American Journal of Pharmaceutical Education **2013; 77 (4) Article 71.**

RESEARCH

Effects of Question Formats on Student and Item Performance

David J. Caldwell, PharmD, and Adam N. Pate, PharmD

University of Louisiana at Monroe College of Pharmacy, Monroe

Submitted October 8, 2012; accepted December 9, 2012; published May 13, 2013.

Objective. To determine the effect of 3 variations in test item format on item statistics and student performance.

Methods. Fifteen pairs of directly comparable test questions were written to adhere to (standard scale) or deviate from (nonstandard scale) 3 specific item-writing guidelines. Differences in item difficulty and discrimination were measured between the 2 scales as a whole and for each guideline individually. Student performance was also compared between the 2 scales.

Results. The nonstandard scale was 12.7 points more difficult than the standard scale ($p=0.03$). The guideline to avoid "none of the above" was the only 1 of the 3 guidelines to demonstrate significance. Students scored 53.6% and 41.3% ($p<0.001$) of total points on the standard and nonstandard scales, respectively.

Conclusions. Nonstandard test items were more difficult for students to answer correctly than the standard test items, provided no enhanced ability to discriminate between higher- and lower-performing students, and resulted in poorer student performance. Item-writing guidelines should be considered during test construction.

Keywords: multiple-choice questions, examination, test item-writing guidelines, test construction, assessment

INTRODUCTION

In higher education, few faculty members receive formal training in how to construct objective test items, yet the mainstay of summative assessment is the multiple-choice examination.[1-5] This format allows the statistical analysis necessary to determine essential psychometric characteristics such as reliability (a quality that supports the format's routine use), as well as item difficulty and discrimination. Many articles have been published on the accepted characteristics of good vs bad test items, and guidelines exist that provide the framework for writing examination questions. Deviations from examination item-writing guidelines may result in undesirable changes to item statistics, eg, in discrimination or in the percentage of students answering correctly.[6] Many potential factors, such as faculty members' lack of familiarity with these guidelines and reluctance to alter personal examination writing habits, and a relative lack of experimental data on examination item performance, may contribute to a low level of guideline acceptance and application among educators. If item-writing guidelines are valid, and nonstandard items (defined as those that violate any of the item-writing guidelines used in this study) affect test scores, conventional item construction techniques in higher education may be contributing to an underestimation of students' true knowledge, skills, and abilities, as well as negatively impacting student progression.

Instruction in this nearly universal assessment technique is often omitted from the training of higher-education faculty members. In academic pharmacy, new educators have often received postgraduate training related to education through residency programs. Sixty-eight percent of the residency programs accredited by the American Society of Health-System Pharmacists (ASHP) reported having some form of teaching opportunity for their residents.[7] However, training in test construction is not included in the ASHP standards for either postgraduate year 1 or 2 programs.[8,9] McNatty and colleagues examined the teaching experiences included in approximately 800 surveyed residency programs, but instruction on testing, and specifically test-item construction, were not reported.[10] Although some programs reported "formal training in teaching and learning," training in classroom assessment was not mentioned. Experienced faculty members are even less likely to have received formal instruction in test construction, and may have full teaching loads that leave little time for the development of test-writing skills.

Compilations of expert opinions on good vs bad item characteristics exist in educational measurement textbooks. Examination item construction variables such as

Corresponding Author: David Caldwell, PharmD, 1800 Bienville Drive, Monroe, LA 71201. Tel: 318-342-1689. Fax: 318-372-5290. E-mail: dcaldwell@ulm.edu

question format and number of answer options per question may affect item difficulty and discrimination.[11] Variable findings describing these effects have been published in the fields of medical and nursing education.[12-14]

In a previous study, the authors retrospectively examined the effects of guideline use on examination item statistics for all examination items administered in a single course. The objective was to determine if variation from item guidelines resulted in differences in student performance or item quality. Examination items were classified into 2 scales: standard (adhered to all 31 guidelines of Haladyna, Downing, and Rodriguez) and nonstandard (broke 1 or more of the 31 guidelines). There was a 7.4% difference in difficulty between the 2 scales, with the nonstandard scale being more difficult ($p=0.01$), and no significant difference in item discrimination measured by average scale point biserial correlation ($p=0.06$).[15] These and the aforementioned findings supported the need for a more systematic evaluation of the effects of guideline use on item quality.

