Designing effective multiple-choice questions for assessing learning outcomes

UDC 371.275
DOI 10.18485/infotheca.2018.18.1.2

ABSTRACT: Nowadays, multiple-choice question tests (MCQ tests) present a popular tool for assessing learning outcomes because they are flexible, relatively easy to implement and grade, and able to assess large content in a short time. More important is the fact that these questions are versatile and reliable, which increases their attractiveness. Additionally, the examiners can use question sets in their assessments that are already prepared for previous courses or are available online. On the other hand, many authors argue that MCQs are commonly used to assess cognitive skills on lower levels as defined by Bloom's taxonomy. However, the comprehensive analyses confirm that this type of assessment tool has the capacity to evaluate certain higher-order thinking.

KEYWORDS: Multiple-choice question tests, levels of cognition, difficulty index, discrimination index

PAPER SUBMITTED: 10 May 2018
PAPER ACCEPTED: 28 May 2018

Jasmina Jovanovska
jasmina.armenska@gmail.com
St. Kliment Ohridski
Faculty of Pedagogy
Ss. Cyril and Methodius
University in Skopje, Macedonia

1 Introduction

Assessment is the systematic collection and analysis of information to improve student learning (Stassen et al., 2001). The measurement of student learning through assessment is important because it provides useful feedback to both teachers (examiners) and students about the extent to which students are successfully meeting course learning objectives. It is also useful for teachers in developing the rationale for pedagogical choices in the classroom.[1] The

[1] Michael R. Fisher, Jr., "Student Assessment for Teaching and Learning", Center for Teaching, Vanderbilt University (on-line)

most widely used traditional assessment tools are multiple-choice question tests, true/false tests, short answers and essays.[2] Brown and Knight (1998) asserted that utilizing a mixture of different tools, improves the reliability of the assessment. However, this approach is quite challenging because the examiners should be able to properly weigh the scores produced by the different tools of assessment.

Nowadays, muliple-choice question tests (MCQ tests) are commonly utilized by teachers, schools, universities and assessment organizations because they present an effective and efficient way to assess learning outcomes (a detailed explanation about MCQs' advantages, as well as their limitations, is given in the following two sections). Each *multiple-choice question*, also known as *item*, consists of a given problem (known as a *stem*), and a list of suggested solutions (known as *alternatives*). The alternatives usually include one correct answer (the best alternative), as well as several incorrect or inferior alternatives, known as *distractors* (Brame, 2013). Student's task is to select the alternative that presents the best answer for the given problem. The purpose of the distractors is to appear as plausible solutions of the given problem for those students who have not achieved the objective being measured by the item. Conversely, the distractors must appear as implausible solutions for the students who have achieved the objective.

This paper is organized as follows. The present section provides an overview of the MCQs' advantages and limitations in the process of assessing learning outcomes. Section 2 is dedicated to the standard protocols for increasing MCQ's validity. At the end, the potential of MCQ tests for evaluating higher-order thinking skills is highlighted, including some recommendations how to construct such tests.

1.1 Advantages of multiple-choice questions

Multiple-choice questions tests have certain advantages, as well as limitations, just as any other type of test items. Examiners must be aware of these features in order to use multiple-choice questions effectively. Below are presented the most important advantages of using MCQ tests as an assessment tool (Burton et al., 1991; Chan, 2009; Dikli, 2003; Towns, 2014).

Versatility. MCQ tests are applicable in many different subject-matter areas and can be used to assess various levels of learning outcomes (from simple

[2] Maryellen Weimer, "Advantages and Disadvantages of Different Types of Test Questions", Faculty Focus, (on-line)

recall of knowledge to more complex levels, such as application, analysis and evaluation). However, these tests cannot be applied in each testing because students are choosing from a set of potential answers. For example, they are not an effective way to test students' ability to organize their thoughts or express their creative ideas (Section 1.2).

Reliability. Reliability is defined as the degree to which the test consistently measures the learning outcomes. Appropriately written MCQ tests are more reliable than the tests including other types of questions. For example, they are less susceptible to guessing than true/false questions. Also, their scoring is easier to understand than short-answer test scoring because there is no need to resolve partial and misspelled answers. Furthermore, MCQ test assessment is more objective than the assessment including essay questions. The essay test scores can be affected by the examiner's inconsistencies and are not immune to the influence of bluffing and writing ability factors, which can lower their reliability.

