DOCUMENT RESUME

ED 395 037 TM 025 057

AUTHOR Buser, Karen
TITLE Basic Precepts in Test Construction.
PUB DATE 25 Jan 96
NOTE 16p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 25-27, 1996).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Decision Making; *Educational Planning; *Evaluation Methods; Scores; Student Evaluation; *Test Construction; Test Content; Test Format; Test Items; Test Reliability; Test Use; Test Validity

ABSTRACT

Most seasoned test developers recognize the importance of thoughtful decision making when constructing a test. Unfortunately, many classroom achievement tests are created by novice test developed who have not received sufficient instruction in item writing (G. Gulliksen, 1986; R. J. Stiggins, 1991). The result is often a test that is poorly constructed and scores that may not be reliable and valid for the purposes intended (Stiggins and N. J. Bridgeford, 1985). Three basic precepts in test construction are outlined: (1) The test developer must identify the purpose of the test; (2) The test developer must identify a plan for the test, reflecting substantive content and the cognitive processes necessary for completing the item task; and (3) The test developer must identify an appropriate format for the test. Adherence to these precepts will assist even beginning test developers to construct appropriate measures for evaluation of local instruction. (Contains 1 chart and 17 references.) (Author/SLD)

Basic Precepts in Test Construction

Karen Buser

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, January 25, 1996.

Abstract

Most seasoned test developers recognize the importance of thoughtful decision making when constructing a test. Unfortunately, many classroom achievement tests are created by novice test developers who have not received sufficient instruction in item writing (Gulliksen, 1986; Stiggins, 1991). The result is often a test that is poorly constructed and scores that may not be reliable and valid for the purposes intended (Stiggins & Bridgeford, 1985). The benefits of testing are directly affected by the test constructor's planful decisions regarding the purpose of the test, the plan for use of results, and the format of the testing measure. This paper attempts to outline three basic precepts in test construction. Adherence to these precepts will assist even beginning test developers to construct appropriate measures for evaluation of local instruction.

Basic Precepts In Test Construction

Most seasoned test developers recognize the importance of thoughtful decision making when constructing a test. Historically, tests in school classrooms have served the purposes of diagnosis, motivation, and measuring achievement (Wood, 1960). Unfortunately, many classroom achievement tests are created by novice test developers who have not received sufficient and practical instruction in item writing (Gulliksen, 1986; Stiggins, 1991). The end result is often a test that is poorly constructed and scores that may not be reliable and valid for the purposes intended (Stiggins & Bridgeford, 1985).

Problems of inappropriate testing practice have been an age-old issue for educators. As Ruch wrote in 1924,

> We are met with the situation today that large numbers of teachers and school officers are justly suspicious of the worth of the typical written examination, without possessing adequate knowledge of the technique for eliminating these faults and dangers. (p. 2)

Perhaps this is the reason that parents rated informal sources of information "...as more useful than standardized tests for learning about their 'child's progress in school'..." (Shepard & Bliem, 1995).

The present paper attempts to outline three basic precepts in test construction. Adherence to these precepts will assist even beginning test developers to construct more appropriate measures for evaluation of local instruction.

Precept Number One: *The test developer must identify the purpose of the test.*

Identifying the purpose of the test will drive other decisions concerning the construction of the test. The first decision of the test developer is to determine who will be tested with the measure. For example, a test designed to measure minimum competency within a population will be constructed differently than a test designed to select top applicants for a competitive program.

Secondly, a test developer must decide exactly what will be measured. Tests will vary according to their measurement of knowledge and behavior from cognitive and psychological domains. According to Crocker and Algina (1986), translating psychological constructs into specific test items has historically been a private, informal and largely undocumented process. These authors continued,

> Typically the test developer will conceptualize one or more types of behavior which are believed to manifest the construct and then simply try to "think up" items that require these behaviors to be demonstrated. (p. 67)

Principles for creating items which are representative of the construct being measured will be elaborated in the next section of this paper.

A final consideration for the purpose of the test is the determination of what is to be gained from the testing information. How will the results be used? A pre-test to define instructional gaps will look different than a post-test which assesses relative strength. A diagnostic measure will help an instructor see why students are making the kinds of mistakes they are making. The types of items and their construction will depend largely on the purposes for which the test results will be used.

