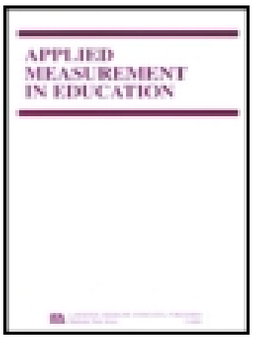

Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: www.tandfonline.com/journals/hame20

A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment

Thomas M. Haladyna, Steven M. Downing & Michael C. Rodriguez

To cite this article: Thomas M. Haladyna, Steven M. Downing & Michael C. Rodriguez (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, Applied Measurement in Education, 15:3, 309-333, DOI: 10.1207/S15324818AME1503_5

To link to this article: https://doi.org/10.1207/S15324818AME1503_5

APPLIED MEASUREMENT IN EDUCATION, *15*(3), 309–334

A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment

Thomas M. Haladyna
College of Education
Arizona State University West

Steven M. Downing
Department of Medical Education
College of Medicine
University of Illinois at Chicago

Michael C. Rodriguez
Department of Educational Psychology
University of Minnesota

A taxonomy of 31 multiple-choice item-writing guidelines was validated through a logical process that included two sources of evidence: the consensus achieved from reviewing what was found in 27 textbooks on educational testing and the results of 27 research studies and reviews published since 1990. This taxonomy is mainly intended for classroom assessment. Because textbooks have potential to educate teachers and future teachers, textbook writers are encouraged to consider these findings in future editions of their textbooks. This taxonomy may also have usefulness for developing test items for large-scale assessments. Finally, research on multiple-choice item writing is discussed both from substantive and methodological viewpoints.

Critics have often noted that item writing is an immature science (Cronbach, 1970; Haladyna, 1999; Nitko, 1985; Roid & Haladyna, 1982). These critics described item writing as a loosely organized set of guidelines mainly transmitted via text-

Requests for reprints should be sent to Thomas M. Haladyna, Arizona State University West, PO Box 37100, Phoenix, AZ 85069–7100. E-mail: tmh@asu.edu

books. The authors of these textbooks are testing specialists who based their advice on personal experience, wisdom, and limited empirical research. Haladyna and Downing (1989a) examined 46 measurement textbook passages dealing with how to write multiple choice items. They produced a set of 43 item-writing guidelines. They found that some guidelines had a strong consensus from these testing specialists. Some guidelines were given lesser attention. Several guidelines were controversial. Coverage of these guidelines in these books varied from very comprehensive to very limited. Commonly, authors did not logically or empirically justify the guidelines they presented.

In a second, complementary study, Haladyna and Downing (1989b) evaluated the results of 93 empirical studies bearing on these guidelines. They found that most of these studies addressed about half the guidelines with the other half unstudied. We might argue that many of these unresearched guidelines were common sense that did not justify research. For example, consider the guideline: "Use novel material to test higher level learning." Most educators place a great value on teaching and testing higher level thinking. Thus, this item-writing guideline may be a consensus value among testing specialists without the need for research.

One of these 43 item-writing guidelines concerns the use of various multiple-choice (MC) item formats, which can vary considerably in structure and cognitive demand (Haladyna, 1999; Martinez,1999). Several important systematic reviews were done to examine the usefulness of these MC item formats (Albanese, 1993; Downing, 1992; Frisbie, 1992; Haladyna, 1992a, 1992b). Since the appearance of these earlier reviews, new research also bears on the validity of using many of these MC formats. These earlier reviews of the effectiveness of various MC item formats suggested that some formats are less desirable than others. However, there has been no systematic study of what writers of measurement textbooks advocate with respect to MC item formats.

Classroom assessment is changing. Most measurement textbooks cited in Appendix A advocate a varied approach to classroom assessments that recognizes the need to teach and assess knowledge, skills, and abilities. Perceived overreliance on the MC format to measure the recall of knowledge instead of higher level learning has resulted in disenchantment with MC testing. A natural aftermath is the increasing use of performance testing that seems better suited for testing complex mental abilities like writing and mathematical problem solving. The standards-based movement has also promoted an approach to teaching and testing in which we recommend a greater variety of assessment approaches. Despite the increased emphasis on performance testing, the MC format continues to play an important role in classroom and large-scale assessments, and this importance is evident in all 27 textbooks on educational measurement reviewed for this study.

Studies by Crooks (1988) and Stiggins, Griswold, and Wikelund (1989) showed that teachers have a difficult time assessing complex abilities, such as writ-

ing or mathematical problem solving. Although we emphasize developing these complex abilities in students, we lack the technology of test-item writing for assessing these complex abilities. Therefore, we need better item formats and clear, valid guidelines to help us write test items to better assess complex student learning.

The setting for this study is classroom assessment, which is also the setting for the measurement textbooks featured in this study and listed in Appendix A. However, we see implications of this study for large-scale assessment programs, which may involve promotion, graduation, certification, licensure, training, or program evaluation.

This study examines and evaluates two sources of evidence bearing on the validity of 31 MC item-writing guidelines intended for teachers and others who write test items to measure student learning. These two sources are measurement textbooks and research. One of these guidelines addresses the validity of using different types of MC item formats. So a separate part of this article deals with textbook endorsements and use of these formats in these textbooks and research on the validity of each MC format. The specific question driving this study was: What is the validity of using each MC item-writing guideline?

METHOD

The original taxonomy proposed by Haladyna and Downing (1989a) was reorganized while conducting this study. This revised taxonomy appears in Table 1. We intended the changes to help users better understand each guideline in its context. For example, eight guidelines are provided in *Content Concerns;* two guidelines were listed under *Formatting Concerns;* three guidelines were listed under *Style Concerns;* four guidelines were listed under *Writing the Stem;* and fourteen guidelines are presented in *Writing the Choices* with six variations for Guideline 28, which deals with providing clues to the right answer.

Our strategy for validating each guideline was to consider two sources of evidence. The first source was the collective opinions of textbook authors. The second source was empirical research. We made our collective judgment about the validity of each guideline after considering both sources of evidence.