METHODS

This study was approved by the university's institutional review board. Three item-writing guidelines that received mixed endorsement in the educational measurement literature and published research examined in the paper by Haladyna, Downing, and Rodriguez were selected to undergo analysis in this study because of the likelihood that they would increase item difficulty and negatively impact student performance:

(1) Word the stem positively; avoid negatives such as *not* or *except.* If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface (63% for, 18% against, 19% uncited).

(2) Develop as many effective choices as you can, but research suggests 3 is an adequate number (70% for, 4% against, 26% uncited).

(3) None-of-the-above (NOTA) should be used carefully (44% for, 48% against, 7% uncited).[5]

Two sets of 15 corresponding questions were developed to test the effects of these rules on student and item performance. Five faculty members who taught first-year students each provided 3 pairs of questions to be included in the analysis, 1 pair for each guideline. One question in each pair adhered to the corresponding item-writing guideline and was defined as *standard scale*. The other question in each pair differed only by breaking the corresponding guideline in a specific way and was defined as *nonstandard scale* (Table 1). Five faculty members donating 3 questions to each scale resulted in 15 pairs of directly comparable items – 5 questions per guideline.

Table 1. Example Correlating Items From the Nonstandard and Standard Scales[a]

Guideline	Nonstandard	Standard
Use positives, no negatives	All of the following organisms have significant toxin production EXCEPT a) *Clostridium perfringens* b) *Escherichia coli* **c) *Neisseria meningitidis*** d) *Streptococcus pyogenes*	Which of the following organisms has significant toxin production? a) *Bacillus cereus* b) *Moraxella catarrhalis* c) *Neisseria meningitidis* **d) *Streptococcus pyogenes***
Write as many plausible distractors as you can	What test is used to determine if a patient is hypochromic regardless of erythrocyte size? a) Blood smear b) Hematocrit c) Mean corpuscular hemoglobin **d) Mean corpuscular hemoglobin concentration** e) Red cell distribution width	What test is used to determine if a patient is hypochromic regardless of erythrocyte size? a) Hematocrit b) Mean corpuscular hemoglobin **c) Mean corpuscular hemoglobin concentration**
Use carefully *None of the above*	Which of the following is a B-cell specific neoplasm? a) Acute lymphocytic leukemia b) Chronic myeloid leukemia **c) Hodgkin lymphoma** d) Non-hodgkin lymphoma e) None of the above	Which of the following is a B-cell specific neoplasm? a) Acute lymphocytic leukemia b) Chronic myeloid leukemia **c) Hodgkin lymphoma** d) Non-hodgkin lymphoma

[a] Correct answers are in bold.

These items were administered to students on the first mile marker examination,which is a formative assessment administered at the end of the spring semester of the first year. The 100-item test covers material determined by instructors to be overarching in the first year. General content covered includes calculations, drug information retrieval, ethics and law, immunology, medicinal chemistry, microbiology, parenterals, pathophysiology, and pharmaceutics. The examination is comprehensive, questions are written by the faculty members who taught the topics addressed, and the college's curriculum and assessment committees each vet the questions before the items are included on the examination. The 15 pairs of study items described above were split into standard and nonstandard scales and added to the end of the examination, resulting in 2 versions of the examination, each with 115 questions. The study items were scored but not factored into students' examination grade. Because all examination items that were part of the study were placed at the end of the examination, the effects of test fatigue and access to testing cues were equalized between the 2 groups.

Randomization occurred by allowing students to self-select seats in the testing auditorium. A standard or nonstandard form was alternatingly distributed at each seat. Examination security was maintained by the presence of 4 examination proctors, as well as by varying the item numbering on the 2 versions of the examination and the corresponding answer sheet. All examinations and answer sheets were accounted for at the end of the testing period. Characteristics of students in each group are presented in Table 2.

One-sided independent *t* tests were used to analyze the differences in item difficulty and item performance between the 2 scales, as well as student performance on each scale. Statistix 9 (Analytical Software, Tallahassee, FL) was used for these analyses. Item difficulty is defined as the percentage of students answering the item correctly. Item discrimination is measured by the point biserial correlation, which is defined as "the correlation between right/wrong scores that students receive on a given item and the total scores that the students received when summing up their scores across the remaining items."[16] Point biserial correlations were calculated by the LXR*TEST grading software (Applied Measurement Professionals, Inc., Georgetown, SC) and reported based on the 15 study items in each scale.