Validity. Validity is defined as the degree to which the test measures the learning outcomes it aims to measure. Because MCQ is usually answered more quickly than an essay question, tests based on MCQs can focus on a relatively broad course material, thus increasing the validity of the assessment (Bacon, 2003).

Efficiency. The usage of MCQ tests is very important for the examiners because they allow easy and quick evaluation. These tests are particularly essential for the examiners who cover multiple courses with large number of enrolments. MCQ test assessment expedites the reporting of students' results, thus allowing the examiner a quick insight of their achievements and an opportunity to give additional clarifications and instructions before the course is completed.

1.2 Limitations of multiple-choice questions

Despite the aforementioned advantages, the assessment of learning outcomes with MCQ tests is often criticized. The rest of this subsection presents an overview of the MCQ tests' limitations (Burton et al., 1991; Chan, 2009; Dikli, 2003; Towns, 2014).

Versatility. Certain researchers emphasize that the MCQ tests evaluate student's ability to memorize, rather than understand, apply and analyze information (Walsh and Seldomridge, 2006). However, it is obvious that these tests can also be used to assess higher-order thinking. This can be achived by including questions that focus on higher levels of cognition, presented in the well-known Anderson and Krathwohl's taxonomy (the revised Bloom's taxonomy) (Bloom, 1977; Anderson et al., 2001). The stem of such question types presents a problem, which can only be resolved with analysis and application of particular principles from the examined area. The alternatives can also contribute to this process, through the necessity to be evaluated. However, the process of developing MCQ tests for assessing higher-order thinking, requires more skills and capabilities than developing questions that evaluate simple recognition and memorization (Palmer and Devitt, 2007).

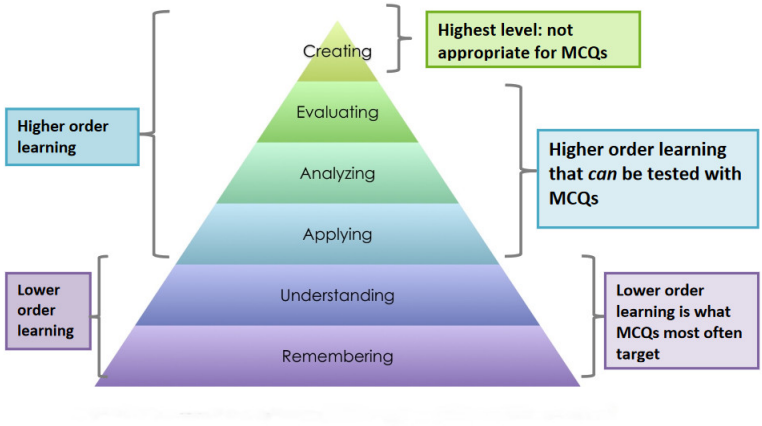

Figure 1. Suitability of MCQs to different levels of cognition of Anderson and Krathwohl's taxonomy

Figure 1 outlines Anderson and Krathwohl's taxonomy[3] for the cognitive domain, which is broken into six levels of objectives. It also includes annotation about the suitability of the MCQ tests for assessing the presented levels.

[3] Workshop: Designing Effective Multiple-Choice Questions, Teaching and Learning Services, McGill (on-line).

Furthermore, Table 1 briefly describes the levels' complexity and specificity. According to Anderson and Krathwohl's taxonomy, MCQs are not appropriate for testing only the highest level of cognition (creating). The reason is that creating requires students to put parts together in a new way, or synthesize parts into something new and different, creating a new form or product. This process is the most difficult mental function in the present taxonomy.[4]

Table 1. Description of the six levels of complexity in the Anderson and Krathwohl's taxonomy

Level	Definition
Remembering	Recalling information
Understanding	Identifying examples of a given term, concept, or principle. Interpreting the meaning of an idea, concept or principle.
Applying	Using information, rules and procedures in concrete situations.
Analyzing	Breaking information into parts to explore patterns and relationships. Analyzing charts, data to support conclusions.
Evaluating	Justifying a decision or a course of action.
Creating	Generating new ideas or products.

Reliability. MCQs are less susceptible to guessing than true/false test items, but they are still affected to a certain extent. The guessing factor reduces the reliability of MCQs scores somewhat, but increasing the number of items offsets this reduction in reliability.

