American International Journal of Humanities and Social Science, Vol. 5 No 3; August 2019
ISSN 2415-1254 (Online), ISSN 2415-1505 (Print)
Published by Center for Global Research Development

Writing Multiple Choice Items that are Reliable and Valid

Dr. Patricia E. Allanson
Liberty University
United States of America

Dr. Charles E. Notar (Emeritus)
Jacksonville State University
United States of America

Abstract

*This article is written to help educators construct high-quality, reliable, and valid **multiple-choice** test items to evaluate students' ability to demonstrate proficiency in the learning objectives of the content domain under study. Using **item**-writing guidelines based upon the research literature allows an educator to create assessments that provide greater ability to discriminate between high- and low-achieving students. This article provides the most science based advanced rules for writing multiple choice test items.*

Introduction

A major role in the classroom teachers' practice is to assess student learning. Gareis and Grant (2015) define this practice as "the ability to create and use valid and reliable assessments as a classroom teacher to facilitate and communicate student learning" (p.178). An ideal assessment would measure how well students know, comprehend, apply, analyze and evaluate the material they are currently studying (Lumley & Craven, 2004; Crossley, Humphris & Jolly, 2002). Cohen and Swerdlik (1999, p. 215 as cited in Vacc, Loesch, & Lubik, 2001) states "The creation of a good test is not a matter of chance it is the product of the thoughtful and sound application of established principles of test construction." Although there are many different types of testing items available for use in assessing learning, Multiple-choice questions (MCQs) are universal and the most commonly used (Moreno, Martínez, & Muñiz, 2006; Xu, Kauer, & Tupy, 2016). Multiple-choice questions are used in high stakes testing to assess year end learning, as a quick method to assess daily learning, and at educational institutions with larger than ever class sizes. However, Pettijohn and Sacco, (2007) see this practice as a conflict with the belief that **multiple-choice test items are** inferior to essay and other open-ended assessments that measure and promote learning. Many educators are reluctant to use MCQs because they hold the belief that MCQs are best used for measuring lower-level objectives or the simple recall of facts, however, well-constructed MCQs have the applicability to measure higher-level thinking such as comprehension, analysis and cognitive reasoning (Torres et al., 2011).

Multiple Choice Questions

Torres, et al. (2011) defines MCQs as one where "students are asked to select one alternative from a given set of alternatives in response to a question stem" (p.1) consisting of three parts: a stem, a correct answer, and distractors. Multiple-choice questions are generally used to measure learning achievements as an objective test item compared to essay test items that are considered subjective in nature. Research has shown that both test item formats, objective and subjective, are valuable in measuring student learning, however, it is important to note situations where one is more suitable than the other. Both can be used to measure educational achievement and the ability to apply principles, to think critically and solve problems.

The use of subjective test items is appropriate when the test group is small and items will not be reused, to encourage writing development, and assessing student attitudes rather than measuring achievement. Objective test items are best used with large groups with the potential of reusing the specific test items, and when it is necessary to obtain reliable scores that provide impartiality from outside influences (Clay, 2001). Regardless, the test item format must align with learning objectives to ensure high test validity – testing what is supposed to be tested. Brookhart (2015) states that "The effectiveness of multiple-choice questions—or any kind of questions—depends on their relationship to what students are trying to learn" (p. 36) and what the instructor is trying to assess.

Advantages and Disadvantages of Multiple-Choice Questions

There are many advantages and disadvantages of using MCQs in classroom assessments. Parmenter (2009) discovered that students prefer MCQs over essay type items as MCQs provide opportunities for higher scores and prove to be a more valid type of assessment rather than its subjective counterpart. Smith and Karpicke (2014) state there is little or no advantages of answering essay type questions over MCQs. Multiple-choice questions are widely used for end of course summatives and standardized tests. They can also be used to diagnose learning gaps through formative assessments, and as an analysis to identify student thinking - why certain answers were selected (Brookhart, 2015). Other advantages include: measuring a greater variety of educational objectives and subject matter at different cognitive levels; ease of administration with large groups and consistency of scoring; reliable in assessing student progression; reducing grading bias and test taking anxiety; and easily implemented by computer which allows for immediate results or feedback (Begum, 2012; Boland, Lester, & Williams, 2010; Brookhart, 2015; Burton et al., 1991; Morrison & Free, 2001; Zimmaro, 2004; Zeidner, 1987 as cited in Xu, Kauer, & Tupy, 2016).

