View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HKU Scholars Hub

The HKU Scholars Hub The University of Hong Kong 香港大學學術庫

Title	A Framework for improving the quality of multiple-choice assessments
Author(s)	Tarrant, M; Ware, J
Citation	Nurse Educator, 2012, v. 37 n. 3, p. 98-104
Issued Date	2012
URL	http://hdl.handle.net/10722/149231
Rights	This is a non-final version of an article published in final form in Nurse Educator, 2012, v. 37 n. 3, p. 98-104

A Framework for Improving the Quality of Multiple-Choice Assessments

Marie Tarrant, RN, MPH, PhD; James Ware, MA MB, FRCS, DMSc

Authors' Affiliations: Associate Professor, School of Nursing, Li Ka Shing Faculty of Medicine, 21 Sassoon Road, Hong Kong; Director, Medical Education and Postgraduate Studies, The Saudi Commission for Health Specialties, Riyadh, Saudi Arabia.

Corresponding Author: Marie Tarrant, School of Nursing, 4/F, William M. W. Mong Block, Li Ka Shing Faculty of Medicine, 21 Sassoon Road, Hong Kong (tarrantm@hku.hk).

Conflicts of Interest: The authors report no conflicts of interest. The authors alone are responsible for the content of this article.

Sources of Funding: None

Keywords: multiple-choice questions, multiple-choice tests, test construction, quality assurance, assessment.

Abstract

Multiple-choice questions are frequently used in high-stakes nursing assessments. Many nurse educators, however, lack the necessary knowledge and training to develop these tests. This paper will discuss test development guidelines to help nurse educators produce valid and reliable multiple-choice assessments. These guidelines for multiple-choice test development can be divided into three categories: (1) pre-test planning, (2) test development practices, and (3) post-test review.

Introduction

Multiple-choice (MC) questions are a widely used selection-type test format (1). Single best-answer MC questions consist of a question or a problem (*the stem*), two or more choices from which examinees must choose the correct option (*the distractors*) and one correct or best response (*the key*) (2). MC items allow teachers to efficiently assess large numbers of candidates and to test a wide range of content and learning objectives (3, 4). Well-constructed MC questions are able to test higher levels of cognitive reasoning and can accurately discriminate between high- and low-achieving examinees (3, 5). Consequently, MC questions are frequently used in high-stakes assessments in nursing and other health science disciplines.

Most non-certification level tests taken by nursing students are developed in-house by nursing faculty members who teach the courses. Few nurse educators, however, have adequate preparation and knowledge of how to develop high-quality MC tests. Educators usually either develop the test items themselves or rely on item test banks as a source of questions, both of which may result in less than optimal test quality. Thus, there can be substantial deficiencies in tests prepared by course teachers (6). Because student learning is largely driven by tests, careful test construction is an important skill for educators to develop (7). We have reviewed and synthesized the research literature related to the issue of increasing quality in MC tests. The purpose of this paper is to present a set of clear guidelines for both novice and experienced nurse educators responsible for test development to help them produce high quality MC tests. The focus

is on important and key test development practices and not on logistic or other organizational factors related to testing. The guidelines are divided into three categories: (1) pre-test planning, (2) test development practices and (3) post-test review.

Pre-test Planning

Provide Training for Item Writers

Discipline-based higher education in nursing means that many educators have not had any formal training in assessment methods and test construction. Nurse educators are often hired because of their clinical expertise. This expertise, however, does not ensure that they can develop high quality tests in their own discipline. In addition, test development procedures are largely passed down from senior to junior academics and are often not evidence-based (8). Without appropriate training, most novice item writers will develop low-quality test items that test only factual recall or trivial content (9).

Studies across disciplines have shown that teacher-produced tests are much improved by prior training in writing MC items. In one study, where the quality of test produced in-house at three medical schools was examined, it was found that items produced by educators who had received training were of much higher quality than those written by untrained educators (10). A review of MC items in certification-level accounting examinations, the majority produced by trained staff, found the quality of the tests to

be of considerably higher quality than that to be found in textbooks or test banks (11). However, writing good MC items is time-consuming and difficult work. It is therefore necessary for academic institutions that employ expert clinicians as academic staff to provide the training and instruction they need to become capable members of the nursing faculty (12). If training is not immediately available, Case and Swanson (13) have produced an excellent manual that covers many key issues in developing high-quality MC items.