Table 2 presents the probabilities of scoring 50% or higher on MCQ test by blind guessing the correct answers (P). The results are obtained using the Binomial distribution:[5]

[4] Leslie Owen Wilson, "Anderson and Krathwohl – Bloom's Taxonomy Revised", The Second Principle (on-line)

[5] Ronald E. Walpole et al., *Probability & Statistics for Engineers & Scientists*, Prentice Hall (on-line)

Table 2. The probability of scoring 50% or higher by blind guessing, depending on the number of MCQs

Number of 4-alternative multiple-choice questions in the test (n), with only one correct answer	**Probability of scoring 50% or higher by blind guessing (P)**
5	0.1035
10	0.0781
20	0.0139

$$P = \begin{cases} \sum_{k=\frac{n}{2}}^{n} \binom{n}{k} p^k (1-p)^{n-k}, & \text{if n is even} \\ \sum_{k=\lfloor \frac{n}{2} \rfloor + 1}^{n} \binom{n}{k} p^k (1-p)^{n-k}, & \text{if n is odd} \end{cases}$$

where $\binom{n}{k} p^k (1-p)^{n-k}$ is the probability to get questions right, and $p = \frac{1}{4}$ is the probability to get an individual question right. As can be noted from table 2, the probability P equals to 0,0139, when the test consists of 20 MCQs. If the total number of students is equal to N, it can be expected that $p \cdot N$ of them will score 50% or more by blind guessing. This means that, for example, if 100 students are tested with 20 MCQs (each containing 4 alternatives, and only one of them presents the correct answer), then the expected number of students to pass the test by blind guessing is one.

Difficulty of construction. The key to take advantage of MCQs' strengths (presented above), is to construct good multiple-choice items. However, good MCQs are generally more difficult and time-consuming to write than other types of questions. This is particularly evident for the process of determining plausible distractors, which requires a certain amount of skills. These capabilities, however, may be increased through study, practice and experience.

2 Increasing MCQ validity by implementing standard protocols

Versatility and reliability are inherent qualities of MCQs, but their validity cannot be assumed due to the possibility of the student to guess the correct answer even when he or she lacks the expected knowledge. Therefore,

standard **prevalidation** and **postvalidation** protocols are recommended to increase the validity of MCQs (Ramakrishnan et al., 2017).

2.1 Standard MCQs prevalidation protocols

Prevalidation is a process that prevents error occurrence in the construction of MCQs by using guidelines and checklists. The most important observation is to construct MCQ test with independent items. This will disable students to use information from one item in order to answer another one, thus reducing the test validity. The most notable guidelines which should be observed when developing effective multiple-choice items are presented below. For better illustration, the guidelines are augmented with multiple-choice question examples from the test-collection described in (Jovanovska, 2018). All of the questions are taken from the Macedonian State Matura, and are publicly available on the State Examination Center's web site.[6] The correct answers are highlighted.[7]

Constructing an effective stem. The following requirements are crucial in the process of constructing an effective stem (Frey et al., 2003):

1. The stem should be meaningful by itself and should present a definite problem. Such a stem guarantees that the item is focusing on assessing learning outcomes (table 3).
2. The stem should not contain irrelevant information, which can reduce the reliability and the validity of the test results (table 4) (Haladyna and Downing, 1989).
3. The stem should be expressed with a negation only when a significant learning outcome requires it. Studies confirm that students have difficulty in understanding items with negative phrasing (Michael Rodriguez, 1997). If there is a necessity of a negative expression for assessing specific expertise (such as in medicine), then the negation must be emphasized with italics or capitalization (table 5).
4. The stem should be a question or a partial sentence (Statman, 1988). A question stem is preferable because it allows the student to focus on answering the question, rather than memorizing the partial sentence and subsequently completing it with each of the alternatives. Studies

[6] State Exam Center, Bureau for Development of Education (on-line)

[7] In on-line edition they are indicated in red, in printed edition they are given in small caps.

have already confirmed that the cognitive load increases if the stem is constructed with initial or internal blank, Table 6) (Brame, 2013).

