In the Name of Him, the Most High

دانشگاه علوم پزشکی و خدمات بهداشتی درمانی
شهید صدوقی یزد

Vice Chancellor for Education, Shahid Sadoughi University of Medical Sciences and Health Services, Yazd

Center for Studies and Development of Education of Medical Sciences

---

Guidelines and Principles for Designing Multiple-Choice Questions in Medical Science

Excerpts from Millmen Indices

Faculty Members' Questions Evaluation Committee of the Center for Studies and Development of Education of Medical Sciences

* In this guide, a brief overview of the essential and practical points in examining multiple-choice questions is mentioned.

How to design better questions?

Designing a question is not a science but a set of rules based on the experience and wisdom of professors and examiners.   [Jan Bromot (1970)]

Question Writing Rules

-Tips related to the general stages of question writing .

-Execution of base/stem/body of the Question .

-Designing choices and general points .

-Notes on the correct option.

-Designing distractors .

General Stages of Question Writing

-Avoid designing integrated and complex options that sometimes take into account more than one point.

-Arrange the options vertically and below each other (for multi-word options), especially for children and adults.

-Horizontal arrangement of options (for short options) to save on paper and reproduction costs.

-Reducing the likelihood of accidental errors increases the validity of the test.

Avoid grammatical, punctuation, spelling and acronyms/abbreviations

-Creating an inappropriate image of test designers and developers in the viewer's mind.

-Failure to take the test seriously undermines the validity of the results in a double

way.

-Lack of proper preparation and planning for test preparation.

-Possibility of inappropriate sampling.

-Diverting the subject's attention from the content.

-Presence of a type of error in test scores.

Avoid asking misleading questions:

-Providing trivial and unimportant content.

-Not distinguishing between testers.

-Painting and glazing the question (showcase).

-Providing more than one correct answer.

-Asking about principles we have not taught.

-Ambiguity and mental confusion of a group of students.

-It is not intended to deceive, confuse or mislead test takers.

General Content of Questions

-Each question addresses an educational goal based on a course description table, or a two-dimensional table (In professional licensing tests, it is the reflection of the desired behaviors for the relevant expert).

-Each question focuses on the actual data or concept, principle or successive steps of performing a practical technique.

-The wording and writing style in each question are appropriate for the target group (the purpose of the test is not to measure the language proficiency of the participants.)

-The content or answer of a question does not lead the participant to find the

correct answer to another question in the same test.

-Test items are completely independent of each other.

By modeling successful questions in previous exams and using them as a question format, we will have a more appropriate design.

-Refraining from using the same phrase or sentence of the textbook.

-Avoiding memorizing points like a parrot.

-Presence of statistical bias on the basis of memorization power of the participants.

-Refraining from designing questions that deal with very specific knowledge.

-Avoiding items that deal with highly abstract issues.

-Refraining from asking questions based on personal opinions .

-Emphasizing higher levels of thought.

-Questions should address the facts, concepts, principles, and steps of performing techniques that are based on public evidence.

Item Stem

The clearest form of an item stem is a question.

-Easy to read.

-Shorter time.

-The least amount of pressure and anxiety.

-The best quality.

-When writing an incomplete sentence, never put a space at the beginning or in the middle of the sentence!

-The blank is always placed at the end or final part of the question.

After writing the base of the question, read it and make sure that the subjects understand exactly what the question asks of them.

-Refraining from unnecessary words.

-Refraining from providing data that has no role in solving the problem (circumlocution, verbosity, prolixity).

-Refraining from writing the base of the question in a negative way.

-Rewrite the question as a few true-false questions.

-Be sure to highlight the negative sign in some way (bold-face letters, underlining the letters and using different letters and larger sizes).

-Always try to state the main idea and the main part of the question in the base/stem of the question.

Options/Distractors

-The more effective distractors, the better (reducing the likelihood of guessing the correct answer).

-It is very difficult to design more than one or two distractors.

-The performance of distractors is important.

-Avoid showcasing and paying attention only to the number of options.

-No interference with each other.

-The different options of each question are the same in terms of content.

-The length of the options should be similar to each other.

-Refrain from writing options such as "both A & B", "I do not know", "None of the above options" or "a, b, and c".

-The options should be positive in terms of form and meaning, if possible.

-Whenever you have to use a negative sign or words like "except or unless", highlight them clearly.

-Unwanted guidance to the correct answer (by sound or by meaning).

-Avoid designing ridiculous options.

-Whenever you use the base of an incomplete question, each of the options must be grammatically placed in the blank to obtain an acceptable language sentence.

-Refraining from using absolute constraints (absolutely, completely, in general, never, always, etc.)

The Correct Option

-Only one correct option.

-Balanced distribution of correct answers among the choices (Random Selection of Answer Option). Use almost identical number of correct answers marked with each of the letters A, B, C, D, and so on.

Distractors

-Always use justified and good-natured distractors.

-Common mistakes of students.

-Use of technical terms, using terms that are correct in themselves but can not be the desired answer.

-Refrain from ridiculous options.

