American Journal of Pharmaceutical Education **2009; 73 (8) Article 146.**

RESEARCH ARTICLES

Relationship Between Assessment Item Format and Item Performance Characteristics

Stephen D. Phipps, PharmD, PhD,* and Marcia L. Brackbill, PharmD

Bernard J. Dunn School of Pharmacy, Shenandoah University

Submitted February 17, 2009; accepted June 12, 2009; published December 17, 2009.

Objective. To evaluate the relationship between assessment item formats (case-based versus noncase-based) and item performance characteristics.

Methods. Assessment items (1,575) were collected from examinations administered in several therapeutics courses over 4 academic years. Items were categorized as either ''case-based'' or ''noncase-based'' and item performance characteristics (discrimination index and level of difficulty) were evaluated.

Results. Noncase-based items represented approximately three-fourths of all items that were evaluated, and demonstrated a higher discrimination index than case-based items. Case-based items were generally lengthier and included more detailed information than noncase-based items; however, they were not more difficult and exhibited a lower discrimination index. Secondary analyses revealed that 5-foil multiple-choice items are more difficult and have a higher discrimination index compared to 4-foil items.

Conclusion. The format used for an examination/test item (case-based or noncase-based) has an impact on item performance characteristics.

Keywords: item analysis, assessment, discrimination index, case-based, psychometrics

INTRODUCTION

Construction and selection of appropriate assessment items is an essential task for instructors. The critical issue throughout that process is how best to formulate items to optimally assess comprehension. Assessment items should correspond to the content (lecture material and readings) and the learning objectives and suitably match the instructional method(s) that are used to convey that material. For example, if course content is primarily delivered by working through calculation problems, the assessments should be comprised largely of calculation-type items.

There are numerous types of item formats used for student assessment, including traditional multiple-choice, K-type multiple-choice, true-false, matching, short answer, and essay items.[1] In addition to format, items vary in the level of abstraction based on the content and query of the item, ranging from knowledge (least abstract) to evaluation (most abstract).[2] The 6 levels of abstraction in Bloom's taxonomy correspond to various competencies that can be evaluated in an assessment: knowledge, comprehension, application, analysis, synthesis, and evaluation.[2] The level of abstraction that an item possesses is thus influenced by both the item format and its content.

Many educational disciplines often use ''case-based'' items on student assessments. In health care-related disciplines, case-based items commonly incorporate patient-specific information that is critically important in determining appropriate and accurate therapeutic decisions such as drug and dosage selection. As compared to traditional noncase-based assessment items, case-based items are theoretically wider in scope, require assimilation of more content, and are taxonomically more abstract (ie, categorized at the higher cognitive levels of synthesis and/or evaluation) than lower-level categories such as comprehension and simple recognition. Thus, it would seem intuitive and advantageous to incorporate case-based items on assessments in such courses as therapeutics.

In addition to their use for student assessment, case-based items represent an item format prevalent in licensing examinations for several health professions, including medicine (United States Medical Licensing Examination or USMLE),[3] dentistry (National Board Dental Examination or NBDE),[4] and pharmacy (North American Pharmacist Licensure Examination or NAPLEX).[5] In the case of

Corresponding Author: Stephen D. Phipps, PharmD, PhD. Department of Pharmaceutical Sciences, Lipscomb University College of Pharmacy, One University Park Drive, Nashville, TN 37204. Tel: 615-966-7123. E-mail: steve.phipps@lipscomb.edu
*Affiliation at time of study.

the NAPLEX, case-based items constitute the primary format of the examination:

> A majority of the questions on the NAPLEX are asked in a scenario-based format (ie, patient profiles with accompanying test questions). To properly analyze and answer the questions presented, you must refer to the information provided in the patient profile. Interspersed among these profile-based questions are "stand-alone questions," whose answers are drawn solely from the information provided in the question.[5]

After an assessment, an important and valuable endeavor for instructors is evaluating items to determine which ones were "good" and which ones were "poor." Two common psychometric parameters that can be evaluated for each item are the discrimination index (item discrimination) and the level of difficulty (item difficulty). The level of difficulty is simply the percentage of examinees who correctly answered an item. Unanimous agreement as to what constitutes an ideal level of difficulty does not exist; however, a desirable range for this psychometric parameter is 25%-30% at the lower end, to 75%-80% at the higher end.[6,7] Items with a difficulty level of less that 25% are typically regarded as very difficult, whereas items with a difficulty level of greater than 75% are considered moderately easy to easy.[7]