Although MCQs are versatile and easy to score, they are difficult to construct, especially for higher level structured questions, and prove to be a time-consuming endeavor, (Tulving, 1967; as cited in Roediger & Marsh, 2005). Funk and Dickson (2011) found that MCQs can overestimate or provide inaccurate information as to learning outcomes. As with any type of test item, MCQs have limitations such as the inability to measure student creativity, articulation of ides and explanations, or justifying answer responses which is best left to be assessed by subjective test items (Burton et al., 1991), and allows for guessing or ruling out answers regardless of actually knowing the content (Funk & Dickson, 2011). Multiple-choice questions also eliminate the ability to receive partial credit for knowing some of the content being assessed.

A major disadvantage with MCQs is that very few educators are trained or possess skills in writing quality MCQs despite guidelines in numerous publications. Technology that produces automatic question creations and textbooks exams are also guilty of poor construction traits (Gierl et al., 2016; Gutl et al., 2011; Tarrant & Ware, 2008). According to Ory (n.d), "MC items need writing ability of the teachers and reading ability of the students" (p. 7). Unfortunately, these poorly constructed MCQs tend to show up in high-stakes summative assessments and can be disadvantageous for high achieving students compared to borderline students (Khan et al, 2013; Tarrant & Ware, 2008).

Parts of a Multiple-Choice Question

As noted, there are many concerns with using MCQs as an assessment practice, however, there are practical ways to overcome these negative apprehensions. Learning how to correctly format and create quality MCQs, that have the ability to assess deeper-level thinking, can lead to enhanced student learning and performance, an increase in instructor's efficiency, and a better understanding of assessment outcomes.(Simkin & Kuechler, 2005; Tractenberg, Gushta, Mulroney, & Weissinger, 2013 as cited in Xu, Kauer, & Tupy, 2016). The decision to use MC tests or include MC items in a test should be based on what the purpose of the test is and the use that will be made of its results. As with any test item construct, assessment must be based on performance objectives with action verbs (i.e. Bloom's Taxonomy, 1956; Anderson, et al., 1961 revised Bloom's Taxonomy) that are observable and measurable.

Typically, MCQs are created with two parts: A stem, and answer choice options. The stem is comprised of a question, problem, or an incomplete statement and should be meaningful, definitive, focused on a learning outcome, and not contain irrelevant information. The stem is followed by answer choices with one correct option and several other options. These other options are called distractors and serve the purpose of appearing "as tempting solutions to the problem, plausible competitors of the answer for the students that do not achieved the objective measured by the test item" (Torres et al., 2011, p. 5).

Well written distractors will characteristically be selected by lower achieving students and ignored by those of higher academic abilities. Like all types of assessment items, well- written MCQs follow a set of guidelines using clear and simple language asking a direct question with plausible answers. Brookhart (2015) suggests that these plausible choices "should reflect common errors in student thinking so that even wrong answers give students and teachers information about what students know and can do" (p. 36).

Domyancich (2014) describes four types of MCQs:

- "Algorithmic - These items require the application of a memorized routine.
- Lower-order cognitive skills (LOCS) - These items test recall or application of knowledge in a familiar situation.
- Conceptual - These items require the demonstration of the basic understanding of scientific themes.
- Higher-order cognitive skills (HOCS) - These items require the ability to link concepts in an unfamiliar situation" (1347-1348).

There are several different varieties of MCQs (Burton et al., 1991). The authors will focus on the more commonly used varieties as listed below with a brief description of each:

- **Single Correct Answer**: Only one alternative option is correct where the remaining options are incorrect or distractors. According to Vacc, Loesch, and Lubik, 2001, this is the most commonly used variety of MCQs.
- **Best Answer**: Similar to single correct answer format with the exception that one answer is clearly more correct than the other options, and students are directed to identify or select the "best answer".
- **Negative**: The opposite of single correct answer variety where one option is incorrect and the remaining options correct.
- **Combined Response**: One or more alternative options are correct and identified by selecting sets of letters or numbers (i.e.: 1, 2, and 3; a and b only).