Check the Millmen questions for a qualitative review

1. Does each question address a key issue based on the test budget?

2. Is each question clear and as short as possible?

3. Is the stem/base of the question not negative, and if it is inevitably negative, is the negative sign prominent in a way?

4. By covering the options, can the basis of the question alone be understood or completed?

5. If it has more than one correct answer, are compounds like "the best" used?

6. Is the blank not at the beginning but in the final sections of the basic question?

7. Does each option test only one point?

8. There is no grammatical problem at the base of the questions and options, and does the combination of each option and the base of the question provide an accepted linguistic form?

9. Has repetition of recollections been avoided based on the question?

10. Are distractors interesting for uninformed subjects?

11. In terms of length, complexity and practical points, are the options appropriate and have the common words in all options been transferred to the question?

12. Are the question base and options appropriate in terms of thematic content and form?

13. Is the number of options in each question is the same ((3, 4, 5, or 6)?

14. Have you refused to use the same word of the textbook in the item?

15. Have you ever mentioned a concept different than other options?

16. Has the use of absolute constraints (always, never, only ....) been avoided?

17. Does the term or concept that needs to be defined come in handy and do the options complement its characteristics?

18. Does the distribution of correct answers between different options follow any particular pattern and is it merely random?

19. Is the agenda clear for each part of the test?

20. In each part of the test, is the arrangement of all the options in one direction (vertical/horizontal with appropriate distance)?

21. Is each question independent of the others?

22. Has the use of all/none of the cases or "Both A and B" and the like been refused?

23. Is the base of the question and the relevant options on one page?

24 . Are the distractors opposite to the correct answer not easily removable?

25. Aren't similar or guiding words based on the question and not in any of the common options?

27. At the end of the question, there is no correct answer to another question?

Criteria for Selecting Questions

Top questions should meet the following criteria by recognizing the question selection committees at the university and secretariat levels:

-Taxonomy (level of knowledge used in the question)

-Selected questions from Bloom's cognitive level must necessarily be posed in taxonomy 1, 2, and 3 (Bloom Cognition Levels).

-Three of Bloom's six levels are applicable to multiple-choice questions (MCQs).

-Taxonomy I (Recalling): Questions based on memory and superficial knowledge: Example: What are the most common manifestations of X disease?

-Taxonomy II (Comprehension): Questions based on understanding and comprehension: For example, when a case is introduced and diagnosis is questioned, taxonomy II is used.

-Taxonomy III: Application of knowledge: Items based on the application of knowledge. For example, when a case is introduced in the question and the most probable treatment is asked, the examinee must first diagnose the disease or problem and then find the best possible treatment for the case. In this state, taxonomy III is used.

-In postgraduate clinical tests, most items (more than %80) should be developed using Taxonomy II and Taxonomy III.

-As a general rule, not more than %50 of the test items are allowed to use memory-based items using Taxonomy I.

Observance of Structural Rules of MCQs (Millmen Principles)

To guarantee the construct validity of MCQs, Millmen's principles about stem and options of MCQs have been developed the observance of which has been agreed upon by all experts in the field of achievement tests of residents promotion.

No.	Subject	Yes	No
1	Is the greatest part of information entailed in the stem?		
2	Does the item examines a particular learning objective?		
3	Is the wording used in the stem and options clear and straightforward?		
4	Has the use of negative signs in the stem been avoided?		
5	Has the use of "all of the above", "none of the above", "both a and b", etc. been avoided?		
6	Has the use of opposite choices been avoided?		
7	Are positive words used in the stem or if negative signs are used, are they highlighted in a way or the other?		

8	Are the test items developed independently from each other?		
9	Are the options matched for length, lexical structure, and writing style?		
10	Are repetitious phrases are avoided in the item as far as possible?		
11	Is the spelling of words used in the stem and options correct?		
12	Are the options arranged vertically?		

The selected items should have observed Millmen principles. Some items are elucidated below.

Item 1 of checklist: The bulk of information ought to be included in the stem and not in the choices. On this basis, developing MCQ items such as the following is not appropriate in terms of constructive validity.

-Regarding ……., which of the following is correct/wrong?

Item 2 of checklist: In an academic achievement test, it is better for an MCQ to investigate the SOB. In other words, it not proper to intermingle distractors from immunology, diagnosis, physiology, etc.

Item 3 of checklist: The diction of the stem and distractors should be clear and straightforward. For example, phrases like "better" or "more appropriate" should be avoided to ward off ambiguity.

Item 4 of checklist: If the stem is negative, no negative choices should be used (double negative). For example, the following item is not acceptable in terms of MCQ principles:

-A 5-year-old boy has presented this morning with fever, vomiting, and headache. Examinations show…… Which of the following is NOT correct for this patient?

a. …….

b. …….

c. For this patient, meningitis is NOT diagnosed.

d. ....

Item 5 of checklist: In writing MCQs, word strings like "all of the above, none of the above, both a and b, etc." should not be used.

Item 6 of checklist: Two opposing terms should be avoided in choices. For instance:

-A 40-year-old man presents with ...... Which of the following is correct regarding the paraclinical findings of ......examination of the patient?

a. ....

b. Serum sodium level is normal.

c. Serum sodium level is not normal.

d. ....