The discrimination index is a statistical index of item quality and reflects the degree to which the item was able to differentiate between examinees who scored well and those who scored poorly on an assessment.[6,8] The discrimination index is a calculated value ranging from -1.0 to 1.0, and indicates the extent to which the item correlates with overall examinee performance on the examination, eg, a high (positive) discrimination index reveals that the item was correctly answered by those examinees who performed well on the overall assessment. Typically, items with discrimination index values of 0.30 and above are regarded as good items.[9] The higher the discrimination index of an item, the greater the probability that selection of the correct response was due to content knowledge rather than chance.

Item difficulty and item discrimination are the 2 psychometric parameters that "are used to qualify and determine inclusion of an item in the NAPLEX test bank or pool"; however, the National Association of Boards of Pharmacy (NABP) does not divulge what the acceptable ranges and values are for these item parameters in making those determinations.[10] As a point of reference, the criteria for what are deemed effective items on the NBDE are a level of difficulty between 40% to 80% and discrimination index of ≥ 0.15 (on Part I) and ≥ 0.08 (on Part II).[11]

The objective of the current study was to evaluate the relationship between assessment item formats (case-based versus noncase-based) by analyzing and comparing their respective performance characteristics of discrimination index and difficulty level. Although the literature is replete with articles on case-based (or problem-based) instruction and learning, the authors are unaware of any previous research that has evaluated and compared case-based assessment items with noncase-based (or stand-alone) items. Our hypothesis was that case-based items and noncase-based items will exhibit dissimilar performance characteristics; specifically, that case-based items would demonstrate both a higher level of difficulty and a greater discrimination index. Additionally, there are varied opinions and recommendations regarding the utilization of different item structures (eg, standard multiple-choice and K-type multiple-choice), as well as on the number of response options (foils) that items should include. Therefore, the secondary objectives of the study were to evaluate the performance characteristics of (1) K-type multiple-choice versus standard multiple-choice items, and, (2) 4-foil multiple-choice items versus 5-foil multiple-choice items.

METHODS

Performance characteristics of multiple-choice items on therapeutics examinations administered in the second and third years of the doctor of pharmacy (PharmD) curriculum during 4 academic years (2004-2005 through 2007-2008) were collected. All examinations were either midterm or final examinations and were 2 hours in length; approximately one-fourth of the examinations were cumulative in nature. The course modules represented in this study included cardiovascular, infectious diseases, hematology-oncology, neurosensory, and psychiatry. Data for analyses included those assessments that were available and retrievable from computer-based grading systems (Scantron, Irvine, CA, and Questionmark-Perception, Norwalk, CT). The item characteristics and data that were obtained included item format, level of difficulty, discrimination index, and number of foils; neither individual student scores nor item author information were collected or used in any of the analyses. The authors reviewed each item and categorized the format as either case-based or noncase-based (examples are provided in Appendix 1). Items were categorized as case-based consistent with the description in the NAPLEX Bulletin.[5] Exclusion criteria included duplicate items, incorrectly keyed items, items with more than 1 correct response, true-false items, and items with fewer than 4 or greater than 5 foils. A sample size for the number of evaluable items was calculated based on a significance criterion of 95% ($\alpha = 0.05$) and a power of 80% ($\beta = 0.20$). A minimum of 251 items for each format (case-based and noncase-based) was

necessary in order for the study to be adequately powered. Data were analyzed using SPSS (v.15; Chicago, IL); *t* tests for independent samples were performed to evaluate between-group differences. This study was approved by the Shenandoah University Institutional Review Board.