Multiple-Choice Questions that Promote Higher Order Thinking

In addition to using varieties of MCQs, it is important to address different levels of cognitive processing according to Bloom's Taxonomy (1956) and Anderson et al.'s (2001) revision of Bloom's Taxonomy. Knowledge items are concerned with remembering and recalling material (e.g., *Which of the following mountains is the tallest in the world?*); Comprehension items measure the extent to which students are able to describe and explain the material (e.g. *Which of the following describes how equations differ from expressions?*); Application items measure whether students can apply material in a novel context (e.g. *Using the repair flowchart shown here, what should you check if the processor stops working?*); Analysis items use learned material breaking them in to parts and then recomposing to determine how the parts relate to one another (e.g. *Given that the student solves the problem in X, Y, Z manner, which of the following learning disability is most likely?*); Evaluation items use knowledge to make judgement based on set criteria and standard (e.g. evaluating the validity of an argument, or the most appropriate policy) (Scully, 2017).

Haladyna, Downing, and Rodriguez (2002) note that in order to avoid testing for simple recall (measuring beyond the knowledge level), test items must contain novelty or familiarity of language used during the instructional process. Morrison and Free (2001) suggest utilizing questions that require respondents to answer in a manner that maintains a higher level of discrimination such as answering what is the best, most important, first, highest priority, etc and focus on application of concepts. Another strategy to promote critical thinking uses real-life scenarios or problem-based questions (Iwaoka et al. 2010; Ling-Na et al. 2014), or requires respondents to interpret data, images, or diagrams (Azer, 2003; Kong et al., 2009) aligned with course's objectives (Bassett, 2016). Question that promote higher level thinking may also ask respondents to identify an illustrated theory or predict an outcome based on a hypothesis (Davis, 1993). Psychologist Steven Pinker's mind games (2014) provides a prime example of MCQs that promote critical thinking.

Best Practices for Creating Multiple-Choice Questions

Many proponents agree that writing the stem portion of a MCQs is easier to construct than the options (distractors). Begum (2012) states that the stem must directly relate to the learning objectives assessed and at the same level of learning. Each question in the assessment should progress in an easy to difficult order .

Test items must be grammatically consistent, focused and give adequate information in order to be answered correctly without having to read the alternatives/options. (Begum, 2012; DeChamplain, 2010; Suski & Banta, 2009). Previous research suggest the the stem should be written in question format either as an incomplete statement or direct question, and stated in a positive manner (Sireci, Wiley & Keller, 1998; Haladyna & Downing, 1993; Kehoe, 1995). The stem must include relevant information void of excessive verbiage, and of an appropriate difficulty level. Suski and Banta (2009) suggest writing stems that ""Remove all the barriers that will keep a knowledgeable student from answering the item correctly...[and] Remove all clues that will help a less-than-knowledgeable student answer the item correctly" (p. 170). Researchers also suggest restricting the use of negative verbiage and absolute terms (i.e. always, never, all, none, not, seldom, rarely, occasionally, sometimes, few, many, or except) in the stem or avoid using them altogether. If the use of negatives in the stem is unavoidable, it is best to underline, boldface or capitalize the negative word (Begum, 2012; McCoubrie & McKnight, 2008; Torres et al., 2011). Irrelevant clues (may, could, can) and imprecise terms (seldom, rarely, occasionally, sometimes, few, many) should be avoided as well as they lead to correct options (Begum, 2012; Torres et al., 2011).

Once the stem is created, the difficult task of creating good answer choices follows. Literature also refers to these answer choices as alternatives, distractors, elections, choices, and options, however, the authors will use distractors when referring to such. Multiple-choice item distractors consist of three or four distractors (incorrect choices) and the correct response. The primary function of distractors is to "distract" students who may be uncertain of the answer. Good multiple-choice items depend on effective distractors and best assessed in terms of the quality of such (Begum, 2012; Hansen & Dexter, 1997). Distractors must include one choice that is clearly the only correct response with the remaining alternatives appearing as plausible to the student (Miller & Erickson, 1985). If correctly formatted, a good distractor will entice a low achiever to select it, but be ignored by high achievers (Haladyna, 1999 as cited in Torres et al., 2011).

The distractors should be sufficient in number to reduces the chance of guessing which is a main issue in MCQ construction. Typically, 4-5 distractors are recommended, and most commonly used in standardized testing, however several studies have suggested that three distractors is optimal (Baghaei & Amrahi 2011; Rodriguez, 2005). Research also noted that three distractors offer additional benefits – faster time to create items, administer the assessment, and instructors' ability to cover more content on the assessment. Students were also able to answer questions fives times faster than those completing assessments with four to five distractors (Schneid, Armour, Park, Yudkowsky, & Bordage, 2014 as cited in Xu, Kauer, & Tupy, 2016).