Item 7 of checklist: If a negative word or a term with negative weight is used in the stem, it should be highlighted by bolding, italicizing, underlining, or capitalizing.

Item 8 of checklist: Items should be developed independent of each other; i.e., not more than one item should be based on a given case.

Item 9 of checklist: The length of the choices ought to be almost equal and their grammatical structure should also be similar as far as possible. So, items with one or two short and one or two long choices or with items with figures and numbers in some distractors but not in others are not appropriate regarding principle of test development. For instance:

-A 20-year-old woman presents with dermal lesions and...... Which of the following is correct regarding this patient?

a. The selected drug is steroid.

b. In the case of negative result of Double-Stranded DNA Test, investigation of other criteria should be considered.

c. …..

d. ……

In this item, choice b is very long and it is better to write all choices in English.

Item 10 of checklist: Repetitious phrases should be avoided in the stem and choices. For example:

-A 4-year-old boy, weight……., presents with …….findings. Which of the following is correct?

a. The rate of Creatinine clearance is …….

b. The rate of Creatinine clearance is …….

c. The rate of Creatinine clearance is …….

d. The rate of Creatinine clearance is …….

The stem should be modified in this way:

-Which choice shows the rate of patient's Creatinine? Numbers should be given in the choices.

Items with Positive Discrimination Index

Item discrimination index indicates the degree to which the item has been able to discriminate among the stronger and weaker students with a range of -1 to +1. In other words, in an ideal situation, the ID of an item is positive if more stronger students have given correct answer to an item. Thus, an item is not appropriate if both strong and weak students have answered it correctly to the same degree (ID=0) or if the weaker students have answered it correctly to a greater degree than the stronger students (ID<0). Hence, the ID of an item will be greater if more stronger students with higher scores in the test have answered it correctly compared to weaker students.

The greater the number of stronger candidates who answered the question correctly than the number of weaker candidates who answered the question correctly, the

greater the numerical coefficient. Each question must be answered more correctly by the candidates who scored higher on the whole test.

Item Difficulty Index

The difficulty index of each question is the number of people who answered the question correctly. It is represented with P. For example, if 67% of the participants answered the question correctly and with the test letter, the difficulty coefficient of the question will be 67%. The simpler the question (the more people have answered it), the higher the difficulty index: The point to note is that the difficulty index of the question is a test-dependent property. A question that has a difficulty index of 55% in one test, may have a difficulty factor in another subject (with 0 participants/another).

The Relationship between the Item Difficulty and Item Discrimination

The simpler the question and the more volunteers answer it correctly (the higher the difficulty factor), the smaller the difference between the strong and the weak groups and the item discrimination will go down. Similarly, when the question has a low item difficulty and a large number of people have not been able to answer it correctly, the difference between the strong and weak groups will still decrease and the item discrimination will decrease, too.

Validity and Reliability of the Test

The reproducibility of the measurement results or the similarity of the measurement is called reliability and the accuracy of the measurement is validity. Validity of the test means that the test can measure what it is supposed to measure. For example, if an individual's skills and knowledge in a specific area is to be tested in the residents admission test, the test is valid if it measures the skills and

knowledge of the test in the person who is more knowledgeable. If the test has validity, it discriminates the person with more knowledge and skills than others who do not have enough skills and knowledge in all or some areas of the field under measurement.

This definition indicates that if the purpose of the test is to assess a person's clinical abilities, the design of theoretical and non-practical questions reduces the validity of the test, or if the test is such that a person with other abilities than knowledge such as intelligence or logic will be able to answer the questions, that test has sufficient validity. Also, the test should be "comprehensive and obstructive". Also, the test should cover as much as possible all the topics that should be questioned and avoid topics that are not within the scope of the test objectives.

Reliability

Reliability means repeatability of results. This concept in the test means that if the test is repeated with the same participants and conditions, the people who got higher scores in the first test will have higher scores in the repetition of the test and the total scores of the people will remain largely constant. In other words, in order for the test to be reliable, if the participants' scores were obtained in the second test performance, their scores should be positively correlated with each other.

Naturally, the stronger this correlation, the more reliable the test. Different tests are prepared for different purposes. There are different credentials, and depending on the purpose for which the test was created, its reliability is determined according to that purpose.

Factors Affecting Test Reliability

1. Number of questions: The more questions, the higher the reliability.

2. Quality of questions: In the number of questions equal to the test, the quality of the questions will be more reliable.

3. Type of variable to be measured: The reliability of tests that measure a simple variable such as maintaining a multiplication table is higher than the reliability of tests that measure complex variables, e.g., clinical competence as clinical adequacy.

4. Differences in the academic level of the candidates.

5. Test conditions

Sources :

Presented at the meeting of power point committee members:

-Sabouri Kashani, Ahmad Specialist in Basic Sciences of the Ministry of Health and Medical Education, 2008.

-Educational Measurements and Principles of Designing Multiple Choice Test Questions in Clinical Sciences, Secretariat of the Medical and Specialized Education Council, 2000.

-Azizi, Fereydoun: Education of Medical Sciences, Challenges and Perspectives, 2003.