RESULTS

Of the 1,575 unique items that met the inclusion criteria for evaluation, 76% were noncase-based items (Table 1). Standard multiple-choice items were predominant (90%) compared to K-type multiple-choice (10%). The majority of evaluated items (58%) were constructed with 5 foils as compared to those items with 4 foils (42%). Case-based items were not different with respect to level of difficulty ($p = 0.75$), but they demonstrated a significantly lower discrimination index ($p < 0.01$; Table 2). When items were compared based on item structure, K-type multiple-choice items had a higher level of difficulty than standard multiple-choice items ($p < 0.01$), but did not exhibit a significant difference in their respective discrimination index values (Table 3). The number of foils that an item possessed had a significant impact on both level of difficulty and discrimination index (Table 4). As compared to items with 4 foils, 5-foil items were more difficult ($p < 0.001$) and exhibited a higher discrimination index ($p < 0.001$). Of the 1575 evaluated items, 579 items (37%) had discrimination index values ≥ 0.30.

DISCUSSION

The format of the majority of items evaluated in this study was noncase-based (Table 1). The higher levels of abstraction, however, might be better addressed by a case-based format in which the basic levels (eg, knowledge and comprehension) are foundational to the higher competency levels of analysis, synthesis, and evaluation. In a health care professional education program (eg, pharmacy), exposing students to clinically relevant patient-based items would seem rational and justifiable. These clinical scenarios are frequently constructed from actual patient cases and present students with the challenges and thought processes involved in "real life" therapeutic decisions. A possible benefit to using case-based items is that they afford students the opportunity to put therapeutic decision-making competencies into practice. Although some case-based items may be better able to address the higher levels of abstraction, their construction is more challenging and time-consuming and this may explain why there were fewer case-based items (24%) in the current study. Although using clinical cases in therapeutic examinations may address learning and course objectives, it does not allow authentic performance assessment since the plan established is not applied to actual patients.[12]

Interestingly, case-based items were not different from noncase-based items with respect to difficulty level and their discrimination index was lower (Table 2). Although this difference in discrimination index was significant between the 2 item formats, it was relatively small (0.250 and 0.227). This may not correspond to a psychometrically significant difference, particularly since there is not uniform agreement as to what cutoff values delineate a "good" item from an "acceptable" item.

The mean discrimination index values for each category comparison in the present study was below 0.3. Items that have a lower discrimination index may be reflective of core course/content objectives that were repeatedly emphasized in class. As such, those items may not discriminate as highly as other items, but are nonetheless important to include on assessments in order to evaluate comprehension of those learning objectives. For example, hyperkalemia is an adverse effect of angiotensin-converting enzyme inhibitors that is conveyed and reiterated by both basic science and clinical faculty members during the

Table 1. Summary of Descriptive Statistics for Evaluated Items (N = 1575)

Variable	No. (%)
Format	
Case-based item	383 (24)
Noncase-based item	1192 (76)
Structure	
Multiple-choice item (K-type)	152 (10)
Multiple-choice item (standard)	1423 (90)
Number of Foils	
4	662 (42)
5	913 (58)

Table 2. Item Performance Characteristics Based on Item Format

Performance Characteristic	Case-Based Mean (SD)	Noncase-Based Mean (SD)	*P*
Level of Difficulty	76.51 (19.2)	76.86 (18.5)	0.75
Discrimination Index	0.227 (0.1)	0.250 (0.1)	<0.01

Table 3. Item Performance Characteristics Based on Item Structure

Performance Characteristic	Multiple-Choice: Standard Mean (SD)	Multiple-Choice: K-type Mean (SD)	*P*
Level of Difficulty	76.28 (18.4)	72.01 (20.4)	<0.01
Discrimination Index	0.246 (0.1)	0.231 (0.1)	0.21

Table 4. Item Performance Characteristics Based on Number of Foils

Performance Characteristic	Four (4) Foils Mean (SD)	Five (5) Foils Mean (SD)	*P*
Level of Difficulty	79.16 (17.9)	75.04 (19.0)	<0.001
Discrimination Index	0.221 (0.1)	0.262 (0.1)	<0.001

Cardiovascular Therapeutics Module. Although an item related to this core concept may not exhibit a high discrimination index value, its inclusion on an assessment is still worthwhile and appropriate to evaluate student recognition of this clinically important adverse effect.

One question that arises based on these findings is whether instructors and/or the instructional methods provide students with the skills and abilities to evaluate a patient case and to discern between relevant and extraneous information. Additionally, from a pedagogical and assessment perspective, it is appropriate for instructors to evaluate the relationship between their teaching methodology and the types of items that are used to assess that particular content (eg, was case-based instruction—or problem-based learning—offered in proportion to the items that reflected that format?).