Textbook Review

Two of the coauthors were randomly assigned to read each passage identified in Appendix A bearing on writing MC items. Each coauthor classified whether each guideline was cited and supported, cited and not supported, or not cited. The page number on which the guideline was cited or applied was noted. This method is essentially the same used by Haladyna and Downing (1989a). In instances in which

TABLE 1
A Revised Taxonomy of Multiple-Choice (MC) Item-Writing Guidelines

Content concerns

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications (two-way grid, test blueprint).
2. Base each item on important content to learn; avoid trivial content.
3. Use novel material to test higher level learning. Paraphrase textbook language or language used during instruction when used in a test item to avoid testing for simply recall.
4. Keep the content of each item independent from content of other items on the test.
5. Avoid over specific and over general content when writing MC items.
6. Avoid opinion-based items.
7. Avoid trick items.
8. Keep vocabulary simple for the group of students being tested.

Formatting concerns

9. Use the question, completion, and best answer versions of the conventional MC, the alternate choice, true-false (TF), multiple true-false (MTF), matching, and the context-dependent item and item set formats, but AVOID the complex MC (Type K) format.
10. Format the item vertically instead of horizontally.

Style concerns

11. Edit and proof items.
12. Use correct grammar, punctuation, capitalization, and spelling.
13. Minimize the amount of reading in each item.

Writing the stem

14. Ensure that the directions in the stem are very clear.
15. Include the central idea in the stem instead of the choices.
16. Avoid window dressing (excessive verbiage).
17. 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.

Writing the choices

18. Develop as many effective choices as you can, but research suggests three is adequate.
19. Make sure that only one of these choices is the right answer.
20. Vary the location of the right answer according to the number of choices.
21. Place choices in logical or numerical order.
22. Keep choices independent; choices should not be overlapping.
23. Keep choices homogeneous in content and grammatical structure.
24. Keep the length of choices about equal.
25. *None-of-the-above* should be used carefully.
26. Avoid *All-of-the-above*.
27. Phrase choices positively; avoid negatives such as NOT.
28. Avoid giving clues to the right answer, such as
    a. Specific determiners including always, never, completely, and absolutely.
    b. Clang associations, choices identical to or resembling words in the stem.
    c. Grammatical inconsistencies that cue the test-taker to the correct choice.
    d. Conspicuous correct choice.
    e. Pairs or triplets of options that clue the test-taker to the correct choice.
    f. Blatantly absurd, ridiculous options.
29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.
31. Use humor if it is compatible with the teacher and the learning environment.

two reviewers disagreed on a guideline citation, Haladyna noted the page reference and reread that page for a decision. In each instance of a disagreement, consensus was achieved through this process.

Research Studies

We reviewed studies published from 1990 to the present, picking up where the Haladyna and Downing (1989b) review ended. The studies reported here came from conference proceedings, review of electronic databases of educational and psychological articles, and references provided by each article identified. The 19 studies that evaluated differences in item and test statistics for different formats were narrowly focused in subject matter and age of examinees. Studies included tests given to medical, dental, and nursing students; a selection test for entry-level police officers; tests in psychology and communications undergraduate courses; a biology test, the American College Testing (ACT) Assessment, and a science test for middle school students. Several of these empirical studies addressed more than one guideline. Four review articles were published bearing on these guidelines. All studies are listed in Appendix B.

MC Item Formats

Guideline 9 concerned the validity of seven MC item formats. The same procedure was used. We identified formats as supported, not supported, or uncited in the textbooks reviewed. With instances of disagreement between two reviewers, Haladyna adjudicated.

Judgments About Validity of Each MC Format

Each of the 31 item-writing guidelines was subjected to the same validation procedure. We evaluated and classified the authors' treatments of each guideline as cited and supported, cited and not supported, or not cited. The MC item formats were treated the same way. We reviewed research on each guideline. The summative judgment by Haladyna, Downing, and Rodriguez for each guideline was based on these two sources of evidence. The validity of any guideline may change depending on a new, compelling, logical argument and the collective old and new evidence bearing on this argument.

WHAT IS THE VALIDITY OF EACH ITEM-WRITING GUIDELINE?

Table 2 presents a summary of our tabulations for 31 guidelines. Guideline 10, which deals with the validity of various MC item formats, was omitted because it is treated separately in another section.

TABLE 2
Frequency of Citation for Each Item-Writing Guideline

Guideline	*For (%)*	*Uncited (%)*	*Against (%)*
1. Singe content and behavior	74	26	0
2. Important, not trivial content	78	22	0
3. Use novel material	85	15	0
4. Keep items independent	48	52	0
5. Avoid over specific/general	15	85	0
6. Avoid opinions	26	74	0
7. Avoid trick items	67	33	0
8. Simple vocabulary	70	30	0
9. Format vertically	37	52	11
11. Edit and proof	33	67	0
12. Correct grammar	52	48	0
13. Minimize reading	67	33	0
14. Clear directions	82	18	0
15. Central idea in stem	100	0	0
16. Avoid window dressing	52	48	0
17. Use positive, no negatives	63	19	18
18. Write as many plausible distractors as you can	70	26	4
19. One right answer	70	30	0
20. Vary location of right answer	52	48	0
21. Logical/numerical order	67	33	0
22. Choices not overlapping	30	70	0
23. Choices homogeneous	67	33	0
24. Choice length equal	85	15	0
25. Use carefully *None of the above*	44	7	48
26. Avoid *All of the above*	70	7	22
27. Avoid NOT in choices	70	30	0
28. Avoid clues	96	4	0
29. Make distractors plausible	96	4	0
30. Use common errors of students	70	30	0
31. Use humor sparingly	0	85	15

Unanimous Author Endorsements

Although the number of guideline citations ranged considerably among textbooks, nearly all guidelines received unanimous endorsements when they were cited. These unanimously endorsed guidelines are 1–8, 11–16, 19–24, and 27–30. The percentage of textbooks citing and endorsing these guidelines ranges from 15% (5—Avoid overly specific/general content) to 100% (15—Central idea in the stem). About 47% of these guidelines were cited in more than 70% of these textbooks, suggesting that we have a common core of MC item-writing guidelines. In this section we discuss guidelines in which disagreement exists or empirical evidence exists that bears on the validity of the guideline or both.