Develop a Test Blueprint

When developing MC items, focus on important topics only – usually common or important clinical problems. Tests should focus on the learning objectives of the course, not trivia. A test blueprint will help with this task. A blueprint is simply a grid or table that maps the course objectives and content to be tested, and is a necessary step in producing a valid and reliable test (see Figure 1). A test blueprint will precisely outline the proportion of test questions to be allocated to the various content areas and the cognitive level the questions are written at (9). The weighting of exam content is usually designed to approximate the weighting of course content (14). Further guidelines exist in the literature for developing test blueprints (14-16). Nurse educators should review the course objectives, particularly the verbs used, and ensure the test items are written to be consistent with the skills students are expected to demonstrate after completion of the course.

Test Development Practices

Write Questions to Test Higher Cognitive Thinking

Nurses are required to process a great deal of complicated information to arrive at the right decisions on patient care (17). Such complex cognitive abilities must consequently be tested during their education to ensure that students can later operate at a high level of cognition whenever necessary (18). However, MC questions often do not reach this level. A review of items in nursing textbooks, specifically of their cognitive levels, found that 72.1% were assessing only knowledge and comprehension (17), the two lowest levels of Bloom's Taxonomy (19). A further review, of the quality of 2770 test items used over a five-year period in one nursing school found that 91.1% were written at the knowledge and comprehension levels (20).

It is often argued that the MC format, by its nature, can only deal with the repetition of factual material (21). But MC items can and should be written to assess higher-level cognitive processes such as application and analysis (13, 22, 23). Although there is no research literature to indicate what proportion of test items should be written at higher cognitive levels (17), it is probably safe to assume that the weighting of tests in nursing, whose practice requires higher degrees of cognition, should reflect that fact.

Use Clinical Vignettes

Test items should present clinical decision-making tasks within the education and experience of examinees and the use of well-constructed problem-solving clinical

vignettes will help with this process and increase the likelihood that questions are testing higher cognitive levels rather than just recall of isolated facts (24). A vignette includes a description of the patient and/or situation and some subjective and objective data, some of which is pertinent to answering the question. The use of a vignette requires students to go beyond simply recalling information. Students have to apply their memorised knowledge to make a judgment or to solve a patient problem – a situation that is similar to what they will face in real life (13). An additional advantage of using a clinical vignette is that more than one question can be constructed using the same clinical scenario. Real life clinical cases provide an excellent source of clinical vignettes to use in developing MC questions.

Write Only Plausible Options

In teacher-produced MC tests, the question stem often receives far more attention from the writer than the distractors, with the result that the latter are often ineffective. Clearly, however, developing plausible options to the correct answer is of great importance for a high-quality test (25). When educational outcomes are being assessed in a classroom by means of a MC test, the distractors must be effective, each one centring (where possible) on widespread errors about the correct answer (26). A high proportion of questions on teacher-generated tests however, have one or more implausible distractors (20, 27, 28). In a four-option item, it may be hard to come up with three distractors that are of more or less equal likelihood, and thus 'fillers' are added. It is commonly believed that MC items must have at least four or five options. A

question with only two good distractors, however, is preferable to one with additional filler options added only to make up some pre-determined number of options (5, 29). Such implausible distractors can be easily spotted by even the weakest examinees, and are therefore usually rejected outright.

Research has repeatedly shown that, in most situations, a three-option MC item is preferable (i.e., one containing the key plus two distractors) (30). One reason 3-option items are endorsed is that most 4- and 5-option items have at most only two plausible distractors. Haladyna and Downing (26) reviewed functioning distractors on four, 5-option MC assessments in one medical school and found that only 1.1 to 8.4% of all the items had three distractors that functioned appropriately. Little difference was found in item difficulty and discrimination between questions containing two, three, or four functioning distractors. In a review of seven, 4-option tests in nursing, only 5.7 to 26.1% of all items had three functioning distractors (31). Furthermore, when 4-option and 3-option tests were compared, there were no substantial changes in mean test score, pass rates, test reliability, item difficulty or item discrimination (28).

