JNPD Journal for Nurses in Professional Development • Volume 29, Number 2, 52–57 • Copyright © 2013 Wolters Kluwer Health | Lippincott Williams & Wilkins

Assessment of Item-Writing Flaws in Multiple-Choice Questions

CE

Rosemarie Nedeau-Cayo, MSN, RN-BC ❍ Deborah Laughlin, MSN, RN-BC ❍ Linda Rus, MSN, RN-BC ❍ John Hall, MSN, RN

This study evaluated the quality of multiple-choice questions used in a hospital's e-learning system. Constructing well-written questions is fraught with difficulty, and item-writing flaws are common. Study results revealed that most items contained flaws and were written at the knowledge/comprehension level. Few items had linked objectives, and no association was found between the presence of objectives and flaws. Recommendations include education for writing test questions.

In the hospital setting, multiple-choice questions (MCQs) are used to test large numbers of staff in a cost-effective manner. Farley (1989a) stated that multiple-choice tests are "desirable because they assess a broad range of content in a short period of time. They are objective, accurate, easily scored and readily adap to a variety of content" (p. 10). These statements prove to be valid as long as the MCQs are well written. Poorly written test questions do not validate learning and waste resources in the form of time, money, and productivity for the employer and employee.

BACKGROUND

Efforts to establish quality test-question criteria have been the subject of multiple articles and textbooks since the early 1900s. Haladyna and Downing (1989a) completed a comprehensive study of MCQs and established 43 rules for writing questions. They used 46 authoritative sources representing educational measurement to develop definitive item-writing rules in three categories: general test writing, stem construction, and options. These authors found consensus from their sources on 33 of the 43 rules. Of the 43 rules, the authors were able to find research on only 23 rules. Of the criteria that have been identified, few have been empirically studied (Downing, 2005; Rodriguez, 2005). Thus, these authors concluded that, although there are many sources that provide direction on writing MCQs, there was little empirical basis for the development of test questions. One could say that rules that define test-writing criteria are considered guidelines.

Nurse academicians who sought to study MCQs found that neither test bank questions (Masters, Hulsmeyer, Pike, Leichty, Miller, & Verst, 2001) nor instructor-developed test questions (Tarrant, Knierim, Hayes, & Ware, 2006) yielded quality items. Masters et al. (2001) examined 2,913 textbook test bank questions and found 76.7% violated test-writing guidelines. In addition, 47.3% were written at the knowledge level, the lowest cognitive level, according to Bloom's (1956) taxonomy. Tarrant et al. (2006) examined 2,770 instructor-developed test questions and found that 46.2% contained at least one violation of accepted guidelines and 91% were written at a knowledge/comprehension level. A search of PubMed, CINAHL, and references from all reviewed articles revealed no research regarding the use of MCQs in staff development, and only the two identified articles in nursing academia.

The lack of studies on testing in nursing undergraduate or hospital-based education is surprising as the topic is addressed in education textbooks and has been studied in other disciplines. Test questions have been studied in undergraduate medical education (Case & Swanson, 2003; Downing, 2005; Palmer & Devitt, 2007), continuing medical education (Braddom, 1997), undergraduate pharmacy (Schultheis, 1998), and radiology (Collins, 2006). Regardless of discipline, item-writing flaws (IWFs) in test question construction have been noted. For example, Downing (2005) found that 35%–65% of test items in medical education were flawed.

Researchers have identified potential reasons for lack of quality questions. Vyas and Supe (2008) noted that limited time and education of faculty in preparing MCQs contribute to flaws in writing quality items. Tarrant et al. (2006) noted that few nurse educators have formal preparation in constructing MCQs. Farley (1989b) identified a trend in graduate nursing programs to prepare "clinical" experts as opposed to programs focused on educational expertise.

Rosemarie Nedeau-Cayo, MSN, RN-BC, is Staff Development Specialist, Bronson Methodist Hospital, Kalamazoo, Michigan.

Deborah Laughlin, MSN, RN-BC, is Education Services Instructor, Bronson Methodist Hospital, Kalamazoo, Michigan.