Avoid Trick Items

Although this guideline is unanimously endorsed by textbook authors, Roberts (1993) provided some empirical evidence for this guideline's validity. He did a comprehensive review of this topic and completed the only known study on trick items. He asked 174 students and 41 faculty members what they thought constituted a trick test item. Overwhelmingly, respondents believed that trick items were deliberately used on tests to deceive students. The defining characteristics of trick items, in order of prevalence, included intention, trivial content, too fine answer discrimination, stems with window dressing, multiple correct answers, content presented opposite from instruction, and high ambiguity. Conventional MC item formats were most likely to be a trick item (48% of the respondents) rather than true-false (TF), matching, essay, or short-answer items. When 13 trick items were written with 12 nontrick items in an introductory statistics test, students were not able to differentiate between the two. However, trick items were more difficult than were nontrick items, lending some credibility to their trickiness.

We should obviously avoid trick items. Roberts (1993) concluded that textbook authors do not give trick items much coverage, owing perhaps to the fact that trick items are not well defined and examples of trick items are not commonplace. His conclusion is consistent with results reported in Table 2 showing that 33% of the textbooks failed to mention trick items. Another good point is that the accusation of trick items on a test may be an excuse by students who lack knowledge and do poorly on a test. We continue to support the idea that trick items should never be used in any test and recommend that the basis for teaching about trick items be consistent with Roberts's findings.

Simple Vocabulary

The achievement construct being measured can be affected by the reading demand placed on the student by the item. Difficult vocabulary places some students at risk. Of the textbooks discussing the vocabulary of MC items, all agree that vocabulary should be appropriate for students being tested. Abedi, Lord, Hofstetter, and Baker (2000) have experimented with simplified language in MC items with promising results for English language learners, supporting Guidelines 12 and 13. The team that developed the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) has dedicated an entire chapter to the problem of testing English language learners. Simplified language is an effective way to reduce the influence of reading ability, a source of construct-irrelevant variance when the achievement test is intended to measure something else. So we continue to support this guideline.

Format Vertically, not Horizontally

Published tests are usually vertically formatted for easy reading, with the stem and options following in a column. Two-column printing is also very common in published tests along with vertical presentation of items because test takers can more easily read items appearing this way. Also, the presentation of items is more compact. For the sake of saving space in a measurement textbook, some publishers may inadvertently format items horizontally. Sometimes, authors seem to favor this format. Also, we have no research evidence to argue that horizontal formatting might affect student performance. Nonetheless, we side with the authors who format their items vertically. Textbook publishers may want to conserve on space when printing these books, but they do not do a service here to the quality of the appearance of MC tests.

Put the Central Idea in the Stem

All textbook authors believe that we should state the central idea of the test item in the stem. The key item-writing fault that violates Guideline 15 is the unfocused stem:

Validity

A. Should refer to tests not to test scores.
B. Bears on the interpretation or use of a test score.
C. Is as important as reliability.

Downing, Dawson-Saunders, Case, and Powell (1991) evaluated the effects of unfocused stems. They reported no significant differences in difficulty or discrimination, although the effects were compounded with the simultaneous use of heterogeneous options. They suggested that students studying for a profession, such as medicine, have enough experience with MC items that may ameliorate effects of poor item writing. We continue to support this guideline and argue that unfocused stems should be avoided.

Use Positives, not Negatives, in the Stem

The survey of authors' treatments of MC item writing shows that 63% support this guideline, and 19% think using negatives is all right, with another 19% not discussing this guideline. Three of the four researchers found no difference in item difficulty due to the use of negatively worded stems (Downing et al., 1991; Rachor & Gray, 1996; Tamir, 1993). Harasym, Price, Brant, Violato, and Lorscheider (1992) found negatively worded stems to be less difficult, although this result was inconsistent and less important than item content and format (sin-

gle- vs. multiple-response MC items). However, Tamir found that for items written at a higher cognitive level, negation made items significantly more difficult. Rachor and Gray and Downing et al. found no difference in discrimination. Downing et al. and Harasym et al. reported no significant change in reliability due to stem orientation.

Harasym et al. (1992) recommended the use of the multiple true-false (MTF; see Table 3) format instead of negatively worded single-response items in their evaluation of medical education tests. This result was subsequently reinforced in a later direct comparison of single-response negatively worded items and corresponding multiple-response positively worded items (Harasym, Doran, Brant, & Lorscheider, 1993).

The use of negatively worded items should be done with caution, when necessary to measure a relevant objective (e.g., what to avoid or what is not true), and with the negative term highlighted in some way. Under most circumstances, we suggest that a stem should be worded positively, but if a stem must contain a negative term, it should appear in capital letters and boldface.

Write as Many Plausible Distractors as You Can

Textbook authors are somewhat divided. Most like the idea of writing as many plausible distractors as possible, and a few textbooks emphasize four options, which may be a standard in the testing industry. Surprisingly, some textbooks do not mention the desirable number of options.

The desirable number of options for a MC test item has been an issue of considerable interest to researchers. In fact, this guideline has received the most empirical study among all other guidelines. Seven research studies of the number of options were reported since 1989. Results regarding the effect on difficulty of MC items as a function of the number of options were mixed. Five recent studies found that decreasing the number of options results in decreases in difficulty (Landrum, Cashin, & Theis, 1993; Rogers & Harley, 1999; Sidick, Barrett, & Doverspike, 1994; Trevisan, Sax, & Michael, 1991, 1994), whereas two others found increases in difficulty (Cizek & Rachor, 1995; Crehan, Haladyna, & Brewer, 1993). Cizek and Rachor also found an increase in discrimination with a decrease in number of options, whereas Crehan et al. found no change in discrimination. Results on reliability were in opposition to most previous results. Trevisan et al. (1994) and Sidick et al. reported an increase in reliability with fewer options, although Trevisan et al. (1994) reported no difference. Trevisan et al. (1991) and Rogers and Harley were the only ones to evaluate a decrease in criterion-related validity evidence with fewer options. Andrés and Castillo (1990) and Bruno and Dirkzwager (1995) also supported earlier theoretical work supporting the utility of three-option items in terms of information obtained.