To make distractors more plausible, use students' most common errors or misunderstandings as options. Use words that have verbal associations with the item stem (e.g., gastrointestinal, stomach; cardiac, heart etc.) or textbook language that has the appearance of truth. All distractors should be homogeneous and parallel (e.g., all drugs, all diagnoses, all treatments etc.), equally plausible while incorrect or inferior to

the correct answer, attractive to the uninformed, similar in length and construction to the correct answer, and grammatically consistent with the stem. Distractors should distract the uninformed, but they should not result in trick questions that mislead knowledgeable examinees.

Write a Sufficient Number of Items

To maximize sampling of course content, MC tests require a large number of items. In addition, test reliability is increased with more test items. If MC test items are, on average, moderately discriminating, at minimum of 50 to 60 items are needed to achieve a high level of reliability (>.80); if the average item discrimination is low, then at least 100 items are needed (32). As it is more difficult to achieve high reliability and adequate content sampling with a low number of items, teachers are encouraged to write as many items as is feasible. One way to increase the number of test items is to write 3-option items instead of 4- or 5-option items. Because 3-option items perform equally as well as items with more options and they are more efficient to write and administer, teachers can write more items with three options in the same time required to write items with four or more options. In addition, examinees can answer 12.4 extra 3-option MC questions in the same amount of time as 100 4-option items (33).

Distribute Correct Answers Randomly and Evenly

Correct answers should be evenly distributed among the available options and arranged in a random pattern (9, 34, 35), with the exception of numerical options, which should

be arranged in either ascending or descending order. If there is an even distribution of correct responses in an MC test, each option would be correct approximately 25% of the time on a 4-option test and 20% on a 5-option test. When answering questions to which they do not know the answers, examinees often revert to the rule of `when in doubt, choose C.'

In fact, both examinees and item writers have a bias toward the middle position, so unless specific attention is paid to the organization of the correct answers, the correct answer will more often be a middle option which in turn, examinees will tend to select more frequently (36). This has been demonstrated in several studies. In one study, among over 1000 4-option MC questions, option C was the most frequent correct answer (27.6%) and option A the least (21.1%)(37). A study of 5-option MC items found that option E was correct only 5% of the time (38). Randomizing and balancing the position of the correct answer so that there is an equal frequency and distribution of the correct response is an important principle for item writers and test developers to remember and follow.

Screen for and Remove Item Writing Flaws

A common issue affecting MC questions in teacher-generated tests is the presence of item-writing flaws. Item writing flaws can be simply described as violations of conventional item-writing principles that can affect a examinee's test performance, making items either more or less difficult to answer correctly (1, 39, 40). Although a full

description and discussion of such flaws is beyond the scope of this paper and can be found elsewhere (39, 41), a brief summary of common mistakes made by novice item writers is provided in Table 1.

Other violations include the introduction of either linguistic or cultural bias into the test items. Linguistic bias is present when test items contain complex or unnecessary information that can increase the item difficulty (42) and cultural bias is present when items include references to the dominant culture which may not be well understood by members of other cultures (43). With the increasing diversity of baccalaureate nursing programs in most countries, it is increasingly important to ensure that MC test items are free of bias and do not disproportionately disadvantage students from diverse backgrounds (43).

In four high-stakes medical school examinations, Downing (27) found that 33–46% of the MC items were flawed and that 10-15% of examinees who failed would have passed if those flawed items had not been present. Another review of 10 examinations given to nursing students found that 47.3% of all items were flawed and over one-half of these flaws were related to linguistic or structural bias, which can make test items more difficult (44). Furthermore, if the flawed items had been removed from the test, fewer lower-achieving examinees would have passed the test (90.6% vs. 94.3%) and more higher achievers would have obtained a score of ≥ 80% (20.9% vs. 14.5%) (44).

Avoid or Limit Use of Items from Commercial Item Banks

Studies have documented the poor quality of MC questions in textbooks and commercial test banks provided by text book publishers. Masters et al. (17) found 2233 item-writing flaws in 2913 questions in test banks accompanying nursing textbooks. Similar examples occur in other disciplines: 75% of MC questions in accounting test banks were found to contain at least one item-writing flaw (11) and about 60% of items in instructor guides accompanying introductory psychology textbooks were also flawed (37). Test banks are often provided to teachers as an incentive to adopt a textbook for the course but textbook authors often do not have formal preparation in MC item construction or are not the persons actually developing the test bank items. Hence, questions derived from textbooks and test banks are as likely to contain item-writing flaws as those developed by educators.