Linda Rus, MSN, RN-BC, is Education Services Manager, Bronson Methodist Hospital, Kalamazoo, Michigan.

John Hall, MSN, RN, is Education Services Instructor, Bronson Methodist Hospital, Kalamazoo, Michigan.

The authors have disclosed that they have no significant relationship with, or financial interest in, any commercial companies pertaining to this article.

ADDRESS FOR CORRESPONDENCE: Rosemarie Nedeau-Cayo, MSN, RN-BC, Bronson Methodist Hospital, 601 John Street, Box 44, Kalamazoo, MI 49007 (e-mail: cayor@bronsonhg.org).

DOI: 10.1097/NND.0b013e318286c2f1

Well-written test questions begin with identifying the objective of the lesson and focus on relevant content (Braddom, 1997). The test provides feedback to the student on the content learned. Educators inadvertently test on irrelevant information in an effort to construct discriminating questions, which results in unfair tests (McCoubrie, 2004). Collins (2006) noted that well-written test questions produce meaningful test scores and measurement of student achievement. Downing (2005) suggested that, because of flawed MCQs, as many as 10%–15% of students could fail a test they should have passed.

Bloom's (1956) taxonomy is a well-recognized framework in nursing education as a means for defining levels of educational objectives. Well-written questions should be congruent with the level of the objectives. As noted by Masters et al. (2001), knowledge, comprehension, application, and analysis can be tested with MCQs. Tarrant et al. (2006) simplified the taxonomy, creating two levels: K1 represented basic knowledge and comprehension, and K2 encompassed application and analysis.

There are no clear theories regarding effective test-question construction. Haladyna, Downing, and Rodriguez (2002) stated: "The scientific basis for writing test questions appears to be improving but very slowly. We still lack widely accepted, question-writing theories supported by research with resulting technologies for producing many questions that measure complex types of student learning that we desire" (p. 327). Direct correlation between objectives, content, and quality MCQs is, in essence, good educational design.

OBJECTIVE

The objective of this study was to examine the frequency of multiple-choice IWFs and the relationship between IWFs, presence of objectives, and cognitive level in organizationally developed test questions within a learning management system at a midsize acute care hospital. MCQs at the study institution had never been examined to ensure meeting standard criteria.

Definitions useful in this study are included in Table 1.

METHODS

A systematic/constructive replication study based on the work of Tarrant et al. (2006) was used. In a "systematic extension or constructive replication, the study is done under distinctly new conditions. The investigation team identifies a similar problem but formulates new methods to verify the first researcher's findings. The aim of this type of replication is to extend the finding of the original study and test the limits of generalizability of such findings" (Burns & Grove, 2005, p. 74). The sample at the study hospital was composed of 405 computer-based learning (CBL) modules/tests written by multidisciplinary content experts. Duplicate questions were removed resulting in 3,509 test questions used for the study.

The tool to define IWFs, created by the investigators, was derived primarily from the work of Tarrant et al. (2006) who identified 19 IWFs consistent with guidelines formulated by Haladyna et al. (2002). In addition, the cognitive level of each test question (K1 or K2) was included on the tool based on Tarrant et al.'s compression of Bloom's (1956) taxonomy. The presence of objectives and the distribution pattern of the correct responses were also collected. The proposal was ruled exempt upon submission to the hospital's institutional review board.

Interrater reliability was established through review of 20 test questions by the four investigators using the collection tool. Investigators reached consensus of at least 90% when identifying IWFs, cognitive level, and correlation of objectives with test questions. A pilot study of the tool was conducted on 200 MCQs. The four investigators reviewed 50 questions each. Four questions were randomly selected, and the investigators individually and collectively identified the IWFs and cognitive level. Results were discussed, and consensus was determined. Because of the frequency of true/false items, a parameter was added to the tool to separate them from MCQs. True/false items were not part of this study. The study moved into the full study phase when agreement was reached regarding the accuracy of findings. All MCQs were reviewed, and for each 200 questions, four were randomly reviewed collectively by the investigators to maintain interrater reliability.