Haladyna and Downing (1993) evaluated the distractors from four standardized MC tests. They found that about two thirds of all items had one or two effectively performing distractors, with between 1% and 8% of all items having three effective distractors. Although no relation existed between the number of effective distractors and item difficulty, items with more effective distractors were more discriminating. Overall, the modal number of effective distractors per item was one.

We support the current guideline, but we also think that three options are sufficient in most instances. The effort of developing that fourth option (the third plausible distractor) is probably not worth it. If the use of four options is preferred, empirical research has established that it is very unlikely that item writers can write three distractors that have item response patterns consistent with the idea of plausibility.

Place Options in Logical or Numerical Order

This guideline is unanimously supported in textbooks. Not only is it logical to place options in logical or numerical order, but it improves the appearance of the test. Huntley and Welch (1993) evaluated performance of 32 mathematics items during an ACT Assessment pretest session in which items were written in two formats: (a) distractors in ascending or descending (logical) order and (b) distractors in random order. Although no differences in average item difficulty were noted, discrimination was higher among items with randomly ordered distractors. They concluded that random ordering of options may pose obstacles for lower ability students (based on performance on the ACT Assessment mathematics test that was administered at the same time as the experimental pretest items). These researchers supported logical or numerical ordering of options, and we agree with them.

Keep Options Homogeneous in Content and Grammatical Structure

Like other guidelines, the idea is to make options homogeneous to focus the student on the main idea of the test item better and to improve discrimination of the item. This guideline received unanimous endorsement in 67% of these textbooks, with 33% of these textbooks not mentioning this guideline. Downing et al. (1991) also investigated the use of homogenous versus heterogeneous options, again simultaneously with the use of focused versus unfocused stems in exams used by the National Board of Medical Examiners. No significant differences were found in difficulty or discrimination. Although empirical evidence is lacking, the high degree of consensus among authors favors the continued use of this guideline.

None of the Above Should Be Used Carefully

Authors have split their support for this guideline. Although 44% agree with this guideline, 48% believe that none of the above (NOTA) should not be used. All five research studies investigating NOTA reported that its use increased item difficulty (Crehan & Haladyna, 1991; Crehan et al., 1993; Frary, 1991; Kolstad & Kolstad, 1991; Rich & Johanson, 1990). Three of the four who evaluated discrimination reported no difference (Crehan & Haladyna, 1991; Crehan et al., 1993; Frary, 1991), whereas Rich and Johanson reported an increase in discrimination due to NOTA. Kolstad and Kolstad and Rich and Johanson found no change in reliability.

Frary (1991) suggested that cautious use of NOTA could be compatible with good classroom testing when consistent with instructional objectives and when items would otherwise be too easy. Likewise, Kolstad and Kolstad (1991) suggested item writers keep an open mind regarding NOTA and that it be restricted to "items that can logically include one correct option only" (p. 162). Similarly, Rich and Johanson (1990) found improvement in discrimination and difficulty that approached optimal levels with the use of NOTA.

Gross (1994) argued for logical rather than empirical guidelines for item writing, particularly regarding the use of NOTA. He suggested "any stem or option format that by design diminishes an item's ability to distinguish between candidates with full versus misinformation, should not be used" (p. 125). This idea is not new. Although Ebel (1951) addressed this danger when using NOTA, he also suggested that occasional use of NOTA in items in which the possible incorrect responses are relatively few is appropriate.

Given recent results and these arguments, NOTA should remain an option in the item-writer's toolbox, as long as its use is appropriately considered. However, given the complexity of its effects, NOTA should generally be avoided by novice item writers.

Avoid All of the Above

Most textbook authors (70%) support the all of the above (AOTA) guideline, but 7% did not cite it, and the remaining textbooks supported the use of AOTA. Harasym, Leong, Violato, Brant, and Lorscheider (1998) compared items written in single-response format with the AOTA option as correct with identical items written in MTF format. MTF items were far more difficult than the corresponding single-response AOTA items. Most examinees selected at least two options in the MTF format, whereas nearly all selected AOTA when it was presented. Haraysm et al. suggested that this was due to cuing. Reliability was significantly reduced with the AOTA items, supporting the avoidance of AOTA. We continue to support this guideline to avoid AOTA.

Use Humor Sparingly

A significant review and study of humor in items was done by McMorris, Boothroyd, and Pietrangelo (1997). They defined humor, identified types of humor, considered purposes of humor in testing, and then reviewed the research on humor in testing. Whereas some studies they cited also appeared in the Haladyna and Downing (1989b) review, McMorris et al. concluded that humor is probably a good thing for classroom assessment. Based on the results of their study and the studies cited, we concur that humor is probably a good thing for classroom assessment but only if the overall good exceeds any bad that might come from the use of humor. Humor might be used in the classroom if it reflects the character of the teacher and is compatible with the learning environment of the class. However, in formal testing programs with a high-stakes consequence, such as with graduation, promotion, certification, and licensing, humor should probably not be used.

GUIDELINE 9: THE VALIDITY OF SEVEN MULTIPLE-CHOICE ITEM FORMATS

Table 3 presents seven MC formats that have been used in textbooks and are the subject of research. As noted in Table 1 in Guideline 9, only the complex MC format was not recommended for classroom assessment. All other formats are recommended. This section provides the validity evidence that led to Guideline 9.

Table 4 summarizes the use of these seven formats in the 27 textbooks that we reviewed. In addition to the authors' endorsements of formats, we also evaluated the few studies we found bearing on the use of each format.

Conventional Multiple Choice

The first of these seven formats is the very familiar conventional MC, which is widely used in achievement tests at all levels and with most types of content. All textbooks recommend this format, and all test publishers use the conventional MC item format in their standardized tests. This is truly the quintessential MC format.

Two common conventional MC variations are the question stem and the completion stem. An issue with the use of conventional MC is whether we should state the stem as a question or as part of a sentence, with the options completing the sentence. Haladyna and Downing (1989a) presented an argument supporting the question format based on an essay by Statman (1988). Three authors studied the issue of whether to write the stem as a complete question or a statement that is completed

TABLE 3
Multiple-Choice (MC) Item Formats

Conventional MC

Which of the following most clearly defines the process of pollination?