Table 2. Descriptive Characteristics of the Students[a]

	Standard Scale	Nonstandard Scale
Gender, No. (%)		
Female	36 (65.4)	32 (59.3)
Male	19 (34.5)	22 (40.7)
Preprofessional degree, No. (%)	18 (32.7)	18 (33.3)
PCAT, Mean (SD)	56 (16)	56 (18)
Prepharmacy GPA, Mean (SD)	3.4 (0.4)	3.3 (0.4)

Abbreviations: PCAT = Pharmacy College Admission Test; GPA = grade point average.

[a] There was no significant difference between the groups in any available characteristic.

RESULTS

One hundred nine students took the mile marker examination. Fifty-five students completed the version of the examination with standard form items and 54 students completed the version with nonstandard form items. Levels of Bloom's taxonomy represented by each scale were knowledge (60%), comprehension (20%), and application (20%). A summary of the difficulty and discrimination analyses can be found in Table 3. The difference in average percent correct between the standard and nonstandard scales was 12.3% (53.6 and 41.3, respectively, $p<0.001$). Separate analyses of the individual guidelines failed to identify significant results except for: avoid "none of the above." In the scale comparison for this guideline, an average of 39.2 students (71%) correctly answered the standard scale item compared to an average of 25.2 (47%) students who correctly answered the nonstandard scale item ($p=0.044$), and there was no difference in average item discrimination ($p=0.22$).

DISCUSSION

This analysis provides striking initial evidence that common deviations from item-writing guidelines can result in poorer student performance on questions with no increased benefit of differentiating higher- from lower-performing students. Guideline deviations in this study appeared to increase item difficulty without the benefit of increased discrimination, which makes a case for consideration of item-writing guidelines during test construction. Additionally, student performance on standard vs nonstandard questions was significantly better—in this study, the difference was more than a letter grade based on a 10-point scale.

Of the individual guidelines analyzed, only use of "none of the above" resulted in significant differences. The case against "none of the above" can be made based on the cognitive processes it requires of a student. If "none of the above" is the correct answer to a question, the behavior that is being tested is at least in part the student's ability to recognize incorrect answers; knowledge of the correct answer is not an absolute requirement. If "none of the above" is not the correct answer, it may be incorrectly

Table 3. Summary of Comparative Item Statistics per Guideline

	Average Scale Difficulty, %[a]			Average Scale Point Biserial Correlation[b]		
Guideline	**Standard**	**Nonstandard**	***P***	**Standard**	**Nonstandard**	***P***
Stem negation	37.6	31.8	0.24	0.312	0.312	0.5
Plausible distractors	52.2	44.6	0.21	0.333	0.287	0.16
None of the above	71.4	46.8	0.044	0.274	0.335	0.25
Whole scale	53.7	41.0	0.04	0.306	0.311	0.45

[a] Percentage refers to percentage of students answering the question correctly.
[b] Reported point biserial correlations calculated by the 15 study items in each scale.

chosen by students who are able to identify a theoretical answer that is more correct than any of the choices provided. The effects of using a "none of the above" test item can therefore be quite complex.

The other guidelines under investigation included stem negation and optimal number of distractors. Stem negation items included questions in which stems contained words such as "not" or "except." The standard scale versions of these questions were formed by simply rewriting the negative item positively. Items assessing the differences in numbers of plausible distractors were written to contain either 3 options (standard scale) or 5 options (nonstandard scale). In this study, neither guideline analysis resulted in significant results. However, 5 comparisons per guideline may be too few to determine true differences if these guidelines result in more subtle effects than use of the "none of the above" item, for example. Phipps and colleagues described the effects of number of foils on pharmacy examinations and found that 5 options were more difficult but also more discriminating than 4.[12] However, because of the difficulty of writing additional functional distractors, test authors' time may be better spent by limiting item choices to 3.[5]

