Item-Writing Rules: Collective Wisdom

Bruce B. Frey
Stephanie Petersen
Lisa M. Edwards
Jennifer Teramoto Pedrotti
Vicki Peyton

Department of Psychology and Research in Education
School of Education
University of Kansas
Lawrence, KS

Running Head: Item-Writing Rules

Word Count:
2195

Contact: Bruce Frey bfrey@ku.edu, 785-864-9706
1122 West Campus Road, Room 643
Department of Psychology and Research in Education
Lawrence, KS 66045

Abstract

In student assessment, teachers place the greatest weight on tests they have constructed themselves and have an equally great interest in the quality of those tests. To increase the validity of teacher-made tests, many item-writing rules-of-thumb are available in the literature, but few rules have been tested experimentally. In light of the paucity of empirical studies, the validity of any given guideline might best be established by relying on experts. This study analyzed twenty classroom assessment textbooks to identify a consensus list of item-writing rules. Forty rules for which there was agreement among textbook authors are presented. The rules address four different validity concerns- potentially confusing wording or ambiguous requirements, the problem of guessing, test-taking efficiency, and controlling for testwiseness.

Item-Writing Rules: Collective Wisdom

Classroom assessment is an integral part of teaching (Chase, 1999; Popham, 2002; Trice, 2000; Ward & Murray-Ward, 1999) and may take more than one-third of a teacher's professional time (Stiggins, 1991), yet there are few research-based rules to guide teachers in this activity. Teachers of classroom assessment must rely on advice, opinion, experience, and common sense to direct their students in constructing classroom tests that produce reliable and valid scores. In the absence of empirical research, what rules can educational researchers provide for those who produce classroom assessments? The purpose of this study was to analyze 20 popular classroom assessment texts to identify, through group consensus, the recommended practices (or rules-of-thumb) for writing paper-and-pencil objectively-scored classroom assessments. Additionally, recommended practices consistent with the few empirically-based research studies that do exist were identified.

Review of the Literature

Most classroom assessment involves tests that teachers have constructed themselves. It is estimated that 54 teacher-made tests are used in a typical classroom per year (Marso & Pigge, 1988) which results in perhaps billions of unique assessments, yearly, world-wide (Worthen, Borg, & White, 1993). Regardless of the exact frequency, teachers regularly use tests they have constructed themselves (Boothroyd, McMorris, & Pruzek , 1992; Marso & Pigge, 1988; Williams, 1991). Further, teachers place more weight on their own tests in determining grades and student progress, than they do on assessments designed by others, or on other data sources (Boothroyd, et al., 1992; Fennessey, 1982; Stiggins & Bridgeford, 1985; Williams, 1991). Many teachers believe that they need strong measurement skills (Wise, Lukin & Roos, 1991), and report that they are confident in their ability to produce valid and reliable tests (Oescher & Kirby,

1990; Wise, et al., 1991). Other teachers, however, report a level of discomfort with the quality of their own tests (Stiggins & Bridgeford, 1985) or believe that their training was inadequate (Wise, et al.). Indeed, most state certification systems and half of all teacher education programs have no assessment course requirement or even an explicit requirement that teachers have received training in assessment (Boothroyd, et al.; Stiggins, 1991; Trice, 2000; Wise, et al.). In addition, teachers have historically received little or no training or support after certification (Herman & Dorr-Bremme, 1984). The formal assessment training teachers do receive often focuses on large-scale test administration and standardized test score interpretation, rather than on the test construction strategies or item-writing rules that teachers need (Stiggins, 1991; Stiggins & Bridgeford, 1985).

A quality teacher-made test should follow valid item-writing rules, but as many researchers point out, empirical studies establishing the validity of item-writing rules are in short supply and often inconclusive, and, “item writing-rules are based primarily on common sense and the conventional wisdom of test experts” (Millman & Greene, 1993; p. 353). Even after decades of psychometric theory and research, Cronbach (1970) bemoaned the almost complete lack of scholarly attention paid to achievement test items. Twenty years after Cronbach’s warning, Haladyna and Downing (1989a) reasserted this claim, stating that the body of knowledge about multiple-choice item writing was still quite limited and added recently that “item writing is still largely a creative act” (Haladyna, Downing & Rodriguez, 2002, p. 329). The current empirical research literature for item-writing rules-of-thumb is most often of two kinds: (a) studies which look at the relationship between a given item format and either test performance or the psychometric properties of the test; and (b) studies which have evaluated the quality of teacher-made tests by applying some set of item-writing standards or criteria.