Constructing effective alternatives. The process of creating item's alternatives should fulfill the following recommendations (Frey et al., 2003):

Table 3. An example of an item with a meaningless stem and its improved formulation

Item with a meaningless stem	**Item's improved formulation**
Which of the following thesis is true:	Phenomenology is the science of?
A. phenomenology is the science of the circumstances in nature	A. the circumstances in nature
B. phenomenology is the science of the beauty things	B. the beauty things
C. PHENOMENOLOGY IS THE SCIENCE OF THE PHENOMENA IN CONSCIOUSNESS	C. THE PHENOMENA IN CONSCIOUSNESS
D. phenomenology is the science of the goodwill principles	D. the goodwill principles

Table 4. An example of a stem with irrelevant information and its improved formulation

A stem with irrelevant information	**Stem's improved formulation**
When Aristotle was 18 years old, he came to Athens to study at Plato's Academy. As a best student, associate and lecture, afterward, he became a tutor of:	Whose teacher was Aristotle?
A. ALEXANDER THE GREAT	A. ALEXANDER THE GREAT
B. Philip of Macedon	B. Philip of Macedon
C. Socrates	C. Socrates
D. Al-Farabi	D. Al-Farabi

Table 5. An example of a stem with a negation and its improved formulation

A stem with a negation	**Improved formulation**
Which of the following does NOT belong to the five pillars of Islam?	The five pillars of Islam, as Muslim's religious and ethical obligations, include all of the following except:
A. Shahada	A. Shahada
B. Salat	B. Salat
C. Zakat	C. Zakat
D. Halakhah	D. Halakhah

Table 6. An example of a stem with an internal blank and its improved formulation

A stem with internal blank	**Improved formulation**
Together with Holbach, _____ is one of the most important French encyclopedists.	Who is one of the most important French encyclopedists along with Holbach?
A. Helvetius	A. Helvetius
B. Camus	B. Camus
C. Proudhon	C. Proudhon
D. Derrida	C. Derrida

Table 7. An example of implausible alternatives for a given stem

Implausible alternatives (B and D)
Which philosopher is the most influential representative of the modern intuitionism?
A. Henri Bergson
B. Albert Einstein
C. Norbert Wiener
D. John Kennedy

1. All alternatives should be plausible. Implausible alternatives don't present functional distractors and should not be used. The common students' mistakes provide the best source of distractors (table 7).

2. The alternatives should be stated clearly and concisely and should be mutually exclusive. Students consider that the items containing alternatives with an overlapping content can undermine the confidence of the evaluation.

3. The alternatives should not provide clues which rule them out. Otherwise, the sophisticated students can reveal the correct answer easily. Therefore, it is important that the alternatives are similar in length, use the same expression style and have a grammar consistent with the stem.

4. The alternatives "all of the above" and "none of the above" should be avoided when designing multiple-choice items. If the alternative "all of the above" is used as a correct answer, the student who can identify more than one alternative as correct, can select the correct answer even though he or she is not sure about the other alternatives. The same argument is true when the alternative "none of the above" is used as a correct answer. In both cases, it is possible to apply partial knowledge to correctly answer the item.

5. The general assumption in the process of designing multiple-choice questions is that the order of the alternatives is completely irrelevant, until answers are randomly assigned to positions or equally distributed among them (Attali and Bar-Hillel, 2003). In that sense, Hohensinn and Baghaei (2017) examined if the item difficulty depends only on the item stem, or it is influenced by the position of the correct answer. The analysis confirmed that the position of the correct answer has a very small effect on the multiple choice item difficulty and the common practice of distributing correct options randomly is valid. Haladyna et al. (2002) presented a taxonomy of guidelines for creating MCQs. The part referring to the positions of the alternatives emphasizes that the alternatives should be given in a logical order (such as alphabetical or numerical) to avoid biases towards certain positions.

6. The number of alternatives can vary among multiple choice questions, as long as all the alternatives are plausible. There is no strong evidence that confirms significant differences in the item difficulty and the reliability of the test results between the questions that contain two, three or four distractors (Haladyna, 2004).

2.2 Standard MCQs postvalidation protocols

Postvalidation helps to identify MCQs with questionable validity so that they can be appropriately modified before reusing or discarded. Item analysis is a postvalidation procedure which characterizes every MCQ by assigning numerical values, such as: difficulty index, discrimination index and distractor analysis. Based on the standard acceptable limits for these numerical values, MCQs can be accepted, modified and revalidated, or discarded.