In constructing the answer choices, items should be plausible but clearly incorrect, parallel in form, and follow the normal rules of grammar and punctuation. For example, stems in question form should have alternatives that begin with capital letters (Hansen & Dexter, 1997; McCoubrie & McKnight, 2008). Distractors that are grammatically incorrect may lead students to reject distractors as the stem is stated. Words from the stem should not be reused in the options to avoid cueing in the test-wise student nor should the central idea be present (Campbell, 2011; Frary, 1995; Haladyna, Downing, & Rodriguez, 2002; Pachai, DiBattista, & Joseph, 2015). Kehoe (1995) recommends including as much information in the stem, but providing as little as possible in the distractors. he distractors should also be arranged in logical order and independent of one another (Vacc, Loesch, & Lubik, 2001; Haladyna, Downing, & Rodriguez, 2002). McCoubrie and McKnight (2008) suggest avoiding the use of frequency and absolute terms such as rarely, usually, often, commonly, always, never, and all. Frary (1995) also recommends not using negative distractors following a negative stem. Several researchers suggest that the use of distractors that utilize student misconceptions increases the difficulty of the question and adds overall value of results for the instructor. (Begum, 2012; Haladyna, Downing, & Rodriguez, 2002; Timmermann and Kautz, 2015; Vacc, Loesch, & Lubik, 2001). Most researchers strongly advise against using "all of the above" or "none of the above" (AOTA or NOTA), or using it as a distractor sparingly (Begum, 2012; Frary, 1995; Haladyna, Downing, & Rodriguez, 2002; Hansen & Dexter, 1997; Sireci, Wiley, & Keller, 1998). Using NOTA is often used in cases where designers are unable to create another plausible distractor. This type of distractors may be useful in a students ability to detect incorrect options or imply that the options vary in degree of correctness (Hansen & Dexter, 1997; Kehoe, 1995). Some defensible cases where NOTA distractors are acceptable would be in answers that require computation. Several authors also warn about convergence of distractors and logical clues especially when using unbalanced multi-part answers (Campbell, 2011; McCoubrie & McKnight, 2008).

After the distractors are written, test developers will want to pay attention to the location and order of the answers arranging them systematically (alphabetically, chronologically, and numerically) (Begum, 2012; Haladyna, Downing, & Rodriguez, 2002; Harasym, Leong, Violato, Brant, & Lorscheider, 1998 as cited in DiBattista, Sinnige-Egger, & Fortuna, 2014; Kehoe, 1995; Miller & Erickson, 1985). Randomizing the order of questions and using multiple versions will helps to prevent cheating. This is especially an easy task with online test banks. The order of question can also be set up based on difficulty level, which according to Weinstein and Roediger (2010, 2012) does not lead to significant differences on performance, but increases students' perception and optimism of their own performance. Tellinghuisen and Sulikowski's (2008) research confirmed that performance with MCQs questions strongly depends on the placement of answer alternatives. Balch (1989 as cited in Pettijohn & Sacco, 2007) did find that sequential ordering had an impact on student performance levels as students were able to recall context as it was learned thus enhancing recall. At this point in the writing process, it is time to ask if the question passes the "cover test" which refers to a student being able to answer the question correctly even when the answer and distractors are covered up (Campbell, 2011). Campbell also warns that it is "too late to teach" and not to attempt to 'teach' within a question.

Multiple choice questions may be simpler to answer than the constructing process especially when dealing with levels of higher order thinking. Barlow (2014) summarizes the best practices to follow when constructing MCQs as noted in current research (see also Haladyna, Downing & Rodriquez, 2002 for an empirically validated set of guidelines to apply).