Reviewing these studies for an agreed upon list of classroom assessment rules, however, is not overly fruitful, as few rules present themselves.

Haladyna and Downing (1989a; 1989b) and Haladyna, Downing & Rodriguez (2002) have cataloged guidelines for multiple-choice, matching and alternate-choice (e.g. true-false) items with at least some evidence of validity by examining textbook endorsement and empirical studies. Though the authors did find empirical support for general advice such as "avoid trick items" and many studies testing particular rules, only four specific rules on their final revised inventory were supported without contradiction across studies and two of those were supported by the existence of only one study. Our search of additional recent literature (1989 to present) found little beyond Haladyna's, et al, exhaustive review (2002) and focused on the same few empirically validated rules (Klein & Klein, 1998; Knowles & Welch, 1992).

Several studies have evaluated the quality of teacher-made tests by applying test construction standards. Fleming and Chambers (1983), Marso & Pigge (1988, 1989) and Oescher & Kirby (1990) analyzed teacher-made tests for violations of item-writing rules. Among these studies, it was consistently found that the large majority of teacher-made tests had a sizeable number of flaws. By inference, it is clear that these studies applied item-writing and test format conventions as the standard against which quality was judged, but, for the most part, it is not clear what rules were chosen as standards and how those rules were derived. Consequently, it is difficult to produce a list of classroom assessment rules from these studies. In light of little data-driven guidance, we chose to distill the collective wisdom of the field of classroom assessment, by reviewing the aggregate knowledge of experts through analysis of classroom assessment textbooks, with the goal of establishing a list of valid rules for writing objectively-scored items.

Methods

For this study, 20 educational assessment textbooks and standard reference works were reviewed to identify a list of accepted, conventional rules for item construction and test formatting. Within this group, 15 were textbooks produced specifically for classroom assessment training and teacher preparation (Airasian, 2001; Bloom, Hastings & Madaus, 1971; Cangelosi, 2000; Case & Swanson, 1996; Chase, 1999; Gronlund, 1998; Johnson & Johnson, 2002; Kubiszyn & Borich, 2000; Kuhs, Johnson, Agruso & Monrad, 2001; Oosterhof, 1994; Phye, 1997; Popham, 2002; Stiggins, 2001; Trice, 2000; Ward & Murray-Ward, 1999) while the remaining five (Aiken, 1998; Friedenberg, 1995; Millman & Greene, 1993; Popham, 2000; Sax, 1997) were texts or reference works which cover the broader field of testing and educational measurement but include specific advice for constructing achievement test items. Each text was reviewed by one of the authors of this study to identify guidelines, rules, and rules-of-thumb concerning test construction. Different texts, of course, often described essentially the same rule but with different phrasing, and the authors worked as a group to reach agreement on whether differently worded rules were conceptually the same rule. Where disagreement as to conceptual similarity remained, the first author made the classification decision. Only rules concerning objectively scored paper-and-pencil testing formats were chosen for summary, which provided guidelines for four different item formats: multiple-choice, matching, true-false, and completion (or “fill-in-the-blank”) items. While multiple-choice items may occasionally appear in a completion format, the completion item format was defined for this study as non-multiple-choice items which require supplying a very short, objectively-scored answer. To identify the relative importance of each rule, as measured by the frequency with which measurement experts chose to advocate a rule, a list of all rules was compiled and ranked by the number of sources presenting each rule.

Results

Table 1 presents a list of the most commonly found item-writing rules. Rules found in only one source are not included in the table. In addition to listing the rules and indicating the item format to which it applies, the table also indicates which of the rules has received research support. We used the reviews appearing in Haladyna & Downing (1989b) and Haladyna, Downing & Rodriguez (2002) as our sources for this designation.

< **Insert Table 1 about here** >

Discussion

Though there were 40 different item-writing rules identified in this search, the rationales for each rule seem to fall into one or more of a few categories, and all reflect the over-riding concern for the validity of the item responses. The most basic validity concern is addressed by *5. Items should cover important concepts and objectives.* Other rules addressing basic validity concerns can be grouped into four specific areas- potentially confusing wording or ambiguous requirements, guessing, rules addressing test-taking efficiency, and rules designed to control for testwiseness.