A. The joining of egg and sperm cells.
B. The transfer of pollen grains to the pistil.
C. Food is broken down and energy is released.

Alternate-Choice

Which of the following would most effectively slow down the process of respiration in plants?

A. Cold weather
B. Stormy weather

True-False

The capital of Uruguay is Montevideo.

Multiple True-False

You are an expert organic farmer. You know the secrets of growing strong, healthy plants. Which of the following would describe your farming practices? (Mark A if true, B if false.)

1. When you plant some beans you make certain that the beans will be well shaded to receive little to no light.
2. When you plant your seeds you make sure to water them and continue to keep the soil moist.
3. You plant your seeds only when the temperature is appropriate.
4. Because you know how pollination occurs, you spray your crops with insecticides to prevent bees and other insects from harming your crops.

Matching

Match each term on the right with the description on the left.

1. Attracts bees	A. Pollen grains
2. Produces pollen grains	B. Petals
3. Houses the egg cells	C. Flower
4. Seeds are formed	D. Stamen
5. Contains the ovary	E. Ovary
	F. Pistil

Complex MC

Which of the following are fruits?

1. Tomatoes
2. Tomatillos
3. Habanero peppers

A. 1 & 2
B. 2 & 3
C. 1 & 3
D. 1, 2, & 3

(continued)

TABLE 3 (*Continued*)

Context-Dependent Item Set
Imagine you are a delegate from Massachusetts to the Constitutional Convention. You have been authorized to act on behalf of your state.

1. You would most likely approve of the
   A. New Jersey Plan.
   B. Virginia Plan.
2. You would oppose the three-fifths compromise because
   A. Your state, as a rule, is strongly abolitionist.
   B. You will be grossly outrepresented in Congress by northern states.
   C. You want only a single representative house.
3. You support the suggestion that Congress tax
   A. Imports.
   B. Exports.
4. Because of your state's experience with Shays' Rebellion, you feel
   A. Farmers should not have to carry the tax burden for townspeople.
   B. Native Americans must be pacified before there can be peace.
   C. Tories ought to pay reparations.

TABLE 4
Types of Multiple-Choice (MC) Formats and Frequence of Citation in 27 Recent Textbooks

Format	*Description*	*% Cited*
Conventional MC	Stem and 3 to 5 options	100
Alternate-choice	Stem and 2 options	11
Matching	3 to 12 options preceding a group of stems	100
Multiple true-false (MTF)	Stem with 3 to 30 options. Each option is evaluated in terms of its truthfulness	37
True-false	Declarative statement evaluated in terms of its truthfulness	100
Context-dependent item, including the item set (testlet)	A stimulus following by one or more items. Responding to each item is dependent upon the stimulus material. The MTF is actually a unique type of item set	70
Complex MC	A stem followed by choices that are grouped into sets for test-takers to choose	31

by the options. All three found no differences in discrimination between the two formats (Crehan & Haladyna, 1991; Eisley, 1990; Rachor & Gray, 1996). Only Eisley found a difference in difficulty, in which items with stems written as complete questions were more difficult. Eisley also evaluated reliability and found no difference. On the basis of textbook discussions on conventional MC items and

these research findings, it would seem that item writers should use either the question format or the sentence-completion format. However, we prefer the question format over the sentence-completion format for its directness in getting to the central idea of the test item.

Alternate-Choice

The alternate-choice (AC) is essentially a two-option MC item. Ebel (1982) produced a strong argument for why the AC should be used. The AC is easier to write and administer, and it can be used to test a variety of content and cognitive behaviors. In a theoretical article by Lord (1977), the logical argument for the AC format was that most moderate- or high-performing students typically narrow their choices of plausible options to two. A review by Downing (1992) summarized empirical support for this format. A study by Haladyna and Downing (1993) showed that most items naturally are AC when nonfunctioning distractors are eliminated. The limitation imposed by guessing is not a serious one. We can adjust standards for interpreting AC item responses to consider the fact that 50% is a lower bound for performance on AC items. If the test is long enough, guessing has little influence. Textbooks seldom mention this format (11%). Because the AC item is easier to write and may perform as well as conventional MC, the use of the AC format seems very desirable from the item writer's standpoint. So we endorse this item format.

True-False

The True-False (TF) format is very popular. TF items are also referred to as AC (Osterhof, 1999), two-choice (Gallagher, 1998), and binary choice (McMillan, 2001). A slight variation of the TF is when a statement is given and the student chooses an answer from yes or no, right or wrong, correct or incorrect, fact or opinion, or any other bipolar set of terms that can be keyed right or wrong. The TF format has one thing in common with AC, but the two have distinctly different anatomies. Grosse and Wright (1985) detected response styles that influenced student scores among other difficulties with this format. Haladyna (1999) and Downing (1992) also reviewed the accumulating research evidence and found problems with the use of TF items that have not yet been resolved. Frisbie and Becker (1991) provide a very comprehensive review of research on TF items. One of the most comprehensive treatments of TF item writing can be found in Ebel and Frisbie (1991). Nonetheless, this format continues to survive in classroom testing, as any textbook listed in Appendix A will attest.

Multiple True-False

As noted previously, the MTF is an unusual format that has some very positive characteristics (Frisbie, 1992; Frisbie & Sweeney, 1982). This format is a hybrid

MC and TF. Choices follow a leading question or scenario, and the examinee evaluates each choice as true or false, as shown in Table 3. Such items are easier to write than conventional MC. Guessing may be a problem, but with a set of 30 MTF items, the influence of guessing can be reduced, as it is with an equal number of TF and AC items.

Downing, Baranowski, Grosso, and Norcini (1995) found that MTF items produced more reliable scores. They also found that the type of cognitive behavior exhibited in MTF scores was lower level than found in conventional MC items.

Matching

The matching format requires a set of options followed by a set of matching stems (statements, questions, or phrases). All textbooks cited in Appendix A leave no doubt that good matching items can be developed. However, this format has no reported research on its effectiveness, unlike all other formats shown in Table 3. The matching format is not often used in published standardized achievement tests. Given the good account of matching items in all textbooks, we continue to recommend this format for classroom assessment but hesitate to recommend it for high-stakes testing until research supports its use.