Summary of Best Practices for MCQ Writing

Author(s)	Area of Best Practice (Commonalities are in **Bold**)		
	General	Stems	Responses (Distractors)
Suskie (2009)	• Be concise • Define all terms • Avoid unnecessarily complex vocabulary • Avoid "interlocking" items	• Ask a complete question. • Avoid "which of the following" items. • Avoid common knowledge questions • Avoid negative stems • Avoid grammatical clues to the correct answer	• Not all questions need the same number of options.* • Order responses logically • Use vertical responses rather than horizontal • Make all options similar length. • Avoid "None of the above" and "All of the above" • The best distractors identify where students' thinking went wrong, and should be intrinsically possible or true statements
Brunnquell et al. (2011)		• Avoid negative stems • Avoid unfocused or vague stems • Avoid verbal associations between stem and answer	• Avoid "cues" such as "always," "never," "usually," etc. • Avoid "None of the above" and "All of the above" • Make all options similar length.
Case & Swanson (1998)	• Items should focus on important concepts only • Avoid trick questions • Assess application of knowledge rather than recall of facts • Avoid clues for testwise students	• Be clear and concise • Avoid "which of the following" or "Each of the following…except" items. • Avoid "hinging" (i.e. interlocking) items	• Distractors should be homogeneous. • Avoid options with two parts • Order responses logically • Make all options similar length • Distractors should be intrinsically possible or true statements

Table 1: Summary of Best Practices for MCQ Writing (Barlow, 2014, p. 60)

Countermeasures for Cheating

While most college students agree that cheating is unethical, a substantial proportion of them will (or will in the future) cheat in college. Nath and Lovaglia (2009) report admissions of cheating range from 13 to 95 percent in higher education. With students becoming more tech-savvy and online course enrollment at its highest, academic integrity is under fire. Exam cheating is thought to occur more often in online courses compare to traditional face-to-face courses, and students are becoming more proficient at cheating through texting and social media use (Harmon, Lambrinos, & Buffolino, 2010).

There are several strategies to use to combat cheating in both traditional face-to face and online course environments. A common way to discourage cheating is to make more than one version of the test, however Vander Schee (2013) warns that in doing so it may unintentionally compromise student performance on sequencing of items (easy to high level of item difficulty). Another method is to use a seating chart and to number tests thus reducing false accusations if two incorrect answers are obtained and students were not sitting next to each other. For online courses, Harmon, Lambrinos, and Buffolino (2010) suggest instructors modifying their exams by utilizing shuffling tactics and proctoring some of the exams. Item critiquing and asking student to provide rationals or explanations for their answer selections is another way to combat cheating and also provides opportunities for students to practice metacognitive skills (Roediger & Marsh, 2005; Brookhart, 2015). Xu, Kauer, and Tupy's, (2016) review of empirical literature, provides a "teacher-ready" set of guidelines (see Table 2) to optimize the use of multiple-choice question items that focuses on assessment, feedback and efficiency. This includes techniques to combat cheating in both the face-to-face and online environments.

Optimizing Multiple-Choice Question Testing

Domain	Optimization techniques
Assessment quality	Utilize questions designed for higher-order cognitive assessment.
	Discourage students from guessing and/or utilize methods of scoring that penalize guessing.
	Improve quality of exams by conducting item analyses.
	Utilize collaborative testing when appropriate.
Fairness	Use questions that are clearly written.
	Write questions that cover a broad range of topics.
	Use questions that are consistent with the syllabus.
	Inform students about the knowledge to be assessed.
Feedback	Utilize elaborative feedback.
	Timing (i.e., immediate or delayed) of feedback should be based on difficulty level of item and context of the assessment.
	Give students opportunities to self-correct.
	Provide elaborate and timely feedback for online assessments using software.
	Solicit feedback on assessments from students.
Formatting and content	Utilize 3-choice items.
	Avoid questions that use negatives.
	Avoid multipart and giveaway questions.
	Avoid "none of the above" questions.
	Avoid composite answers such as "A and B but not C."
	Avoid "all of the above" questions.
	Choices should be parallel in structure and equal in length.
	Question stems should be as short as possible, but contain all relevant information.
	Randomize answer positions.
Cheating countermeasures	Use alternate test forms.
	Utilize alternate or assigned seating for assessments.
	Provide students with an academic integrity policy.
	Utilize honor codes and academic honesty agreements.
	Draw from a large question bank and randomize question answers and order.
	Defer question feedback until after online assessments close.
	Change exam questions between semesters.
	Utilize "lockdown browsers" for online classes.

Table 2: Optimizing Multiple-Choice Question Testing (Xu, X., Kauer, S., & Tupy, S., 2016, p 148).

