacher has taught, rather than concepts that have not been taught.

4. When the exam paper is not developed scientifically, it is not only invalid, but also unreliable.

RECOMMENDATION TO MCQ EXAM QUALITY ENHANCEMENT

From the above conclusions, it is clear that if the purpose of the exam is to measure understanding of concepts in high-stakes assessments, MCQ must be scientifically developed to meet the requirements of validity and reliability.

One way to produce a scientific exam paper sample is that professional teachers of the subject collaboratively write comprehensive MCQ questions pertinent to the key concepts that students should understand in the subject, basing on relevant textbooks / test banks, and more importantly professional knowledge. If one individual teacher cannot complete the task, several teachers in collaboration with each one responsible for some chapters / topics (cluster sampling) would. If possible, professional teachers may join venture with textbook publishers in producing comprehensive high-quality MCQ test banks.

University teachers, without Doctor of Education degrees, must receive training in higher education teaching to make up for their lack of educational theories / knowledge relevant to assessment—both high-stakes summative assessment of learning and formative assessments for learning.

Assessment is a very important issue in higher education, especially truth-seeking being a central value for any university. For accountability, teachers cannot freely or wrongly label students as good, average, and below average ones, which is quite commonplace in some tertiary institutions. What's more, education administrators / leaders cannot be exempted from understanding academic or educational issues like assessment. The days of "(subject) outsiders ruling over insiders" (subject experts) should be gone.

To accomplish the huge task (MCQ model test bank) recommended above, Goldwater (2005) has proposed a method called "distorted-items" (DI) tests. That is, teachers may excerpt sentences or phrases from the textbooks (in my view, these are important references that should not exclude teachers' professional knowledge and scholarship of teaching) and then distort them to make them false. In so doing, the "whole territory" or a very large pool of conceptual knowledge would be available for sampling. More importantly, in so doing, students are expected to engage themselves in reading and understanding comprehensively on the subject rather than merely a textbook or even a few chapters of it. With education accountability, universities can then produce competent learners in their respective disciplines.

REFERENCES

Best, W.J. and Kahn, V. (2006). Research in education (11th ed.). Pearson Education Inc.

Burton, R. F. (2005). Multiple-choice and true-false tests: myths and misapprehensions. Assessment & Evaluation in Higher Education, 30:1, 65-72.

Chiavaroli, N. and Familari, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs. Bioscience Education, 17.

Ebel, K, L. and Fribie, D. A. (1991). Essentials of educational measurement (5th ed.) NJ: Prentice-Hall

Goldwater, B. C., Grabavac, D. M. and Acker, L. E. (2005). The distorted-item test: A true-false test with a difference. International Journal of Testing, 5:1, 75-80

Kehoe, J. (1995). Writing multiple-choice test items. Practical Assessment, Research & Evaluation, 4(9). Retrieved from http://pareonline.net/getvn.asp?v=4&n=9

Mathews, J. (2006). Just Whose Idea Was All This Testing? The Washington Post, November 14.

Mobalegh, A., and , Barati, H. (2012). Multiple true-false (MTF) and multiple-choice (MC) test formats: A comparison between two versions of the same test Paper of Iranian NUEE. Journal of Language Teaching and Research, 3(5), 1027-1037. Doi:10.4304/jltr.3.5.1027-1037

Pamphlett, R. (2005). It takes only 100 true–false items to test medical students: true or false? Medical Journal, Vol. 27, No.5, pp.468-472.

Phelps, R. (1996). Are US students the most heavily tested on earth? Educational Measurement: Issues and Practice, 15 (3):19–27. Doi:10.1111/j.1745-3992.1996.tb00819.x

Tan, L.T., and McAleer, J.J.A. (2008). The Introduction of Single Best Answer Questions as a Test of Knowledge in the Final Examination for the Fellowship of the Royal College of Radiologists in Clinical Oncology. Clinical Oncology, Vol. 20, Issue 8, pp. 571–576.

Appendix: Public Speaking Final Exam

Decide "true" or "false" for the following 40 statements. Check (√) the better answer on the answer sheet provided.

1. When making a preparation outline, you should state your main points and sub-points in full sentences to ensure that you develop your ideas fully.
2. You should include your specific purpose statement with your preparation outline.
3. According to your textbook, the introduction, body, and conclusion should all be labeled in a speech preparation outline.
4. Stating main points in a word or two is usually sufficient for a preparation outline.
5. "Dogs" would be an appropriate title for a speech to inform an audience about the major breeds of show dogs.
6. A speaking outline is usually longer and more detailed than a preparation outline.
7. The speaking outline is a more complete version of the preparation outline.
8. It is seldom necessary to write out quotations in full in a speaking outline.
9. You should label the body and conclusion in both the preparation and speaking outlines.
10. Main points must be written out in full sentences in both the preparation outline and the speaking outline.
11. If the meaning of a word is clear to you, you can assume that it is also clear to your audience.
12. The denotative meaning of a word includes all the feelings, associations, and emotions that the word touches off in different people.
13. The connotative meaning of a word is more variable, figurative, and subjective than its denotative meaning.
14. A speech dominated by abstract words will almost always be clearer than one dominated by concrete words.
15. As your textbook explains, if you want to sound eloquent, you should use words that are unfamiliar to the audience.
16. Concrete words add to the imagery of language use by creating sensory impressions.
17. "She darted around the bookstore like a hummingbird in a flower garden" is an example of metaphor.
18. "Ask not what your country can do for you; ask what you can do for your country" is an example of antithesis.
19. Antithesis and alliteration are excellent ways to enhance the imagery of a speech.
20. Language needs to be appropriate to a speaker himself or herself, as well as to the audience, topic, and occasion.
21. Using the masculine pronoun "he" in a speech is an inclusive way to designate "all persons."
22. One of the advantages of speaking from a manuscript is that it frees a speaker from the need to establish eye contact with the audience.
23. When speaking impromptu, you should do your best to look calm and confident no matter how nervous you may be feeling.
24. Speaking impromptu and speaking extemporaneously are essentially alike.

25. A faster rate of speech is usually called for when a speaker is explaining complex information.
26. Research has shown that an average speaker who uses visual aids will come across as more credible and better prepared than a speaker who does not use visual aids.
27. One of the advantages of using visual aids in a speech is that their meaning is instantly clear to the audience.
28. Because a picture is worth a thousand words, it is a good idea to pass photographs among the audience in order to illustrate your point.
29. According to your textbook, it is usually a good idea to write or draw on an overhead transparency while you are speaking.
30. A visual aid is only as useful as the explanation that goes with it.
31. "To inform my audience how to create their own Web pages" is a specific purpose statement for an informative speech about a process.
32. Clear organization is less important in speeches about processes than in other kinds of informative speeches.
33. One of the major barriers to effective informative speaking is overestimating what the audience knows about the topic.
34. As your textbook explains, technical language is especially helpful for explaining ideas in informative speeches.
35. Abstractions are especially helpful for clarifying ideas in informative speeches.
36. One reason to use clear and straightforward language even when talking about complex ideas is that listeners must understand your message in the time it takes you to say it.
37. Whenever possible, you should try to enliven your informative speeches by expressing ideas in personal terms.
38. Persuasion is the process of creating, reinforcing, or changing people's beliefs or actions.
39. Of all the kinds of public speaking, persuasion is the most complex and the most challenging.
40. The target audience is that portion of the whole audience that the speaker most wants to persuade.

(END OF TEST)