Do a Pre-Test Review of Items

Although most educators spend a substantial amount of time developing course materials and planning lectures, they often spend less time preparing tests and reviewing them before administration (10, 45). As a result, often tests are administered without being first submitted to an adequate quality review process. Even carefully developed test items written by experienced item writers should be subjected to adequate review prior to administration (46). Therefore, a process that includes peer review by additional content experts and review by an examinations committee whose members are well versed in item writing will help to ensure that test items are of

suitable quality and that they test higher cognitive domains (45). Bush (47) outlined some specific points to note when reviewing multiple choice items prior to test administration (see Figure 2).

Post-test Review

Perform Post-Test Item Analysis

With any assessment measure, the feedback loop requires that the educator also assess the quality of the tools being used in the assessment to ensure that they have achieved the purposes for which they were originally intended (48). After the test has been administered, the performance of each item, and of the test as a whole, should be evaluated using standard item analysis procedures (49, 50). Item analysis involves relating the statistical properties of test items to a response distribution (2). The primary purpose of item analysis is to gain information about the tests items rather than the examinees (51).

Item analysis is one of the most important parts of the quality assurance process. Basic item and test statistics such as item difficulty index (the proportion of examinees answering the item correctly), item discrimination index (the difference between the proportion of high and low achieving examinees who answered the item correctly), test reliability, and mean test score are calculated for further analysis (see Table 2). It has been estimated that more than half the test items that educators write will not produce the intended results (35). Therefore, poorly functioning items (i.e., items that are either

too easy or too difficult) or items with unusual answering patterns (items for which there is no expected increase in the correct answers among examinees with higher scores) can be identified and then either edited or removed from further use (7). In addition, items that unfairly penalize a large proportion of examinees can be excluded from the calculation of the final grade on the current test.

A wide range of software packages are available that can calculate these item and test parameters and present the item analysis results either numerically or graphically (i.e., item response curves or quintile plots). Additionally, most spread sheet software can also be easily programed to perform item analysis. Item analysis is thus a critical step in the development and review of tests as it provides important data for item and test improvement. It is only through this process of item analysis and improvement that tests can be developed in ways that are psychometrically and pedagogically sound (47).

Develop a Bank of High Quality Items

All test items need editing and refinement. The process of developing a new test on each occasion is time-consuming and does not capitalize on previous work. Good MC tests cannot be developed if test items are selected or used indiscriminately and not evaluated after use. Item analysis procedures can help to identify items that perform well, and an item bank can help to organize and categorize these items for quick and easy retrieval in future examinations (52). A number of software programs are available for item banking that allow the users to organize the test items by a number of

different parameters (e.g., item type, cognitive level, body system, discipline, course name, item analysis statistics, or even MeSH terms). All item banking programs must have the capability to securely store items and retrieve items purposely or randomly according to any of the above parameters (53). An item bank reduces the burden of creating assessments by assisting schools to build up a sizeable pool of high quality test items that can be re-used in subsequent tests (24).

Summary

Valid, high-quality assessments requires the establishment of rigorous procedures to review both test quality before and test results after administration. In professional nursing education programmes, educators have to account to many other parties, including licensing authorities, health-care institutions, patients, and the general public. A responsibility therefore exists, in both ethical and legal dimensions, to ensure that assessments are valid, and are assessing what they are supposed to. Few papers specifically outline the steps to be followed to ensure high quality tests. However, the quality of educational assessments of nurses and other health professionals is an issue receiving increased attention as a result of a greater focus on outcome-based assessments in tertiary educational institutions and the increased accountability the public is demanding from institutions that produce health professionals. Hence, clear research-based test development guidelines are required.

References

1. Gronlund NE, Waugh CK. *Assessment of student achievement*. 9th ed. Upper Saddle River, N.J.: Pearson; 2008.

2. Osterlind SJ. *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. 2nd ed. Boston: Kluwer Academic Publishers; 1998.