The number of foils (Table 4) was a highly significant item component that impacted both the level of difficulty and the discrimination index. The higher number of foils resulted in a greater difficulty level and a higher discrimination index. This difference is somewhat intuitive in that a greater number of response options decreases the probability that a mere guess will result in a correct response. Some recommend that items constructed with 3 or 4 well-written options are sufficient and that it is often difficult and time-consuming to construct a viable fifth option.[13] Additionally, there is no apparent advantage to having a uniform number of foils on an assessment and some items logically necessitate fewer options (eg, [a] increase, [b] decrease, [c] no change).[14]

Another finding in this study was that K-type multiple-choice items were significantly more difficult than standard multiple-choice items (Table 3), but not different with respect to discrimination index—findings which are consistent with other research.[15] Although there are arguments against the use of K-type multiple-choice items,[16,17] this type of item format is utilized on the NAPLEX.[17] At least during the timeframe that the K-type format remains a standard component of the licensing examination, their use on student assessments may be justified.

One of the limitations of the current study is that it did not evaluate the type of content that the items addressed (ie, item content ranged from the basic sciences to clinical therapeutics). Although making those distinctions and categorizations would be an arduous task due to the presence of several content disciplines that are interrelated and coexistent within numerous items (because several disciplines may be represented within the content of each question), future studies could evaluate the relationship between item format and item content. The results of such a study might provide valuable insight into the type of item format best suited for assessing specific types of content. Another limitation of the current study was the subjective determination by the authors of how much patient-case information was necessary to categorize the item as a case-based item. Although the case-based items evaluated in this study were of various lengths and depths (Appendix 1), the primary determinant of the format was the presence (or absence) of patient information deemed essential to selecting the best answer.

In this study, 5-foil items were significantly more difficult and better able to discriminate than 4-foil items. Given that the differences between those groups were relatively small, assessments that utilize items with a mix of 4 or 5 response options appears justified. Also, case-based items were lengthier and had more content associated with them, but were not more difficult than noncase-based items. Additionally, case-based items did not discriminate as well as noncase-based items; although the difference was small, it was significant. Based on our findings, we could not conclude that noncase-based items are superior to case-based items or that the use of case-based items should be discouraged or limited. However, the findings do raise questions about the comparative utility of case-based and noncase-based items, and also about the consistency between instructional methodology and the assessment items that are utilized to evaluate comprehension of course content.

CONCLUSION

Item format, item structure, and number of foils affect item performance characteristics. Case-based items were of comparable difficulty but less discriminating. With respect to item structure, K-type multiple-choice items exhibited a greater level of difficulty than standard multiple-choice items, but both types of structure were similar in their ability to discriminate. Items with 5 foils were more difficult and were more discriminating compared to 4-foil items. Evaluation of item performance characteristics is a valuable component in the item-writing process and can provide constructive insight in assessing student comprehension of course content, as well as aiding faculty members in the development of future items.

ACKNOWLEDGMENT

The authors wish to acknowledge Dr. Wallace Marsh and Dr. Robert Kidd for their assistance and contributions to the manuscript.