Data were summarized using descriptive statistics. A chi-square test was conducted to determine the association between IWFs and cognitive level of question and the cognitive level of questions and presence of objectives related to test items. Chi-square test was used to assess whether the correct answers were evenly distributed. Fisher's exact test was performed when the number of events was fewer than

TABLE 1 Study Definitions

Term	Definition
Item	A statement of a problem followed by a list of possible solutions or answers.
Stem	The statement of the problem, usually consisting of 1–2 sentences, that poses the question to the learner.
Options	A list of alternative answers or solutions.
Distractor	Options that are intended to distract the test taker from the correct response. The best distractors are often common mistakes of learners.
Item-writing flaws	Violations of commonly accepted guidelines for writing multiple-choice questions (see Table 2 for details of flaws examined in this study).

TABLE 2 Recommended Guidelines for Writing High-Quality Multiple-Choice Questions (Tarrant et al., 2006)

Name	Description
All options grammatically consistent with stem.	Parallel in style and form; nongrammatically correct options provide cues to the student who easily eliminates distracters that do not flow grammatically with the stem.
Each MCQ should have a clear and focused question.	Teachers should avoid using MCQs with unfocused stems, which do not ask a clear question or state a clear problem in the sentence completion format.
Each MCQ should have the problem in the stem of the question, not in the option.	The options should not be a series of true/false statements.
The basic format for MCQs is the single best answer.	Ensure that questions have one, and only one, best answer.
Avoid gratuitous or unnecessary information in the stem or the options.	If a vignette is provided with the MCQ, it should be required to answer the question.
Avoid complex or K-type MCQs.	K-type MCQs have a range of correct responses and then ask students to select from a number of possible combinations of these responses. Students can often guess the answer by eliminating one incorrect response and all options containing this response or by selecting the responses that appear most frequently in all of the options.
Questions and all options should be written in clear, unambiguous language.	Poorly worded or ambiguous questions can confuse even knowledgeable students and cause them to answer incorrectly.
Make all distracters plausible.	Students who do not know the material increase their chances of guessing the correct option by eliminating implausible distracters.
Avoid repeating words in the stem and the correct option.	Similar wording allows students to identify the correct option without knowing the material.
Avoid providing logical cues in the stem and the correct option that help the student to identify the correct option without knowing the material.	For example, asking students to select the most appropriate pharmaceutical intervention for a problem and only having one or two options, which are actually pharmaceutical interventions.
Avoid convergence cues in options where there are different combinations of multiple components to the answer.	Question writers tend to use the correct answers more frequently across all options, and students will identify as correct the answer in which all components appear most frequently.
All options should be similar in length and amount of detail.	If one option is longer, includes more detailed information, or contains more complex language, students can usually correctly assume that this is the correct answer.
Arrange MCQ options in alphabetical, chronological, or numerical order	No definition.
Options should be worded to avoid the use of absolute terms (e.g., never, always, only, all).	Students are taught that there are often no absolute truths in most health science subjects, and they can eliminate these distracters.
Options should be worded to avoid the use of vague terms (e.g., frequently, occasionally, rarely, usually, commonly).	Lacks precision, and there is seldom agreement on the actual meaning of ''often'' or ''frequently.''
Avoid the use of negatives (e.g., not, except, incorrect).	They poorly assess actual knowledge. If teachers wish to assess contraindications, the questions should be worded clearly to indicate that this is what is being assessed.

Continued

TABLE 2 Continued

Name	Description
Avoid the use of "all of the above" as the last option.	Students can easily identify if this is the correct answer by simply knowing that at least two of the options are correct. Similarly, they can eliminate it by knowing if only one of the options is incorrect.
Avoid the use of "none of the above" as the last option.	It only measures students' ability to detect incorrect answers. If "none of the above" is the correct option, the teacher must be certain that there are no exceptions to any of the options that the student may detect.
Avoid fill-in-the-blank format whereby a word is omitted in the middle of a sentence and the student must guess the correct word.	All options should be placed at the end of the stem.

Abbreviation: MCQ = multiple-choice question.

five. A 5% level of significance was used to evaluate statistical significance. All data analysis was performed using SAS 9.1 (SAS Institute, Inc.).