The Complex Multiple Choice (Type K)

Considering the complex MC, 10 textbook passages (37%) mentioned this format. Of these 10 references, 6 opposed this format, 3 supported it, and 1 gave a mixed endorsement. The MTF format is a replacement for the complex MC, having received some favorable research support in comparison studies (Downing et al., 1995; Frisbie, 1992). In a review of research by Albanese (1993), the complex MC fared poorly. Nnodim (1992) suggested that complex MC items were more difficult than conventional MC items but with no difference in discrimination in human anatomy tests given to medical students. The complex MC is also longer than most other MC formats and requires more administration time, rendering it less efficient than other MC formats. As a result of these discussions of the complex MC and the research studies, this format is not recommended for use.

The Context-Dependent Item and Item Set

The context-dependent item format usually has a scenario, vignette, table, chart, graph, reading passage, or other visual material followed by a single item. Of the textbooks reviewed, 70% show this format and discuss its usefulness for measuring types of higher level thinking such as mathematical problem solving and criti-

cal thinking. The examples provided in these books are exemplary of the variety that exists with this format. The context-dependent item takes up considerable space in a test and requires a longer administration time, two very limiting factors for arguing its use.

An important variation of the context-dependent item format is the item set (Haladyna, 1992a, 1992b). Of the 19 textbooks discussing context-dependent items, 14 gave examples of the item set. Item sets have many other names, including testlets, interpretive exercises, context-dependent item sets, and superitems. The value of the item set may be its ability to test the application of knowledge and skills to a more complex set of behaviors, such as mathematical problem-solving ability, reading comprehension, writing skills (through the interlinear form), and with a stimulus material, such as a picture, chart, graph, or vignette (Haladyna, 1999). Few authors mentioned the local dependence problem with scoring (Haladyna, 1992a). An answer to one item in the item set may affect how one answers other items in that set. No authors used the term *testlet*, even though test analysts regularly refer to item sets as testlets (e.g., Wainer & Kiely, 1987). Worthen, White, Fan, and Sudweeks (1998) provide one of the best discussions of item sets among these textbooks. Linn and Gronlund (2000) also have a good section on context-dependent item sets. We are enthusiastic about the use of item sets for measuring many important types of student learning. The item set format is increasingly being used in large-scale assessments due to its ability to simulate thought processes of a multistep nature that an extended performance test item might measure with greater fidelity.

CONCLUSIONS

About the Guidelines (Excluding Guideline 9)

Textbook consensus and research studies have provided validity evidence supporting the use of most of the MC item-writing guidelines listed in Table 1. These guidelines do not have equal degrees of evidence supporting their use. All guidelines are subject to future validation studies, as our understanding of the technology of MC item writing continues to develop and new research provides more insight into the validity of each guideline.

About Multiple-Choice Formats (Guideline 9)

The conventional MC, TF, and matching formats are the mainstays of classroom assessment. The context-dependent item seems well established for measuring different types of higher level thinking but with known limitations of having a longer time for development and longer administration time. The context-dependent item

set has greater potential for efficiency than the single stand-alone format, but item responses are dependent, which may lead to an overestimate of reliability. The AC and MTF seem much neglected in these textbooks. Research supports the use of both MC formats. The complex MC should not be used for many good reasons offered in this research report.

IMPLICATIONS OF THIS STUDY FOR CLASSROOM AND LARGE-SCALE ASSESSMENT PROGRAMS

Those who will write future measurement textbooks or who teach in teacher-preparation programs or do MC item-writing training have an empirically validated set of guidelines to apply. Educational reform continues to emphasize the use of content standards both for classroom assessment and high-stakes test score uses. Test items need to be linked to content standards both in our classroom assessments and in our high-stakes tests. Instruction should link to these assessments so that students learn and perform according to expectations. The use of a toolbox full of validated item-writing guidelines and MC item formats serve items writers well, particularly teachers who traditionally have had difficulty writing items measuring complex student learning.

Large-scale assessment programs tend to have high standards for item development and employ guidelines consistently. These programs will probably continue to depend on conventional MC and context-dependent item sets as they have in the past. TF and matching formats are unlikely to be used in these assessment programs. Although research suggests that MTF and AC have promising properties to recommend their use, they too are unlikely to be used, perhaps because testing programs tend to be conservative in their use of item formats. The complex MC should not be used.

IMPLICATONS OF THIS STUDY FOR RESEARCH ON ITEM WRITING

In this final section, we discuss (a) the current proposed taxonomy of item-writing guidelines, (b) the need for validated item-writing theories and related technologies for item writing, (c) the need for a validated taxonomy of cognitive behavior that we can use in item-writing research, and (d) systematic research on item formats that provides better guidelines for when we should use MC and constructed-response formats. This final section speculates about how to move MC item-writing science forward.

The Current Taxonomy of Item-Writing Guidelines

The current taxonomy of item-writing guidelines is limited to the extent that it does not advance item writing as a science but merely gives practitioners some useful advice to write items that reflect intents of instruction.

Research on these MC item-writing guidelines has been asystematic. The number of studies reported since the Haladyna and Downing (1989b) research synthesis is small relative to the number of guidelines. Four guidelines receiving the most attention are 17 (the use of negative words in the stem), 18 (the desirable number of options), 25 (NOTA), and 26 (AOTA). The plausibility of distractors (Guideline 29) continues to be an area that is long overdue for study, particularly as it applies to Guideline 18 concerning the number of options.

Need for Validated Item-Writing Theories to Motivate Future Research

Toward a Technology of Test-Item Writing (Roid & Haladyna, 1982) reported on the status of item-writing theories current to that time. None of these theories survived. Frederiksen, Mislevy, and Bejar (1993) edited a volume of essays on new theories and emerging technologies in testing and insights, but item formats were only one of several important issues discussed by various contributors to this edited book. Another edited volume by Bennett and Ward (1993) dealt with important issues related to constructed-response and the MC formats. With the publication of these essays, theories, and research, the scientific basis for writing test items appears to be improving but very slowly. We still lack widely accepted, item-writing theories supported by research with resulting technologies for producing many items that measure complex types of student learning that we desire.