REFERENCES

1. McKeachie WJ. *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers.* 11th ed. Boston, MA: Houghton Mifflin Company; 2002:72-80.
2. Bloom BS. *Taxonomy of Educational Objectives: The Classification of Educational Goals.* New York, NY: David McKay Company, Inc; 1956.
3. United States Medical Licensing Examination (USMLE). USMLE Secretariat, Philadelphia, PA; 2009. http://www.usmle.org/ Accessed September 8, 2009.
4. National Board Dental Examinations (NBDE). The Joint Commission on National Dental Examinations, Chicago, IL; 2009. http://www.ada.org/prof/ed/testing/index.asp Accessed September 8, 2009.
5. North American Pharmacist Licensure Examination (NAPLEX) Registration Bulletin. National Association of Boards of Pharmacy, Mount Prospect, IL; 2008. http://www.nabp.net/ftpfiles/bulletins/NAPLEXMPJE.pdf Accessed September 8, 2009.
6. Kehoe J. Basic item analysis for multiple-choice tests. Practical Assessment, Research & Evaluation. http://PAREonline.net/getvn.asp?v=4&n=10 Accessed September 8, 2009.
7. Sim S-M, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple- choice questions of a para-clinical multidisciplinary paper. *Ann Acad Med Singapore.* 2006;35(2):67-71.
8. Pyrczak F. Validity of the discrimination index as a measure of item quality. *J Educ Meas.* 1973;10(3):227-231.
9. Zurawski RM. Making the most of exams: procedures for item analysis. The National Teaching & Learning Forum. 1998;7(6):http://www.ntlf.com/html/pi/9811/v7n6smpl.pdf Accessed September 8, 2009.
10. Newton DW, Boyle M, Catizone CA. The NAPLEX: Evolution, scope, and educational implications. *Am J Pharm Educ.* 2008;72(2):Article 33.
11. Neumann LM, MacNeil RL. Revisiting the national board dental examination. *J Dent Educ.* 2007;71(10):1281-1292.
12. Garavalia LS, Marken PA, Sommi RW. Selecting appropriate assessment methods: asking the right questions. *Am J Pharm Educ.* 2002;66(2):108-12.
13. Kehoe J. Writing multiple-choice test items. Practical Assessment, Research & Evaluation. http://PAREonline.net/getvn.asp?v=4&n=9 Accessed September 8, 2009.
14. Frary B. More multiple-choice item writing do's and don'ts. Practical Assessment, Research & Evaluation. http://PAREonline.net/getvn.asp?v=4&n=11 Accessed September 8, 2009.
15. Sireci SG, Wiley A, Keller LA. An empirical evaluation of selected multiple-choice item writing guidelines. Ellenville, NY: Annual Conference of the Northeastern Educational Research Association; 1998. ERIC Document Reproduction Service ED 428122.
16. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. *Appl Measure Educ.* 1989;2(1):37-50.
17. Wongwiwatthananukit S, Popovich NG, Bennett DE. Assessing student knowledge on multiple-choice examinations using partial-credit scoring of combined-response multiple-choice items. *Am J Pharm Educ.* 2000;64(1):1-10.

Appendix 1. Examples of Item Format

1.Noncase-based items:

a. Multiple-choice items (standard)

Which of the following most accurately describes the mechanism of action of carbidopa:

a. decreases the active transport of levodopa into the CNS
b. increases dopamine release from dopaminergic neurons
c. inhibits peripheral L-aromatic amino acid decarboxylase
d. an agonist at dopamine (D2R) receptors in the CNS
e. inhibits catechol-*O*-methyltransferase (COMT)

b. Multiple-choice items (K-type)

Which of the following pharmacological effects is/are TRUE for both ACE inhibitors and angiotensin (AT1) receptor antagonists:

I. hyperkalemia
II. decrease afterload
III. increase renin release

a. I only
b. III only
c. I and II
d. II and III
e. I, II, and III

2. Case-based item (shorter case):

KN is a 27 year old white male with new onset cluster headaches. He has no known medical problems and is not on any medications. He is presently in his second cluster period and has experienced 2 headaches per day over the last 4 days. His first cluster period was 3 months ago and it lasted for 9 weeks. Which of the following is an appropriate management strategy for KN's cluster headaches?

a. sumatriptan SQ prn headache + verapamil SR 360mg po QD
b. almotriptan po prn headache + methysergide 2mg po BID
c. oxygen via face mask prn headache
d. lithium 300mg po TID

3. Case-based item (longer case):

TS is a 65 year old WM who presents to his physician today for a check-up.

PMH: significant for HTN
FH: father had an MI at age 70; mother is alive and well; no siblings
SH: smokes 1 pack per day; drinks 3-4 beers per night; diet includes lots of fast food
Exercise: no formal exercise program
Meds: HCTZ 25mg 1 daily
Height: 5′ 9”
Weight: 215 lbs
BP: 120/82 mmHg

Fasting lipid profile: (drawn today)

total cholesterol 250mg/dL
triglycerides 100mg/dL
LDL cholesterol 163 mg/dL
HDL cholesterol 35mg/dL

According to the Updated NCEP ATPIII Guidelines, what is this patient's LDL goal?

a. LDL< 160 mg/dL (optional <130mg/dL)
b. LDL < 100mg/dL (optional <70mg/dL)
c. LDL < 130mg/dL (optional <100mg/dL)
d. LDL < 200mg/dL (optional <160mg/dL)