Potentially Confusing Wording or Ambiguous Requirements

If some respondents understand a question or a set of instructions, and others do not, their responses may vary as a result of that difference, not as a result of different underlying levels of knowledge or skill. Rules proscribing clarity include *1. "All of the Above" should not be an answer option, 2. "None of the Above" should not be an answer option,* (Rules 1 and 2 are placed in this category, though some textbook authors appear to suggest them for reasons having

to do with controlling for testwiseness), *6. Negative wording should not be used, 7. Answer options should include only one correct answer, 11. Stems must be unambiguous and clearly state the problem, 14. Items should use appropriate vocabulary, 15. In fill-in-the-blank items, a single blank should be used, at the end, 19. True-false items should have simple structure, 20. True-false items should be entirely true or entirely false, 25. Matching item directions should include basis for match, 27. Directions should be included, 29. Vague frequency terms (e.g. often, usually) should not be used, 30. Multiple-choice stems should be complete sentences, 37. Complex item formats ("a and b, but not c") should not be used.*

Guessing

If respondents choose a correct answer by chance, instead of knowing the correct answer, there is no validity in that correct response. Some item-writing rules are designed to decrease the chance of guessing correctly by encouraging as many answer options as is reasonable. There are too many answer options if some answer options are so unappealing as not to function as distractors or the test becomes too long for practicality. Rules designed to increase the number of functioning answer options include *3. All answer options should be plausible, 17. In matching, there should be more answer options than stems, 21. There should be 3 to 5 answer options, 34. In matching, answer options should be available more than once, 35. Number of answer options should be < 7 for elementary age tests (in matching),* and *36. Number of answer options should be <17 for secondary age tests (in matching).*

Rules Addressing Test-taking Efficiency

A large set of item-writing rules are designed to make the test-taking process as simple, brief, and free from distraction as possible. These rules all deal with formatting options and include *4. Order of answer options should be logical or vary, 13. Answer options should not be*

longer than the stem, 18. All parts of an item or exercise should appear on the same page, 22. Answer options should not have repetitive wording, 23. Point value of items should be presented, 28. Questions using the same format should be together, 33. Individual items should be short, 38. All items should be numbered, 39. Test copies should be clear, readable and not hand-written, 40. Stems should be on the left, and answer options on the right.

Rules Designed to Control for Testwiseness

Perhaps it is a modern artifact of test construction, but many of the rules consistently recommended in the textbooks we surveyed exist as ways of counteracting testwise respondents with the ability to recognize patterns in answer options, identify unintentional clues, or use other skills unrelated to the level of knowledge or ability which is the intended target of a test. Because different respondents will have different levels of test-taking ability, validity concerns require that items be constructed in ways that prevent the use of these strategies. Rules with this goal include *4. Order of answer options should be logical or vary, 8. Answer options should all be grammatically consistent with stem, 9. Specific determiners (e.g. always, never) should not be used, 10. Answer options should be homogenous, 12 Correct answer options should not be the longest answer option, 16. Items should be independent of each other, 24. Stems and examples should not be directly from textbook. 26. Answer options should be logically independent of one another, 31. There should be an equal number of true and false statements, 32. True-false statements should be of equal length.*

Implications

Some researchers have found that teachers are confident in their test-making skills (Oescher & Kirby, 1990; Wise, et al., 1991), but studies suggest that perceived classroom assessment skill and actual skill are unrelated or even negatively correlated (Boothroyd, et al.,

1992; Marso & Pigge, 1989). Often, little training or resources are available for teachers, and many teachers feel they are not adequately prepared to produce quality classroom assessments. Even if teachers have gone through high quality classroom assessment training, there is an absence of consistent guidelines on the best way to write a test item, the most basic element of classroom assessment. To address this need for item-writing guidelines, we examined 20 classroom assessment textbooks to produce a consensual list of rules for item writing.