Need for a Validated Taxonomy of Cognitive Behavior

The Bloom taxonomy has existed for nearly 50 years but has not been empirically validated. We currently lack a validated taxonomy of cognitive behavior that we widely accept and easily apply to classroom assessment. The important chapter by Snow and Lohman (1989) helped us gain understanding about the future role of cognitive psychology in testing. The edited volume by Frederiksen et al. (1993) provides a glimpse of the pending merger of cognitive and psychometric science.

However, a review of all of the textbooks in Appendix B reveals a variety of approaches to cognitive behavior that may have some commonality but no universality. Cognitive psychology should produce a new cognitive taxonomy that helps us write items that satisfy various cognitive demands.

Systematic Research on Item Formats

Martinez (1999) offers a very good description of cognitive processes measured by test item formats and the meaning we attach to student responses. Martinez concluded that the connection between item formats and cognition is not well established but clearly should be in our item-writing research agenda.

We have seen some current study of the issue of MC versus constructed-response item formats (Bennet & Ward, 1993; Haladyna, 1999; Martinez, 1999; Rodriguez, 2002; Ryan & Demark, 2002). This topic continues to be important as states continue to employ performance formats and experiment with MC formats that promisingly try to measure complex cognitive learning. As we continue to study the characteristics of item formats in relation to student learning, we are discovering more threats to validity.

For example, several researchers noted differential effects among groups of different ability levels—or lack of effects among highly capable students. The use of more complex or unfamiliar item formats may not result in negative consequences for high-ability students but may result in negative consequences for low-ability students. Researchers have suggested that this may be due to test-wiseness (Downing et al., 1991; Rachor & Gray, 1996) or simply due to ability (Huntley & Welch, 1993; Rachor & Gray, 1996). Thus, some item formats may add a construct-irrelevant barrier to performance by low-ability students. If test-wiseness is a component of the knowledge required to correctly answer a set of items, then construct-irrelevant variance is introduced. Test-wiseness may be an important factor related to the usefulness of item formats.

A good point made by DeMars (2000) about item formats is that although the construct being represented by MC and constructed-response test scores may be highly correlated, the cost of this information may differ greatly. Another point is that if the selection of one format favors a certain instructional approach over another, we would need to know the consequences of choosing one format over another. Thus, future research should consider these factors in evaluating the benefits of any item format.

FUTURE RESEARCH METHODS

Experimental studies are desirable for the study of the influence of manipulated variables on item difficulty and discrimination and the kinds of cognitive behaviors elicited by items and different item formats. Within the broad range of potential experimental studies that could be conducted, we argue for a principled and informed approach. Researchers must clearly disclose the methods used to design instruments and the items under investigation. Characteristics of test design may

play a role in the impact of violating a given item-writing guideline. For example, is the effect of NOTA different when it is used as a distractor versus the correct option? When constructing items to test the impact of the number of options, did the researcher write items with five options and sequentially delete options (randomly or based on item statistics) to create the other forms or were options added to create new items?

Similarly, it is important to describe all other relevant features of items to interpret results of any given guideline or item-writing technique under investigation. For example, if a researcher is studying the role of negation in the stem, is the stem a complete question or a statement completed by the options? In the same study, how many options are provided per item? Are the options ordered logically?

Finally, with respect to experimental studies, rarely have researchers evaluated effects of item-writing techniques vis-à-vis individual characteristics, such as age, gender, ability level, or the cognitive demands of individual items. The fact that some researchers have found interesting interactions with these additional variables requires future researchers to address these whenever possible.

A promising way of studying item writing comes from cognitive psychology and involves interviewing students as they take a test (Norris, 1990; Trelease, 1985). The think-aloud procedure can provide insights into the cognitive processes underlying a student's encounter with a test item, but its limitation is the time it takes to collect data and evaluate the findings.

Finally, the research review is still a valuable tool. The review of research on trick questions by Roberts (1993) exemplifies the potential for a review that significantly advances our understanding of a single MC item-writing guideline. The review and discussion of humor in item writing by McMorris, Boothroyd, and Pietrangelo (1997) also show the effectiveness of a concerted effort on a single item-writing guideline.

CLOSING

The science of MC item writing is advancing, but item writing is still largely a creative act that we inexorably link to content standards and instruction. As Ebel (1951) reminds us, item writing is mostly creative:

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

gestions that determines whether good items or mediocre ones will be produced. (p. 185)

REFERENCES

Albanese, M. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28–33.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington DC: American Education Research Association.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J. (1970). [Review of the book *On the theory of achievement test items*]. *Psychometrika, 35,* 509–511.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438–481.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13,* 55–77.

Downing, S. M. (1992). True-false and alternate-choice item formats: A review of research. *Educational Measurement: Issues and Practice, 11*(3), 27–30.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. R. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education, 8,* 189–199.

Ebel, R. L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational Measurement* (1st ed., pp. 185–249). Washington, DC: American Council on Education.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19,* 267–278.

Frederiksen, N., Mislevy, R. J., & Bejar, I. (Eds.). (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 5*(4), 21–26.

Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4,* 67–83.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19,* 29–35.

Grosse, M., & Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement, 45,* 1–13.

Haladyna, T. M. (1992a). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*(1), 21–25.

Haladyna, T. M. (1992b). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5,* 73–88.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1,* 51–78.

Lord, F. M. (1977). Optimal number of choices per item—A comparison of four approaches. *Journal of Educational Measurement, 14,* 33–38.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist, 34,* 207–218.

Nitko, A. J. (1985). [Review of the book *A technology for test item writing*]. *Journal of Educational Measurement, 21,* 201–204.

Norris, S. P. (1990). Effects of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27,* 41–58.

Rodriguez, M. C. (2002). Choosing an item format. In G. A. Tindal and T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 211–229). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Roid, G. H., & Haladyna, T. M. (1982). *Toward a technology of test-item writing.* New York: Academic.

Ryan, J. M., & Demark, S. (2002). Variation in achievement scores related to gender, item format and content area tested. In G. A. Tindal and T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 67–88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–332). New York: American Council on Education and Macmillan.

Statman, S. (1988). Ask a clear question and get a clear answer: An enquiry into the question/answer and the sentence completion formats of multiple-choice items. *System, 16,* 367–376.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26,* 233–246.