Conclusion

Educators have an obligation to ensure that the assessments they create accurately measure learning objectives. Multiple-Choice questions have long been the preferred method and most commonly used in classrooms, providing for both reliable and valid coverage of content (Begum, 2012). Properly constructed MCQs have specific pedagogical benefits that enhance simple recall of facts and provide opportunities to engage in deep processing of higher-order thinking skills (Little, 2012). Creating quality MCQs begins with deliberate thought addressing learning objectives that focus on important concepts of course content and assesses application of knowledge rather than simple recall of isolated facts. Although creating MCQs that assess higher-order thinking skills is a challenging and time-consuming task, it is a possible and worthwhile creative endeavor.

Haladyna, Downing, and Rodrigez (2002) remind us of the words of Ebel (1951):

"Each item as it is being written presents new problems and new opportunities. Just as there can be no set formulas for producing a good story or a good painting, so there can be no set of rules that will guarantee the production of good test items. Principles can be established and suggestions offered, but it is the item writer's judgment in the application (and occasional disregard) of these principles and suggestions that determines whether good items or mediocre ones will be produced" (p. 185).

References

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruickshank, K. A., Mayer, R. E., Pintrich, P. R., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. (Complete edition). New York: Longman.

Azer, S. A. (2003). Assessment in a problem-based learning course: Twelve tips for constructing multiple choice questions that test students' cognitive skills. *Biochemistry and Molecular Biology Education, 31*(6), 428-434.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple-choice items. *Psychological Test and Assessment Modeling, 53*(2), 192-211.

Barlow, P. B. (2014). Development of the biostatistcs and clinical epidemiology skills assessment for medical residents. PhD Dissertation, University of Tennessee. Retrieved from: https://trace.tennessee.edu/utk_graddiss/2676

Begum, T. (2012). A guideline on developing effective multiple-choice questions and construction of single best answer format. *Journal of Bangladesh College of Physicians and Surgeons, 30*(3), 159-166.

Bloom, B. S. (1956). ***Taxonomy*** *of educational objectives: The classification of educational goals.* (1st ed.). Harlow, Essex, England: Longman Group.

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple choice questions. *Academic Psychiatry, 34(*4), 310-316.

Brookhart, S. M. (2015). Making the most of multiple choice. *Educational Leadership, 73*(1), 36-39.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Retrieved from https://testing.byu.edu/handbooks/betteritems.pdf

Campbell, D. (2011). How to write good multiple-choice questions. *Journal of Paediatrics & Child Health, 47*(6), 322-325.

Clay, B. (2001). *Is this a trick question? A short guide to writing effective test questions.* Retrieved from http://www.k-state.edu/ksde/alp/resources/Handout-Module6.pdf

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education, 36*, 800-804.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109–17. doi:10.1111/j.13652923.2009.03425.x

Domyancich, J. M. (2014). The development of multiple-choice items consistent with the AP chemistry curriculum framework to more accurately assess deeper understanding. *J. Chem. Educ., 91*(9), 1347–1351.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424.

Frary, R. B. (1995). *Multiple-choice item writing do's and don'ts*. ERIC Clearinghouse on Assessment and Evaluation Washington DC. (ED398238)

Funk, S. C., & Dickson, K. L. (2011). **Multiple-choice** and short-answer exam performance in a college classroom. *Teaching of Psychology, 38*(4), 273-277.

Gareis, C. R., & Grant, L. W. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning.* New York: Routledge.

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A. P., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education, 29*(3), 196-210.

Gutl, C., Lankmayr, K., Weinhofer, J., & Hofler, M. (2011). Enhanced automatic question creator--EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning, 9*(1), 23-38.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a **multiple-choice** test **item**? *Educational and Psychological Measurement, 53*(4), 999-1010.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. *Journal of Education for Business, 73*, 94-97.

Harmon, O. R., Lambrinos, J., & Buffolino, J. (2010). Assessment design and cheating risk in online instruction. *Online Journal of Distance Learning Administration*, *13*(3).

Iwaoka, W. T., Li, Y., & Rhee, W.Y. (2010). Measuring gains in critical thinking in food science and human nutrition course: The Cornell critical thinking test, problem-based learning activities, and student journal entries. *Journal of Food Science Education, 9*(3), 68-75.

Kehoe, J. (1995). *Writing Multiple-Choice Test Items.* ERIC Clearinghouse on Assessment and Evaluation Washington DC. (ED398236)

Khan, H. F., Danish, K. F., Awan, A. S., & Anwar, M. (2013). Identification of technical item flaws leads to improvement of the quality of single best multiple-choice questions. *Pakistan Journal of Medical Sciences, 29*(3), 715-718.