Difficulty index. The difficulty index is one of the most commonly used statistics for item analysis. It is a measure of the proportion of those students who answered the item correctly, and therefore it is frequently called the p – value. A higher p – value indicates that a greater proportion of the students answered the item correctly, and thus the item is considered as an easier one. The difficulty index is obtained by dividing the number of students who answered the item correctly by the total number of students who answered that item, thus ranging between 0.0 and 1.0 (Crocker and Algina, 1986). Table 8 presents three different item categories depending on the range to which the difficulty index value belongs.

Table 8. Item categories depending on the difficulty index value (Wiersma and Jurs, 1990)

Range of the difficulty index	Item category
$p \leq 0.30$	difficult
$0.30 < p \leq 0.70$	acceptable
$p \geq 0.70$	easy

Discrimination index. The item discrimination index demonstrates how well the item is able to distinguish between students who achieved the learning outcomes and those who did not. In computing this measure, a group of the best performing students is analyzed (the upper group), along with a group of students who did poorly on the overall test (the lower group). To ensure stability, it is preferable that the groups include larger number of students. It is also desirable for these groups to be more diverse, in order to make the discriminations clearer. According to Wiersma and Jurs (1990), the use of 27% of the total number of students in each group, maximizes these two features.

The discrimination index is defined by:

$$D = \frac{U - L}{N}$$

where U and L are the number of students in the upper and lower group, respectively, who answered the item correctly, and N is the number of students in the largest of the two groups. Wood (1960) stated that when more students in the lower group than in the upper group selected the correct answer of the item, then the item has a negative validity. A negative value indicates that the item is not only useless, but also decreases the test validity. Table 9 presents four different item categories depending on the range to which the discrimination index value belongs.

Table 9. Item categories depending on the discrimination index value (Ebel and Frisbie, 1986)

Range of the difficulty index	Item category
$D \geq 0.40$	very good
$0.30 \leq 30 \leq 0.39$	reasonably good (possibly subjected to improvement)
$0.20 \leq D \leq 0.29$	marginal (necessity of revision)
$D \leq 0.19$	poor (necessity of major revision or elimination)

Distractor analysis. The two measures defined above do not consider the characteristics of the item distractors and the way they influence the student's decision to select one of the alternatives. Distractor analysis addresses these issues by examining the quality of the distractors as one important element of the item quality. Each distractor must be plausible and clearly incorrect.

One simple approach of distractor analysis includes determination of the proportion of students who selected each of the alternatives. These proportions can be particularly informative. For example, when the proportion of students who selected a given distractor is greater than the proportion of students who selected the correct answer, then the item should be examined to determine if the correct answer is mistaken. The distractor analysis can also reveal implausible distractors. For example, if the students consistently

fail to select a given distractor, this may be an evidence for its implausibility. Distractors not selected by 5% or more of the students are considered ineffective and should be revised or eliminated (Linn and Gronlund, 2000).

2.3 Relevant research

The conducted research confirmed that there is a significant space for improving the quality of many tests based on multiple choice questions. Analyzing a sample of 60 multiple-choice questions from medical field, Ramakrishnan et al. (2017) concluded that more than one third of all the distractors were not functional, i.e. they were not acceptable. Those distractors should be modified or replaced and tested again, until meeting the defined criteria (achieving distractor effectiveness equal or higher than 5%). Halikar et al. (2016) analyzed 20 multiple-choice questions from the same field (medicine) in detail and noticed that all questions had at least one nonfunctional distractor, while the total number of nonfunctional distractors was 23% from the set of distractors. The results also revealed that the percentage of the acceptable questions, based on the difficulty index and discrimination index, was 35% and 50%, respectively. Therefore, the authors recommended a generation of a pool with valid MCQs, where each question is associated with its index values. Thus, the examiners can choose proper MCQs from that pool for certain testing. In their research, Battista and Kurzawa (2011) highlighted that the examiners need training courses and support, in order to be sure that their MCQ tests are well-designed and have acceptable discriminatory power. The process of creating high-quality MCQ tests is a skill that can be learned (Jozefowicz et al., 2002).

3 Considerations for writing MCQs that test higher-order thinking

Despite the fact that the initial designing of the cognitive skills' taxonomies was accomplished to overcome the distinctiveness of the different domains, the experts agree that the higher-order cognitive processes are inherently domain-specific. Anderson et al. (2001) acknowledged that each major field should have its own taxonomy. The experts are faced with the challenge of operationalizing general taxonomy levels for their specific area (Morrison and Free, 2001). Therefore, the number of MCQ basic construction rules for different cognitive levels is partly restricted. Nevertheless,

some strategies are identified that may help when designing MCQs, which reach beyond mere recall. Following paragraphs offer some recommendations that might facilitate this process.