Precept Number Two: *The test developer must identify a plan for the test.*

Once initial decisions have been made concerning the population for whom the test is intended, which constructs or behaviors will be tested, and what decisions will be made based on the test results, the test developer must formulate a plan for a test that will satisfy these purposes. Tests may be planned from test

blueprints, and/or tables of test specifications or item specifications. A thorough plan will assist the test developer to design a test with a balance of items in proportion to their importance in representing a construct. The plan should also reflect two important properties of items: substantive content, and the cognitive processes necessary to carry out the item task (Crocker & Algina, 1986, p. 72).

Much like a blueprint for an architectural structure, a test blueprint establishes a comprehensive and detailed set of plans for the construction of a test with all the correct components. A well constructed test blueprint will help instructors be certain of the following:

> 1) the content covered on the test is consistent with the content covered during instruction, and 2) that the level of cognitive skill that students need to answer questions on the test is consistent with what is intended. (Worthen, Borg, & White, 1993, p. 251)

Developing a blueprint involves two basic steps. First, the test developer lists the specific objectives to be measured by the test. Then the levels of higher order thinking are assigned to each objective. An abbreviated example of a test blueprint follows. (For fully developed examples, see Bloom, Hastings, & Madaus, 1971.)

Test Blueprint for a Unit of Instruction
on Double Digit Multiplication

Content Objectives	Knowledge	Compre-hension	Application	Total	Percentage
The student will correctly perform 3-digit addition with regrouping	2			2	10
The student will correctly solve problems involving pictorial groupings in multiplication problems	2			2	10
The student will correctly solve single digit multiplication problems		2		2	10
The student will correctly solve double digit numbers multiplied by one-digit multipliers		4		4	20
The student will correctly solve double digit numbers multiplied by double digit multipliers			6	6	30
The student will discriminate and correctly solve a double digit multiplication product from a word problem context.			4	4	20
TOTAL	4	6	10	20	
PERCENTAGE	20	30	50		100

After determining objectives and their cognitive requirements, the test developer can give priority or weight to the most important areas. Some educators (Worthen et al., 1993) recommended

developing the test blueprint before actual instruction occurs, to give instructors clear direction as to what concepts should be taught, and students information on the relative emphasis of content and skills.

Tables of test specifications and item specifications serve essentially the same purpose as test blueprints. Tables of specifications provide information to the test user as well as the test constructor, delineating objectives measured, item characteristics, and level of mastery (Sax, 1989; Schoer, 1970). Developing specifications for each item may also aid the test constructor in avoiding bias and redundancy in items (Kubiszyn & Borich, 1993) while maintaining accuracy in technical construction, grammar and readability (Crocker & Algina, 1986).

Precept Number Three: *The test developer must identify an appropriate format for the test.*

Locally developed measures can be an extremely important part of effective teaching. These measures can be tailored to the specific needs of a class, can be frequently administered, can give quick feedback to instructors, and can assist in identifying individual learner's needs (Worthen et al., 1993, p. 235). The careful planning of the test also requires thoughtful decisions concerning which format will yield the best match for the purposes intended. This section of the paper discusses a variety of test formats with a brief description of each.

True-False formats have historically been popular with local test developers, because of their ease of construction and scoring (Sax, 1989). Critics of true-false tests have maintained that such tests encourage rote learning, expose students to erroneous ideas, and are susceptible to inflated scores due to guessing (Worthen et al., 1993). Careful attention to the development of true-false tests may improve the application of such measures. For example, requiring students to correct a false item to make the item true, or employing a correction for guessing formula may yield more useful results.

Guidelines from Smith and Adams (1972) and Worthen et al. (1993) may assist the novice in writing quality true-false items. A good item should relate to a single idea, and avoid negative wording. Statements should be definitely true, or definitely false. There should be approximately the same number of true items as false items, and the items should be about the same length. The use of superlatives or "specific determiners" (Sax, 1989), which give unintentional clues, should be avoided.

Multiple Choice tests present many advantages as a testing format. Items may be constructed to measure cognition at varying levels of complexity. Compared to true-false tests, guessing effects on multiple choice tests are minimized. Using item analysis, a good multiple choice item may yield valuable information about student misunderstandings, item difficulty, and individual learner differences.

Contrary to popular wisdom among many novice test developers, Kubiszyn and Borich (1993) call good multiple choice items "the most time-consuming kind of objective test items to write" (p. 90). Scannell and Tracy (1975) and Worthen et al. (1993) offer technical advice for the construction of quality multiple choice items. These authors contend that all item alternatives should be plausible to those students who have not mastered the material. The best distractors will assist instructors in determining students' incorrect perceptions, and should be based on the most frequent errors made by students in related classwork. Wording of the distractors should be associated with wording in the item stem, with similarities in vocabulary, content and form. Multiple choice items should have three to five options, with the option completing the item stem statement. Kubiszyn and Borich (1993) add that good multiple choice items may include graphic or tabular material which must be interpreted in context of instruction, and require the student to apply learning to novel situations.