Kong, L. N., Qin, B., Zhou, Y. Q., Mou, S. Y., & Gao, H. M. (2009). The effectiveness of problem-based learning on development of nursing students' critical thinking: A systematic review and meta-analysis. *International Journal of Nursing Students, 51*(3), 458-469. doi: 10.1016/j.ijnurstu.2013.06.009.

Little, J. L. (2012). Optimizing multiple-choice tests as learning events. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 73*(5-B), 2012. pp. 3284.

Lumley, J. S. P., & Craven, J. L. (2004). (Eds). *MCQs in Anatomy.* 3rd ed. New York: Churchill Livingstone; 2004: p-2-4

McCoubrie, P., & McKnight, L. (2008). Single best answer MCQs: A new format for the FRCR part 2a exam. *Clinical Radiology 63*(5), 506-10. doi: 10.1016/j.crad.2007.08.021.

Miller, P. W., & Erickson, H. E. (1985). *How to write tests for students.* National Education Association, Washington, D.C. (ED322157)

Moreno, R., Martínez, R.J., & Muñiz, J. (2006). New guidelines for developing multiple-choice Iitems. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2*(2), 65-72. Publisher: Hogrefe & Huber Publishers.

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*(4), 388-394.

Morrison, S., & Free, K. W. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nurse Education, 40*(1), 17-24.

Nath, L., & Lovaglia, M. (2009). Cheating on multiple-choice exams: Monitoring, assessment, and an optional assignment. *College Teaching, 57*(1), 3-8.

Ory, C. J. (n. d.). Improving your test questions, evaluation and examination service.

University of Iowa. Retrieved from http://www.uiowa.edu/~examserv/Level_2/resources /Technical%20Bulletins/Tech%20Bulletin%2027.pdf

Pachai, M. V., DiBattista, D. K., & Joseph, A. (2015). A Systematic Assessment of "None of the Above" on Multiple Choice Tests in a First Year Psychology Classroom. *Canadian Journal for the Scholarship of Teaching and Learning, 6*(3), Article 2.

Parmenter, D. A. (2009). Essay versus multiple choice: student preferences and the underlying rationale with implications for test construction. *Academy of Entrepreneurship Journal*, *15*(2), 57-71.

Pettijohn, T. F., & Sacco, M. F. (2007). Multiple-choice exam question order influences on student performance, completion time and perceptions. *Journal of Instructional Psychology, 34*(3), 142-149.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Learning, Memory, and Cognition, 31*(5), 1155-1159.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation, 22*(4), 2.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Journal of Innovative Education, 3*, 73–98.

Sireci, S. G., Wiley, A., & Keller, L. A. (1998). *An Empirical Evaluation of Selected Multiple-Choice Item Writing Guidelines.* Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 28-30, 1998). (ED428122)

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*(7), 784-802.

Suski, L., & Banta, T. (2009). *Assessing Student Learning: A Common Sense Guide*. San Francisco: Jossey-Bass.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42,* 198–206.

Tellinghuisen, J., & Sulikowski, M. M. (2008). Does the answer order matter on **multiple-choice exams**? *Journal of Chemical Education, 85*(4), 572-575.

Timmermann, D., & Kautz, C. H. (2015). *Multiple choice questions that test conceptual understanding: A proposal for qualitative two-tier exam questions.* Proceedings of the ASEE Annual Conference & Exposition. 2015, p1-15.

Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2011). Improving **multiple-choice questions.** *US-China Education Review B 1*, 1-11.

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education, 91*(9), 1426-1431.

Vander Schee, B. A. (2013). Test item order, level of difficulty, and student performance in marketing education. *Journal of Education for Business, 88*, 36–42.

Vacc, N. A., Loesch, L. C., & Lubik, R. E. (2001). Writing multiple-choice test items. In G. Walz & J. Bleuer (Eds.) *Assessment: Issues and Challenges for the Millennium.* CAPS: Greensboro, NC.

Weinstein, Y., & Roediger, H. L. III. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition, 38*(3), 366-376.

Weinstein, Y., & Roediger, H. L. III. (2012). The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition, 40*(5), 727-735.

http://dx.doi.org/10.3758/s13421-012-0187-3

Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology, 2*(2), 147-158.

Zimmaro D. (2004). *Writing good multiple-choice exams.* Measurement and Evaluation Center: University of Texas, Austin.