Regarding limitations of this study, the method of randomization was a potential confounder. Because students were allowed to self-select their seats, we could not determine with complete certainty that the 2 study groups (students completing the standard version and nonstandard version of the examination) were truly similar. However, no significant differences in student-specific characteristics were found between the 2 groups (Table 2). The small number of investigational items was a significant limitation for this study. Although the overall scale comparison found significant differences, the individual guideline analyses were potentially negatively affected by the small number of comparisons that could be made for each subgroup. Similarly, the small sample size of items and multidimensional nature of the mile marker examination prevented the calculation and comparison of scale reliabilities. It would also be beneficial to analyze the effects of guideline deviations at each level of Bloom's taxonomy; the distribution of study items in this project, however, makes such comparisons impossible. Considering these measured effects, and because few faculty members in higher education have received formal training in proper test item construction, it may be appropriate for multiple-choice item-writing to become a focus of teaching certificate programs, professional development opportunities, and faculty development initiatives in colleges and schools of pharmacy.

CONCLUSION

Deviations from multiple-choice item-writing guidelines resulted in increased item difficulty without a corresponding increase in item discrimination. Additionally, these deviations resulted in significantly poorer student performance that amounted on average to a letter grade. Although this study illustrates some significant effects of item construction choices, the authors suggest that while consideration should be given to the published item-writing resources, guidelines should remain suggestions that may be accepted or rejected, rather than rules to be enforced.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Veronica Lewis for her valuable contributions and guidance during the entire study process, as well as Drs. Ronda Akins, Benny Blaylock, Chris Gissendanner, Ron Hill, and Seetharama Jois for their willingness to contribute their efforts as item-writers for the project.

REFERENCES

1. Berk RA. A consumer's guide to multiple-choice item formats that measure complex cognitive outcomes. pearson assessment and information. http://www.pearsonassessments.com/hai/images/NES_Publications/1996_12Berk_368_1.pdf. Accessed November 26, 2012.
2. Burton SJ, Sudweeks RR, Merrill PF, Wood B. How to prepare better multiple-choice test items: guidelines for university faculty. brigham young testing services and the department of instructional science. http://testing.byu.edu/info/handbooks/betteritems.pdf. Accessed November 26, 2012.
3. Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences.* 3rd ed (revised). Philadelphia, PA:

National Board of Medical Examiners; 2002. http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf. Accessed November 26, 2012.
4. Professional Examination Services, Inc. A guide to preparing multiple-choice items. 2005. http://www.bpsweb.org/pdfs/itemwritingguide.pdf. Accessed November 26, 2012.
5. 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.
6. Rodriguez MC. The art & science of item-writing: a meta-analysis of multiple-choice item format effects. http://edmeasurement.net/aera/papers/artandscience.pdf. Accessed November 26, 2012.
7. Aistrope DS, Attridge RT, Bickley AR, et al. Strategies for developing pharmacy residents as educators. ACCP Commentary. *Pharmacotherapy.* 2011;31(5):65e-70e.
8. American Society of Health-System Pharmacists. Required and elective educational outcomes and goals, objectives, and instructional objectives for postgraduate year one (PGY1) pharmacy residency programs, 2nd edition. http://www.ashp.org/DocLibrary/Accreditation/PGY1-Goals-Objectives.aspx. Accessed November 26, 2012.
9. American Society of Health-System Pharmacists. Program outcomes, educational goals, and educational objectives for postgraduate year two (PGY2) residencies in an advanced area of pharmacy practice. http://www.ashp.org/DocLibrary/Accreditation/RTP_ObjPGY2AdvanceGoalObj.doc. Accessed November 26, 2012.
10. McNatty D, Cox CD, Seifert CE. Assessment of teaching experiences completed during accredited residency programs. *Am J Pharm Educ.* 2007;71(5):Article 88.
11. Phipps SD, Brackbill ML. Relationship between assessment item format and item performance characteristics. *Am J Pharm Educ.* 2009;73(8):Article 146.
12. 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.
13. Ware J, Torstein V. Quality assurance of item-writing: during the introduction of multiple choice questions in medicine for high stakes examinations. *Med Teach.* 2009;31(3):238-243.
14. 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.
15. Pate AN, Caldwell DJ. Effects of multiple choice item-writing guideline utilization on item and student performance. abstract. *Am J Pharm Educ.* 2012;76(5):Article 99.
16. Varma S. Preliminary item statistics using point-biserial correlation and p-values. Educational Data Systems, Inc. 2012. http://www.eddata.com/resources/publications/EDS_Point_Biserial.pdf. Accessed November 26, 2012.