Application of specific verbs. Morrison and Free (2001) associate certain verbs to the various cognitive processes (table 10). When a particular verb is placed in an item, it may serve as an indicator that the corresponding cognitive level is assessed. However, this strategy should be used carefully because some verbs could be placed in multiple levels, and much depends on the context of the item in which the verb is placed. Nevertheless, this mapping gives an objective and transparent basis for the item developers.

Table 10. Examples of verbs associated with various categories of Bloom's Taxonomy (Morrison and Free, 2001)

Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
Identify	Describe	Apply	Analyze	Compose	Appraise
Define	Differentiate	Calculate	Categorize	Construct	Assess
Know	Discuss	Classify	Compare	Create	Evaluate
List	Explain	Develop	Contrast	Design	Judge
Name	Rephrase	Examine	Distinguish	Formulate	
Recognize	Restate	Solve	Determine	Modify	
State	Reword	Use	Investigate	Plan	

Using realistic scenarios. One of the best ways to promote and evaluate higher-order thinking is to use questions based on realistic situations, especially those that simulate real work experiences (Scully, 2017).

Analysis of visuals. Critical thinking skills can be assessed by asking the students to analyze or interpret information from visuals, which are provided as an integral part of the question stem or the alternatives. In many cases, these visuals, such as diagrams and graphs, simulate the real tasks from different workplaces (Burton et al., 1991).

Request to elaborate the answer. Higher-order thinking can also be evaluated if the students are asked to synthesize what they have learned.

This means that the answers should include explanations that support them (Lord et al., 2009).

4 Concluding comments

The assessment of learning outcomes is crucial for educational improvements. The process of student assessment should align with curricular goals and educational objectives. Identifying the proper assessment strategies for students' progress evaluation within individual programs is as important as establishing curricular content and delivery methods. MCQs, as one of the frequently used assessment strategies, have certain strengths and weaknesses. They are efficient, flexible, objective, easy to implement and grade and can be used to assess large content of the curriculum. However, the development of a good MCQ test for evaluating students' achivements is a very challenging goal. Even when examiners follow a series of guidelines for constructing fair and systematic tests, different factors may influence student's perception of the test items. In order to increase the MCQs' quality, it is vital to analyze items' difficulty and discrimination indices, which might help the test developers to make the test assessment more meaningful.

It is often argued that the multiple-choice items are suitable for assessing only the lower-order thinking skills. Nevertheless, a more accurate assertion may be that the multiple-choice items measuring complex cognitive processes are simply rarely constructed. Adhering to certain strategies, it is possible to construct multiple-choice items that measure processes such as knowledge application, analysis and evaluation.

References

Anderson, Lorin W., David R. Krathwohl and Benjamin Samuel Bloom. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.

Attali, Yiagal and Maya Bar-Hillel. "Guess where: The position of correct answers in multiple-choice test items as a psychometric variable". *Journal of Educational Measurement* Vol. 40, no. 2 (2003): 109–128.

Bacon, Don R. "Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context". *Journal of Marketing Education* Vol. 25, no. 1 (2003).

Battista, David Di and Laura Kurzawa. "Examination of the quality of multiple-choice items on classroom tests". *The Canadian Journal for the Scholarship of Teaching and Learning* Vol. 2, no. 2 (2011). http://ir.lib.uwo.ca/cjsotl_rcacea/vol2/iss2/4

Bloom, Benjamin S. *Taxonomy of Educational Objectives*. New York: David McKay Company Inc., 1977.

Brame, Cynthia J. *Writing good multiple choice test questions*. Vanderbilt University Center for Teaching, 2013. https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

Brown, Sally and Peter Knight. *Assessing Learners in Higher Education*. London : Routledge Falmer, 1998.

Burton, Steven J., Richard R. Sudweeks, Paul F. Merrill and Bud Wood. *How to prepare better multiple-choice test items: Guidelines for University Faculty*. Brigham Young University Testing Services and The Department of Instructional Science, 1991.