3. Downing SM. Assessment of knowledge with written test forms. In: Norman GR, Van der Vleuten C, Newble DI, eds. *International handbook of research in medical education.* Vol II. Dorcrecht: Kluwer Academic Publishers; 2002:647-672.

4. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach.* 2004;26(8):709-712.

5. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ.* 2004;38(9):974-979.

6. Mehrens WA, Lehmann IJ. *Measurement and evaluation in education and psychology*. 4th ed. Fort Worth, TX: Holt, Rinehart and Winston; 1991.

7. Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, van der Vleuten CPM. Quality assurance in test construction: the approach of a multidisciplinary central test committee. *Educ Health.* 1999;12(1):49-60.

8. Nitko AJ. Book reviews: Roid, G. H., and Haladyna, T. M. A Technology for Test-Item Writing. New York: Academic Press, 1982, xii + 247 pp. *J Educ Meas.* 1984;21(2):201-204.

9. Downing SM. Twelve steps for effective test development. In: Downing SM, Haladyna TM, eds. *Handbook of test development*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers; 2006:3-25.

10. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med.* 2002;77(2):156-161.

11. Hansen JD. Quality multiple-choice test questions: Item writing guidelines and an analysis of auditing test banks. *J Educ Bus.* 1997;73(2):94-97.

12. Morrison S, Free KW. Writing multiple-choice test items that promote and measure critical thinking. *J Nurs Educ.* 2001;40(1):17-24.

13. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia, PA: National Board of Medical Examiners; 2001; Available: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf. Accessed November 22, 2011.

14. Coderre S, Woloschuk W, McLaughlin K. Twelve tips for blueprinting. *Med Teach.* 2009;31(4):322-324.

15. Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: methods for developing content-valid student examinations. *Med Teach.* Jul 2003;25(4):414-421.

16. Farley JK. The multiple-choice test: developing the test blueprint. *Nurse Educ.* 1989;14(5):3-5.

17. Masters JC, Hulsmeyer BS, Pike ME, Leichty K, Miller MT, Verst AL. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *J Nurs Educ.* 2001;40(1):25-32.

18. Clifton SL, Schriner CL. Assessing the quality of multiple-choice test items. *Nurse Educ.* 2010;35(1):12-16.

19. Bloom BS. *Taxonomy of educational objectives. Handbook 1: The cognitive domain.* . 1st ed. London: Longman; 1956.

20. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Educ Today.* 2006;26(8):662-671.

21. Pamplett R, Farnhill D. Effect of anxiety on performance in multiple-choice examinations. *Med Educ.* 1995;29(4):297-302.

22. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. *BMJ.* 2003;326(7390):643-645.

23. Su WM. Writing context-dependent item sets that reflect critical thinking learning outcomes. *Nurse Educ.* 2007;32(1):11-15.

24. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Med Educ.* 2004;4(1):23.

25. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ.* 1989;2(1):51-78.

26. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educ Psychol Meas.* 1993;53(4):999-1010.

27. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ.* 2005;10(2):133-143.

28. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Educ Today.* 2010;30(6):539-543.

29. Crehan KD, Haladyna TM, Brewer BW. Use of an inclusive option and the optimal number of options for multiple-choice items. *Educ Psychol Meas.* 1993;53(1):241-247.

30. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educ Meas Issues Pract.* 2005;24(2):3-13.

31. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ.* 2009;9(40).

32. Hopkins KD. *Educational and psychological measurement and evaluation*. 8th ed. Boston: Allyn & Bacon; 1998.

33. Aamodt MG, McShane T. A meta-analytic investigation of the effect of various test item characteristics on test scores. *Public Pers Manage.* 1992;21(2):151-160.

34. Oermann MH, Gaberson KB. *Evaluation and testing in nursing education*. 3rd ed. New York: Springer; 2009.

35. Haladyna TM. *Developing and validating multiple-choice test items*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum; 2004.

36. Attali Y, Bar-Hillel M. Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *J Educ Meas.* 2003;40(2):109-128.

37. Ellsworth RA, Dunnell P, Duell OK. Multiple-choice test items: What are textbook authors telling teachers? *J Educ Res.* 1990;83(5):289-293.

38. Clute RC, McGrail GR. Bias in examination test banks that accompany cost accounting texts. *J Educ Bus.* 1989;64(6):245-247.

39. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ.* 2002;15(3):309-334.

40. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Adv Health Sci Educ.* 2002;7(3):235-241.

41. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. *Appl Meas Educ.* 1989;2(1):37-50.

42. Bosher S. Barriers to creating a more culturally diverse nursing profession: linguistic bias in multiple-choice nursing exams. *Nurs Educ Perspect.* 2003;24(1):25-34.

43. Hicks NA. Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. *Nurse Educ.* 2011;36(6):266-270.

44. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ.* 2008;42(2):198-206.

45. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ.* Feb 2006;11(1):61-68.

46. Baranowski RA. Item editing and editorial review. In: Downing SM, Haladyna TM, eds. *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum; 2006:349-357.

47. Bush ME. Quality assurance of multiple-choice tests. *Qual Assur Educ.* 2006;14(4):398-404.

48. Ebel RL, Frisbie DA. *Essentials of educational measurement*. 5th ed. Englewood Cliffs, N.J.: Prentice Hall; 1991.

49. Farley JK. Item analysis. *Nurse Educ.* 1990;15(1):8-9.

50. Jenkins HM, Michael MM. Using and interpreting item analysis data. *Nurse Educ.* 1986;11(1):101-104.

51. Livingston SA. Item analysis. In: Downing SM, Haladyna TM, eds. *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum 2006:421-441.

52. McDonald ME. *The nurse educator's guide to assessing learning outcomes*. 2nd ed. Sudbury, MA: Jones and Bartlett; 2007.

53. Vale CD. Computerized item banking. In: Downing SM, Haladyna TM, eds. *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum 2006:261-285.

2R. Table (information in 2 or more columns)

Table 1. Guidelines for Avoiding Common Item Writing Flaws in Multiple-Choice Questions	
Guideline	Rationale
Make all options grammatically consistent with the stem	The correct option is more likely to flow grammatically from the item stem, which can cue examinees to the correct answer.
Place numeric option data in either ascending or descending order.	Numeric data that is properly sequenced decreases confusion for examinees and improves the appearance and neatness of the test.
All options should be equal in length and amount of detail.	Item writers often make the correct option longer and include more information to ensure that it is unambiguously correct. When examinees are unsure of the correct answer, a common practice is to select the longest option as they often correctly perceive that it is more likely to be correct.
Make all options equally plausible.	Implausible distractors can often be eliminated by even the weakest examinees and this increases the chances of students guessing the correct option without actually knowing the material. Good test items depend on having effective distractors.
Make sure the stem asks a clear question that can be answered without looking at the options.	The stem should present a clear and focused question that can be understood and answered by knowledgeable examinees without looking at the options. The options should not be a set of true/false statements.
Ensure that items have one, and only one, correct answer.	Single best-answer MC questions should have one, and only one, correct answer.
Do not place information in the stem that is not required to answer the question.	Unnecessary information in the stem that is not required to answer the question increases reading time and can unnecessarily confuse examinees. If a clinical vignette is provided with the question, it should be required to answer the question.
Avoid word repeats in the stem and the correct option.	Repeating the same or similar words in the stem and correct option cues the examinee to the correct answer. For example, using the word ‘cardiac’ in the stem and only in the correct option.
Avoid logical clues in the stem.	Similarly, do not provide information in the stem or in the options that make the correct answer more obvious. For example, if you ask a question about 'pharmaceutical interventions' ensure that all options are actually pharmaceutical interventions

Table 1. Guidelines for Avoiding Common Item Writing Flaws in Multiple-Choice Questions	
Guideline	Rationale
Avoid the use of absolute terms in the options (e.g., always, never, all, only).	Absolute terms provide cues to examinees as most students are aware that these options are rarely correct and thus can easily be eliminated as correct answers.
Avoid the use of vague terms (e.g., frequently, often, occasionally) in the options.	There is seldom universal agreement on the actual interpretation of these terms and they can confuse examinees.
Avoid the use of negative words in the stem (e.g., except, not, incorrect)	Negatively worded stems are less likely to measure important learning outcomes and can confuse examinees.
Avoid the use of 'all of the above' and 'none of the above.'	'All of the above' and 'none of the above' are often used a fillers when item writers have difficulty coming up with a fourth or fifth option. Furthermore, they allow examinees to answer questions based on partial information. If the examinees know that more than one of the options is correct, the 'all of the above' is most likely the answer. Similarly, if the examinees know that at least one option is not correct, they can eliminate 'all of the above' as the correct answer. The use of 'none of the above' can produce a similar cuing effect.