A similar approach to compiling rules was taken by Haladyna, Downing and Rodriguez (2002). Though their textbook sampling included only five of the texts sampled in our review, there is consistency with the present study's list of rules. Of the forty rules presented here, about half (19) were also endorsed by Haladyna and colleagues based on textbook citation, empirical studies or both. This represents substantial agreement, as that study's recommendations included all of the most frequently appearing rules in our review (Rules 1 through 12 on Table 1) and their review did not include rules for fill-in-the-blank items or rules specific to matching items.

In light of the paucity of empirical evidence, a theoretical approach may be the most valid path toward a list of item-writing rules for classroom assessment. We agree with Millman and Greene that, in measurement, some rules "make sense regardless of the outcome of empirical studies on the effect of violating that rule" (p. 353). The validity evidence for the majority of these rules would seem to remain limited to expert consensus, but they provide a solid basis for a consensus list of item-writing guidelines.

References

Aiken, L. R. (1998). *Tests and examinations.* New York: John Wiley & Sons, Inc.

Airasian, P. W. (2001). *Classroom assessment: concepts and applications*. Boston: McGraw-Hill.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook of formative and summative evaluation of student learning*. New York: McGraw-Hill.

Boothroyd, R. A., McMorris, R.F. & Pruzek, R.M. (1992). *What do teachers know about measurement and how did they find out?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA. ERIC Document Reproduction Service No. 351 309.

Cangelosi, J. S. (2000). *Assessment strategies for monitoring student learning*. New York: Addison Wesley Longman, Inc.

Case, S.M. & Swanson, D.B. (1996). *Constructing written test questions for the basic and clinical sciences*. Philadelphia: National Board of Medical Examiners.

Chase, C. I. (1999). *Contemporary assessment for educators*. New York: Addison-Wesley Educational Publishers Inc.

Cronbach, L.J. (1970). Review of *On theory of achievement test items* by J.R. Bormuth. *Psychometrika, 35*, 509-511

Fennessey, D. (1982). *Primary teachers' assessment practices: Some implications for teacher training.* Paper presented at the annual meeting of the South Pacific Association for Teacher Education, Frankston, Victoria, Australia. ERIC Document Reproduction Service No. 229 346.

Friedenberg, L. (1995). *Psychological testing.* Boston: Allyn & Bacon.

Gronlund, N. E. (1998). *Assessment of student achievement. 6th edition*. Boston: Allyn and Bacon.

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

Haladyna, T. M. & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51-78.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Herman, J. L., & Dorr-Bremme, D. W. (1984). *Teachers and testing: Implications from a national study. Draft.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. ERIC Document Reproduction Service No. 244 987.

Hoepfl, M.C. (1994). Developing and evaluating multiple choice tests. *Technology Teacher, 53* (7), 25-26.

Johnson, D. W., & Johnson, R. T. (2002). *Meaningful assessment: A manageable and cooperative process*. Boston: Allyn and Bacon.

Klein, S.P. & Klein, S.P. (1998). Standards for teacher tests. *Journal of Personnel Evaluation in Education, 12*(2), 123-138.

Knowles, S.L. & Welch, C.A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "None of the Above". *Educational and Psychological Measurement, 52*(3), 571-577.

Kubiszyn, T., & Borich, G. (2000). *Educational testing and measurement: Classroom application and practice*. New York: John Wiley & Sons, Inc.

Kuhs, T.M., Johnson, R.L., Agruso, S.A. & Monrad, D.M. (2001). *Put to the test: Tools and techniques for classroom assessment.* Portsmouth, NH: Heinemann.

Marso, R. N., & Pigge, F. L. (1988). *An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. ERIC Document Reproduction Service No. 298 174.

Marso, R. N., & Pigge, F. L. (1989). The status of classroom teachers' test construction proficiencies: Assessment by teachers, principals, and supervisors validated by analyses of actual teacher-made tests. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA. ERIC Document Reproduction Service No. 306 283.

Millman, J. & Greene, J. (1993). The specifications and development of tests of achievement and ability. In Linn, R.L. (ed). *Educational Measurement, 3rd Edition*. Phoenix, AZ: American Council on Education.

Oescher, J., & Kirby, P. C. (1990). *Assessing teacher-made tests in secondary math and science classrooms.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA. ERIC Document Reproduction Service No. 322 169.

Oosterhof, A. C. (1994). *Classroom applications of educational measurement. 2nd edition*. Columbus, OH: Merrill Publishing Company.