C., Chan. *Assessment: Multiple Choice Questions. Assessment Resources*. HKU, University of Hong Kong, 2009. http://ar.cetl.hku.hk/am_mcq.htm

Crocker, Linda and James Algina. *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston, 1986.

Dikli, Semire. "Assessment at a distance: Traditional vs. alternative assessments". *The Turkish Online Journal of Educational Technology* Vol. 2 (2003).

Ebel, Robert L. and David A. Frisbie, *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

Frey, Bruce B., Stephanie Petersen, Lisa M. Edwards, Jennifer Teramoto Pedrotti and Vicki Peyton. "Toward a consensus list of item-writing rules", 2003.

Haladyna, Thomas M. *Developing and validating multiple-choice test items (3rd ed.)*. Lawrence Erlbaum Associates, 2004.

Haladyna, Thomas M. and Steven M. Downing. "Validity of a taxonomy of multiple-choice item-writing rules". *Applied Measurement in Education* Vol. 2, no. 1 (1989): 51–78.

Haladyna, Thomas M., Steven M. Downing and Michael C. Rodriguez. "A review of multiple-choice item-writing guidelines for classroom assessment". *Applied Measurement in Education* Vol. 15,no. 3 (2002): 309–334.

Halikar, Swapnagandha S., Veerendra Godbole and Saurabh Chaudhari. "Item Analysis to Assess Quality of MCQs". *Medical Science* Vol. 6, no. 3 (2016): 123–125.

Hohensinn, Christine and Purya Baghaei. "Does the position of response options in multiple-choice tests matter?". *Psicologica* Vol. 38 (2017): 93–109.

Jovanovska, Jasmina. "Multiple-choice question answering system for Macedonian and English test-collections". PhD. thesis, Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, Republic of Macedonia, 2018.

Jozefowicz, Ralph F., Bruce M. Koeppen, Susan Case, Robert Galbraith, David Swanson et al. "The quality of in-house medical examinations", *Academic Medicine* Vol. 77, no. 2 (2002): 156–161.

Linn, Robert R. and Norman E. Gronlund. *Measurement and assessment in teaching (8th ed.)*. Prentice Hall, Upper Saddle River, NJ, 2000.

Lord, Thomas R., Donald P. French and Linda W. Crow. *Guide to Assessment*. National Science Teachers Association, 2009. http://static.nsta.org/files/PB231Xweb.pdf

Michael Rodriguez, K. "The art and science of item-writing: A meta-analysis of multiple-choice item format effects", 1997.

Morrison, Suzan and Kathleen Walsh Free. "Writing multiple-choice items that promote and measure critical thinking". *Journal of Nursing Education* Vol. 40, no. 1 (2001): 17–24.

Palmer, Edward J. and Peter G. Devitt. "Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions?" *BMC Medical Education* Vol. 7 (2007): 49–55.

Ramakrishnan, Minu, Aditya B. Sathe and Vinayak A. "Item analysis: A tool to increase MCQ validity". *Indian Journal of Basic and Applied Medical Research* Vol. 6, no. 3 (2017): 67–71. http://ijbamr.com/pdf/June201767-71.pdf.pdf

Scully, Darina. "Constructing multiple-choice items to measure higher-order thinking". *Practical Assessment, Research and Evaluation* Vol. 22, no. 4 (2017). http://pareonline.net/getvn.asp?v=22&n=4

Stassen, Marta L., Kathryn Doherty and Mya Poe. *Course-based review and assessment: Methods for understanding student learning*. Office of Academic Planning & Assessment, University of Massachusetts Amherst, 2001.

Statman, Stella. "Ask a clear question and get a clear answer: An enquiry into the question/answer and the sentence completion formats of multiple-choice items". *System* Vol. 16, no. 3 (1988): 367–376.

Towns, Marcy H. "Guide To Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments". *Journal of Chemical Education* Vol. 91 (2014): 1426–1431.

Walsh, Catherin and Lisa Seldomridge. "Critical thinking: Back to square two". *Journal of Nursing Education* Vol. 45, no. 6 (2006): 212–219.

Wiersma, William and Stephen G. Jurs. *Educational measurement and testing (2nd ed.)*. Boston, MA: Allyn and Bacon, 1990.

Wood, Dorothy Adkins. *Test construction: Development and interpretation of achievement tests*. Columbus, OH: Charles E. Merrill Books, Inc, 1960.