FINDINGS

Investigators evaluated 3,509 questions and eliminated 1,018 true/false questions resulting in 2,491 MCQs evaluated in this study. Of the 2,491 multiple-choice items, 386 (15.5%) items contained no flaws, 1,243 (49.9%) items contained one flaw, and 862 (34.6%) questions had more than one flaw. The most frequent IWFs were "all of the above" (n = 713), "more than one or no correct answer" (n = 387), "implausible distractors" (n = 380), "repeating word" (n = 314), "dissimilar length options" (n = 268), and "none of above" (n = 205). The most infrequent IWFs were "convergence cues" (n = 20), "complex or K type" (n = 21), "vague terms" (n = 23), and "unfocused stem" (n = 26; see Figure 1).

When objectives were present, there was a significant association between "objectives present" and "question refers to objective" ($p \leq .0001$). Ninety-seven percent of the questions referred to objectives; however, only 16% of the questions had associated objectives. There was no significant association between "objectives present" and IWFs (p = .3270) or between "question refers to objective" and IWFs (p = .6570). There was no association found between cognitive levels and the presence of objectives (p = .087). Although not statistically significant, MCQs were more likely to relate to the objectives at the K1 level.

The 2,491 MCQs were evaluated for a relationship between cognitive level and IWFs; more than 90% of the items were written at the K1 recall level (n = 2,332, 93.69%). There was a significant association between levels of cognition and IWFs. Most of the items written at the K1 level had IWFs (n = 1,986, 94.4%, p = .0008) as compared with those written at the K2 level (n = 118, 5.6%, p = .0008).

A significant association between cognitive levels and eight IWFs was observed. At the lower cognitive level (K1), six flaws, "grammatical cues" (p = .0129), "repeating word" (p = .0149), "all of the above" (p < .0001), "more than one or no correct answer" (p < .0001), "implausible distracters" (p = .0003), and "negatively worded stem" (p = .0296), were more likely to occur. At the higher cognitive level (K2), two flaws, "options not in sequence" (p < .0001) and "fill in the blank" (p = .0268), were more likely to occur.

DISCUSSION

In this study, 85% of the MCQs had at least one flaw, of which most were contained in the options. Half of the questions contained only one flaw, and the remaining 35% had at least two or more flaws. This is a higher rate of flaws than the 46% and 76.7% found in nursing academia by Tarrant et al. (2006) and Master et al. (2001). Effective distractors are hard to develop when the correct answer is obvious to the test question creator. Education about writing quality MCQs, such as not using "all of the above," could eliminate 29% of the errors. Other frequent flaws that are easy to eliminate

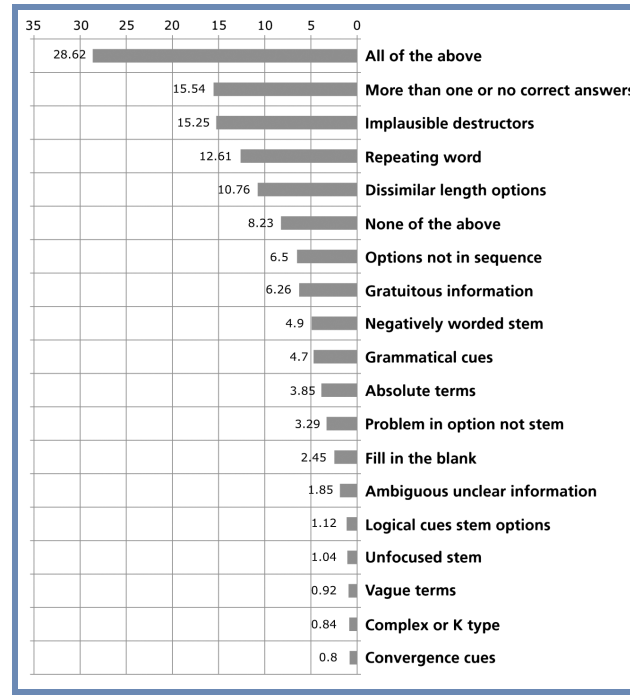

FIGURE 1 Frequency of 19 item-writing flaws.

include "more than one or no correct answer," "implausible distractors," "repeating word," and "none of the above." The fifth most common error, "dissimilar length of options," occurred because the correct answer was the longest.