Phye, G. D (1997). *Handbook of classroom assessment: learning, adjustment, and achievement.* San Diego, CA: Academic Press.

Popham, W. J. (2000). *Modern educational measurement*. Boston, MA: Allyn & Bacon.

Popham, W. J. (2002). *Classroom assessment: What teachers need to know*. Boston: Allyn and Bacon.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation. 4th edition.* Belmont, CA: Wadsworth.

Stiggins, R.J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*, 7-12.

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). New Jersey: Prentice Hall.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*, 271-286.

Trice, A. D. (2000). *A handbook of classroom assessment*. New York: Addison Wesley Longman, Inc.

Ward, A., & Murray-Ward, M. (1999). *Assessment in the classroom*. Belmont, CA: Wadsworth Publishing Company.

Williams, J. M. (1991). *Writing quality teacher-made tests: A handbook for teachers.* ERIC Document Reproduction Service No. 349 726.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*, 37-42.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools*. White Plains, NY: Longman

Table 1

Item-Writing Rules Found in Twenty Classroom Assessment Texts

Item-Writing Rule	Frequency	%	Research Support*
1. "All of the Above" should not be an answer option.	16	80.0	X
2. "None of the Above" should not be an answer option.**	15	75.0	
3. All answer options should be plausible.	14	70.0	
4. Order of answer options should be logical or vary.	13	65.0	X
5. Items should cover important concepts and objectives.	12	60.0	
6. Negative wording should not be used.	11	55.0	
7. Answer options should include only one correct answer.	11	55.0	
8. Answer options should all be grammatically consistent with stem.	10	50.0	
9. Specific determiners (e.g. *always*, *never*) should not be used.	10	50.0	
10. Answer options should be homogenous.	10	50.0	
11. Stems must be unambiguous and clearly state the problem.	10	50.0	
12. Correct answer options should not be the longest answer option.	9	45.0	
13. Answer options should not be longer than the stem.	8	40.0	
14. Items should use appropriate vocabulary.	8	40.0	

Table 1 (Cont.)

Item-Writing Rule	Frequency	%	Research Support*
15. In fill-in-the-blank items, a single blank should be used, at the end.	8	40.0	
16. Items should be independent of each other.	8	40.0	
17. In matching, there should be more answer options than stems.	8	40.0	
18. All parts of an item or exercise should appear on the same page.	8	40.0	
19. True-false items should have simple structure.	6	30.0	
20. True-false items should be entirely true or entirely false.	6	30.0	
21. There should be 3 to 5 answer options.	6	30.0	X
22. Answer options should not have repetitive wording.	6	30.0	
23. Point value of items should be presented.	6	30.0	
24. Stems and examples should not be directly from textbook.	5	25.0	
25. Matching item directions should include basis for match.	5	25.0	
26. Answer options should be logically independent of one another.	5	25.0	
27. Directions should be included.	5	25.0	
28. Questions using the same format should be together.	5	25.0	
29. Vague frequency terms (e.g. *often*, *usually*) should not be used.	4	20.0	

Table 1 (Cont.)

Item-Writing Rule	Frequency	%	Research Support*
30. Multiple-choice stems should be complete sentences.	4	20.0	
31. There should be an equal number of true and false statements.	4	20.0	
32. True-false statements should be of equal length.	4	20.0	
33. Individual items should be short.	4	20.0	
34. In matching, answer options should be available more than once.	4	20.0	
35. Number of answer options should be < 7 for elementary age tests.	4	20.0	
36. Number of answer options should be <17 for secondary age tests.	4	20.0	
37. Complex item formats (“a and b, but not c”) should not be used.	3	15.0	X
38. All items should be numbered.	3	15.0	
39. Test copies should be clear, readable and not hand-written.	2	10.0	
40. Stems should be on the left, and answer options on the right.	2	10.0	

* Though studies were found by Haladyna, Downing & Rodriguez (2002) relevant to many of the rules in this table, the small number of studies concerning some rules and the lack of consistent findings providing persuasive empirical support was reported for only the four rules indicated. In some cases, fairly consistent evidence found that application of a rule, while not harmful, had no effect on a test’s psychometric properties. Support for Rule 21 is inferred from the finding that little is gained by adding additional answer options.

** Two textbooks (10%) supported the use of “None of the Above” as a way of increasing difficulty.