2R. Table (information in 2 or more columns)

Table 2. Description and Interpretation of Item-Analysis Data From Multiple-Choice Tests				
Item/Test Statistic	What is Measured	Range of Values	Interpretation*	Explanation / Rationale
Difficulty Index *(often referred to as the P-value)*	The proportion of examinees who answered the item correctly.	0 to 1.00	• Low difficulty: >.80 • Medium difficulty: .30 to .80 • High difficulty: <.30	The majority of tests items should have medium difficulty levels as they are better able to discriminate between high and low achieving examinees. Items that are either too easy or too difficult cannot discriminate.
Discrimination Index	The difference in the number of high achieving and low achieving examinees who answered the question correctly.	-1.00 to 1.00	• Excellent discrimination:≥ 0.40 • Good discrimination: 0.30 to 0.39 • Satisfactory discrimination: 0.15 to 0.29 • Low discrimination: < 0.15 • No discrimination: ≤ 0	Good test items are answered correctly more frequently by higher achieving examinees. Thus, test items with high discrimination are desired.
Point-biserial Correlation (RPB) Coefficient	The RPB is another measure of discrimination that is similar to, but more robust than, the Discrimination Index. It is the correlation between how well examinees did on the item and their total test score.	-1.00 to 1.00	• Interpreted the same as the discrimination index.	Items with higher RPB coefficients are more discriminating. RPB ≤ 0 indicates items in which lower achieving examinees performed better than higher achieving examinees. These items are problematic and should be reviewed closely.
Distractor Frequency	The proportion of people selecting each distractor.	0 to 1.00	• Functioning distractor: >.05 • Poorly functioning distractor: <.05 • Non-functioning distractor: = 0	Distractors selected by <5% of examinees are so implausible that even the weakest examinees can eliminate them as correct answers. These distractors should be replaced with more plausible options.
Distractor Discrimination	The difference in the number of high achieving and low achieving students who select each distractor.	-1.00 to 1.00	• Discriminating distractor: < 0 • Poorly discriminating distractor: > 0	Good distractors should appeal to a greater number of lower achieving examinees than higher achieving examinees. Thus, the distractors should have a negative discrimination statistic.

Table 2. Description and Interpretation of Item-Analysis Data From Multiple-Choice Tests				
Item/Test Statistic	What is Measured	Range of Values	Interpretation*	Explanation / Rationale
Test Reliability	A reliability coefficient that indicates the homogeneity of the test items.	0 to 1.00	• Good test reliability: >.80 • Acceptable test reliability: .70 to .80 • Poor test reliability: < .70	As with any measure of reliability, test reliability is closely related to the number of test items. Tests with fewer test items will rarely produce acceptable reliability coefficients.

*There is some minor variation in the literature about the cutoff values for some of these item parameters.

3R. Figure (information not in column format)

Course Content (Objectives)	% of Course Content	Cognitive Level				Total N (%)
		Knowledge	Comprehension	Application	Analysis	
Objective 1	30	4	8	10	4	26 (32.5)
Objective 2	25	3	6	4	5	18 (22.5)
Objective 3	15	3	3	4	4	14 (17.5)
Objective 4	20	4	5	2	3	14 (17.5)
Objective 5	10	2	1	3	2	8 (10)
Total	100	16	23	23	18	80 (100)

Figure 1. Example of a test blueprint for an 80-item multiple-choice test

3R. Figure (information not in column format)

Six key areas identified by Bush (47) for peer reviewers to assess when reviewing multiple-choice test items:

1. Is the question clear and unambiguous?
2. Are there uncommon words or phrases that could be replaced with more familiar words with the same meaning?
3. Are any of the distractors too obviously correct or incorrect?
4. Are there any overlapping questions?
5. Do the questions collectively cover the subject matter?
6. Are there enough items on the test to sufficiently cover the subject matter without overlapping questions?

Figure 2. Points for peer-reviewers of multiple-choice test questions