Trelease, J. (1985). *The read-aloud handbook.* New York: Penguin.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

APPENDIX A
Textbooks Reviewed

Airasian, P. (1996). *Classroom assessment* (3rd ed., pp. 96–113). New York: McGraw-Hill.

Carey, L. (1994). *Measuring and evaluating school learning* (2nd ed., pp. 113–156). Boston: Allyn & Bacon.

Chase, C. I. (1999). *Contemporary assessment for educators* (pp. 113–129, 131–145). New York: Longman.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed., pp. 133–153, 154–178, 179–187). Englewood Cliffs, NJ: Prentice Hall.

Gallagher, J. D. (1998). *Classroom assessment for teachers* (pp. 155–205). Upper Saddle River, NJ: Prentice Hall.

Gredler, M. E. (1999). *Classroom assessment and learning* (pp. 78–97). New York: Longman.

Hanna, G. S. (1993). *Better teaching through better measurement* (pp. 134–190). Fort Worth, TX: Harcourt Brace Jovanovich.

Hopkins, K. (1998). *Educational and psychological measurement* (8th ed., pp. 213–253). Needham Heights, MA: Allyn & Bacon.

Hopkins, K., & Antes, R. L. (1990). *Classroom measurement and evaluation* (pp. 174–219). Itasca, IL: Peacock.

Kubiszyn, T., & Borich, G. (1996). *Educational testing and measurement* (6th ed., pp. 81–110). Glenview, IL: Scott, Foresman.

Linn, R. L., & Gronlund, N. (2000). *Measurement and assessment in teaching* (8th ed., pp. 151, 156–168, 169–192, 193–216, 217–234). Columbus, OH: Merrill.
McDaniel, E. (1994). *Understanding educational measurement* (pp. 138–150). Madison, WI: Brown & Benchmark.
McMillan, J. H. (2001). *Classroom assessment* (pp. 139–156, 176–181). Boston: Allyn & Bacon.
Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed., pp. 106–150). Fort Worth, TX: Harcourt Brace Jovanovich.
Nitko, A. J. (2001). *Educational assessment of students* (3rd ed., pp. 135–143, 144–183). Columbus, OH: Merrill.
Oosterhof, A. (1994). *Classroom applications of educational measurement* (2nd ed., pp. 95–108, 127–174). New York: Merrill.
Oosterhof, A. (1999). *Developing and using classroom assessments* (2nd ed., pp. 85–120). Upper Saddle River, NJ: Merrill/Prentice Hall.
Payne, D. A. (1992). *Measuring and evaluating educational outcomes* (pp. 120–135, 136–171). New York: Merrill.
Payne, D. A. (1997). *Applied educational assessments* (pp. 154–197). Belmont, CA: Wadsworth.
Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd ed., pp. 111–134). Needham Heights, MA: Allyn & Bacon.
Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed., pp. 83–115). Belmont, CA: Wadsworth.
Stiggins, R. J. (2001). *Student-centered classroom assessment* (3rd ed., pp. 117–154). Upper Saddle River, NJ: Merrill.
Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed., pp. 443–470). Columbus, OH: Merrill.
Tindal, G. A., & Marston, D. B. (1990). *Classroom-based assessment: Evaluating instructional outcomes* (pp. 52–62). Columbus, OH: Merrill.
Ward, A. W., & Ward, M. M. (1999). *Assessment in the classroom* (pp. 115–144). Belmont, CA: Wadsworth.
Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed., pp. 41–61). Boston: Allyn & Bacon.
Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1998). *Measurement and evaluation in the schools* (2nd ed., pp.209–244). New York: Longman.

APPENDIX B
References on Item-Writing, 1990–Present

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.
Andrés, A. M., & del Castillo, J. D. (1990). Multiple-choice tests: Power, length, and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology, 43*, 57–71.
Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55,* 959–966.
Cizek, G. J., & Rachor, R. E. (1995, April). *Nonfunctioning options: A closer look.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
Crehan, K. D., & Haladyna, T. M. (1991). The validity of two item-writing rules. *Journal of Experimental Education, 59*, 183–192.

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement, 53,* 241–247.

Downing, S. M., Dawson-Saunders, B., Case, S. M., & Powell, R. D. (1991, April). *The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II item characteristics.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Eisley, M. E. (1990). *The effect of sentence form and problem scope in multiple-choice item stems on indices of test and item quality.* Unpublished doctoral dissertation, Brigham Young University, Provo, UT.

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4,* 115–124.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items. *Evaluation and the Health Professions, 17,* 123–126.

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

Harasym, P. H., Doran, M. L., Brant, R., & Lorscheider, F. L. (1993). Negation in stems of single-response multiple-choice items. *Evaluation and the Health Professions, 16,* 342–357.

Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation and the Health Professions, 21,* 120–133.

Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation and the Health Professions, 15,* 198–220.

Huntley, R. M., & Welch, C. J. (1993, April). *Numerical answer options: Logical or random order?* Paper presented at the annual of meeting of the American Educational Research Association, Atlanta, GA.

Kolstad, R. K., & Kolstad, R. A. (1991). The option "none of these" improves multiple-choice test items. *Journal of Dental Education, 55*, 161–163.

Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement, 53,* 771–778.

McMorris, R. F., Boothroyd, R. A., & Pietrangelo, D. J. (1997). Humor in educational testing: A review and discussion. *Applied Measurement in Education, 10,* 269–297.

Nnodim, J. O. (1992). Multiple-choice testing in anatomy. *Medical Education, 26*, 301–309.

Rachor, R. E., & Gray, G. T. (1996, April). *Must all stems be green? A study of two guidelines for writing multiple choice stems.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Rich, C. E., & Johanson, G. A. (1990, April). *An item-level analysis of "none of the above."* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Roberts, D. M. (1993). An empirical study on the nature of trick test questions. *Journal of Educational Measurement, 30,* 331–344.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59,* 234–247.

Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology, 47,* 829–835.

Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation, 19,* 311–325.

Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement, 51,* 829–837.

Trevisan, M. S., Sax, G., & Michael, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54,* 86–91.