Matching exercises are basically multiple choice tests in which examinees associate options in one column with item stems in another column. Matching formats are frequently used to measure knowledge of factual events, dates, persons, etc. Given this format, novice test developers may have difficulty designing items that measure anything other than the lowest level of knowledge or memorization.

Sax (1989) offered suggestions for the development of matching tests. Options should be homogeneous in their nature and content. The tests should contain more options than item stems. Options should be arranged alphabetically or numerically, with the shorter responses in the second column. To extend the level of critical thinking, a test developer might match terminology with new examples of previously instructed content. One might also include novel pictorial material which must be interpreted and matched to the correct option.

Completion or short answer items comprise other common test formats. Not only can these tests be quickly constructed, but they require the examinee to supply the correct answer, thus eliminating the possibility of guessing. However, short answer tests may also take longer to score because students may supply alternative wordings, or long responses, in an attempt to cover the answer.

While short answer tests typically test only basic knowledge, such items can be constructed to yield valuable information. Kubiszyn and Borich (1993) offer the following guidelines for improving the quality of short answer items. Items should require a brief and definitive answer, with the completion occuring near the end of the item statement. Omit only key words from completion items, taking care not to distort the sense of the content. Avoid using verbatim quotes from the text; instead, use items that require application of knowledge previously instructed.

Essay test formats allow the opportunity to test higher cognitive skills than some of the formats mentioned previously. Essay tests are quick to assemble, and are appropriate for small groups of students. Criticism of the essay format mainly stems from the subjectivity required in scoring essay items. Essay responses are also criticized because they are time consuming to score and are subject to "bluffing" (Sax, 1989). In addition, essay answers are dependent on the student's ability to express thoughts clearly and consisely in writing.

Tuckman (1988) and Worthen et al. (1993) offer several recommendations for writing and scoring good essay items. These authors suggest questions which have a narrow focus, to prevent a broad interpretation of possible answers. Specific instruction concerning time limits and amount of information expected should be communicated. Questions should be directly stated and brief in nature.

Holistic scoring of essay items may be accomplished with the aid of a table of specifications for each item. In this way, weights can be assigned for each component of the expected answer. Reading every student's response to one question before moving on will allow for more consistent scoring. Keeping students' names and previously scored items out of sight may help eliminate bias in scoring other items. If possible, an instructor should reread papers, or have a peer read responses before assigning a final score (Sax, 1989).

Conclusion

Testing is valuable if the results contribute to better instruction and improved learning. The technology of the design of beneficial tests and test items continues to emerge with study and research (Roid & Haladyna, 1982). The benefits of testing are directly affected by the test constructor's planful decisions regarding the purpose of the test, the plan for use of results, and the format of the testing measure. Inappropriate decisions regarding testing practice may decrease student motivation, give incorrect information about student learning, and contribute to poor decisions concerning instructional effectiveness and educational practice (Nitko, 1989).

References

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

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Gulliksen, G. (1986). Perspective on educational measurement. Applied Psychological Measurement, 10, 109-113.

Kubiszyn, T., & Borich, G. (1993). Educational testing and measurement (4th ed.). New York: HarperCollins College Publishers.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In Linn, R. L. (Ed.), Educational measurement (3rd ed., pp. 447-474). London: Collier Macmillan.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. New York: Harcourt Brace Jovanovich.

Ruch, G. M. (1924). The improvement of the written examination. New York: Scott, Foresman & Company.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd. ed.). CA: Wadsworth.

Scannell, D. P., & Tracy, D. B. (1975). Testing and measurement in the classroom. Boston: Houghton Mifflin.

Schoer, L. A. (1970). Test construction: A programmed guide. Boston: Allyn and Bacon.

Shepard, L. A., & Bliem, C. L. (1995). Parents' thinking about standardized tests and performance assessments. Educational Researcher, 24 (8), 25-32.

Smith, F. M., & Adams, S. (1972). Educational measurement for the classroom teacher (2nd ed.). New York: Harper & Row.

Stiggins, R. J., (1991). Assessment literacy. Phi Delta Kappan, 72(7), 534-539.

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

Tuckman, B. W. (1988). Testing for teachers (2nd. ed.). New York: Harcourt Brace Jovanovich.

Wood, D. A. (1960). Test construction: Development and interpretation of achievement tests. Columbus, OH: Charles E. Merrill.

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