There are many reasons for the high incidence of IWFs. The study hospital's tests are written by "experts" of their respective discipline such as nursing, environmental services, dietary, or environmental safety and not necessarily by nursing educators. There is no evidence, however, that nursing educators are better prepared to write test questions than other disciplines. Raising awareness and education on how to write MCQs would help eliminate most of flawed test items. Writing quality test items takes time, and educators and content experts have many competing obligations resulting in little time dedicated to test-question writing. Some test writers consider the test itself as a learning opportunity rather than an evaluation and make the test questions "teaching" questions.

When objectives were present, there was a significant association between "objectives present" and the "question refers to an objective." However, 84% of the questions did not have associated objectives. Good instructional design measures mastery of subject matter and is related to learning outcomes. Therefore, test questions should be based on the objectives (Rasmussen, Speck, & Twigg, 1998). Objectives posted online with the test would enable the test taker to ascertain the nature of content to be tested and would facilitate the creation of questions focused on objectives. Many of the departmental experts are not trained in educational design and thus may not understand the necessity of objectives or the importance of connecting objectives to test questions. At the study hospital, many e-learning modules are stand-alone tests with no content. This provides a convenient method to document compliance. These stand-alone tests do not include objectives either, which accounts for the low overall percent of modules with objectives.

The purpose of MCQs is to measure achievement of the objectives. The study hospital requires employees to obtain 100% on all CBL tests. The validity of obtaining 100% on a test about a particular subject could be questioned if the MCQs are not based on objectives. This may contribute to the frustration of test takers who comment that the tests are a waste of time with meaningless test questions. Rather than engaging in learning, employees develop "work around" methods to determine correct answers. Some questions have discernible answers because of multiple flaws, and test takers do not have to know the information to answer correctly. Test savvy employees with knowledge of test taking techniques could readily determine the correct answer without knowing the content.

There were more flaws in items written at the K1 level than those written at the K2 level; however, most items were written at the K1 level. This finding is consistent with Tarrant et al. (2006) who found that most items were written at the K1 level; nearly half of those items had flaws. There are no clear data regarding these phenomena. It is possible that more thought is given to items written at a higher level or that those who wrote higher level questions had more instruction regarding educational design. It is interesting to note that objectives were lacking regardless of the cognitive level of the test question.

Four of the most common flaws ("all of the above," "use more than one or no correct answer," "implausible distractors," and "repeating word") were significantly associated with the lower cognitive level. "Grammatical cues" and "negatively worded stem," although not the most common errors, were also associated with lower level test questions. The inexperienced test writer may be more likely to write lower level and/or flawed questions. Good test questions take time to write, but flawed test questions are quick and easy. For example, options such as "all of the above" allow the test creator to include a lot of information, or "implausible distractors" allow the test writer to develop a distractor when they cannot think of enough quality options. There were only two flaws significantly associated with higher level test questions: "options not in sequence" and "fill in the blank." Test writers may be unfamiliar with these particular MCQ guidelines and inadvertently use these flaws as they write test questions.

The number of distractors could affect the number of flaws. The number of distractors for questions in this study ranged from three to nine. Although there are no clear guidelines, the literature supports limiting to three options as more may simply create improbable distractors or increase the number of item flaws (Rodriguez, 2005). Six of the seven most frequent flaws, such as "all of the above" or "implausible distractors," occur in the options. Limiting options could eliminate some of the errors. Anecdotally, although not an actual identified flaw, the test writer may want to give consideration to an even distribution of correct responses. It has been suggested (Collins, 2006) that certain answer placement (e.g., a or b or c) can be overused, leading the test taker to the correct answer or selecting the correct answer by chance.

Limitations

Despite efforts to ensure interrelator reliability, the potential exists that there was inconsistency between investigators during item analysis. In addition, investigators could miss errors during the evaluation of each MCQ because of interpretation of the stem or distractors, lack of information related to the content of question, analysis fatigue, or other disruptions. Although Haladyna et al. (2002) have extensively researched MCQs, many of the criteria have not been empirically tested. Some errors may be more important to the reliability of the test items than others. Finally, true/false questions, which were frequently used, were not analyzed in this study.

Implications

The findings from this study closely correlated to results in the literature of other healthcare disciplines, which found that most MCQs contain flaws and were written at low cognitive levels. To help test item writers avoid common test construction errors, standards for writing effective MCQs are needed. Criteria should be developed for writing well-constructed test items based on objectives.

A task group was assigned to review all CBLs in the study hospital's e-learning system. Standards were established for existing CBLs as well as the development of new CBL modules to include objectives, content based on objectives, and quality MCQs. Instructional guidelines were developed and posted on the study hospital's intranet, which incorporated information on how to write effective MCQs and learning objectives. Anecdotally, the task group also determined if the readability level was congruent with the intended audience. These tools provided valuable resources for content experts writing CBLs. In addition, an inservice class was presented on how to write effective MCQs, and mentors were assigned to CBL authors.

The 405 CBL modules in the study hospitals learning management system were placed on a 3-year review cycle. A task force of educators was designated to review one third of the CBLs annually for good educational design with a focus on the quality of the MCQs. The CBLs were edited for objectives and content, and all test questions were reviewed for IWFs. Task force members worked with content experts to write objectives and develop quality test questions. The level of test writing competence has been enhanced in the organization through this undertaking.

CONCLUSION

The objective of this study was to examine the frequency of multiple-choice IWFs and the relationship between IWFs, presence of objectives, and cognitive level in organizationally developed test questions. Research has identified that MCQs constructed following established guidelines more accurately validate an individual's learning. On the basis of the results of this study, processes were developed for writing quality test items. These strategies encompassed the inclusion of objectives for each online module and measurement of MCQs against established criteria. Content experts were mentored by educators and provided with resources for constructing well-written modules and test questions. The outcome was quality MCQs that evaluate learning in a system, which is consistent with good educational design.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Karen Morin, DSN, RN, ANEF, FAAN, for reviewing and editing this manuscript; Yusuf K. Bilgic, MS, for his statistical expertise; and David J. Burnham for the development of the data collection tool.

References

Bloom, B. S. (1956). *Taxonomy of educational objectives. Handbook1: The cognitive domain*. London, UK: Longman.

Braddom, C. L. (1997). A brief guide to writing better test questions. *American Journal of Physical Medicine & Rehabilitation*, *76*(6), 514–516.

Burns, N., & Grove, S. K. (Eds.). (2005). *The practice of nursing research: Conduct, critique, and utilization* (5th ed.). St. Louis, MO: Elsevier Saunders.

Case, S. M., & Swanson, D. B. (2003). *Constructing written test questions for the basic and clinical sciences (3rd ed.)* [brochure]. Philadelphia, PA: National Board of Medical Examiners. Retrieved from http://www.nbme.org/publications/item-writing-manual.html

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics*, *26*(2), 543–551.

Downing, S. M. (2005). 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. *Advances in Health Sciences Education*, *10*(2), 133–143.

Farley, J. K. (1989a). The multiple-choice test: Developing the test blueprint. *Nurse Educator*, *14*(5), 3–5.

Farley, J. K. (1989b). The multiple-choice test: Writing the questions. *Nurse Educator*, *14*(6), 10–12.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37–50. doi:10.1207/s15324818ame0201_3

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.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education*, *40*(1), 25–31.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? *BMC Medical Education*, *28*(7), 49–56.

Rasmussen, L., Speck, D. J., & Twigg, P. (1998). Developing and using classroom tests. In D. M. Billings & J. A. Halstead (Eds.), *Teaching in nursing: A guide for faculty* (pp. 385–405). Philadelphia, PA: W.B. Saunders Company.

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.

Schultheis, N. M. (1998). Writing cognitive educational objectives and multiple-choice test questions. *American Journal of Health System Pharmacists*, *55*(22), 2397–2401.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nursing Education in Practice*, *6*(6), 354–363.

Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *The National Medical Journal of India*, *21*(3), 130–133.