DOCUMENT RESUME

ED 328 908 CS 010 442

AUTHOR Powell, Janet L.; Gillespie, Cindy
TITLE Assessment: All Tests Are Not Created Equally.
PUB DATE Dec 90
NOTE 13p.; Paper presented at the Annual Meeting of the American Reading Forum (11th, Sarasota, FL, December 12-15, 1990).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Elementary Secondary Education; Essay Tests; Higher Education; Objective Tests; Reading; *Response Style (Tests); *Student Evaluation; Teacher Made Tests; *Test Construction; *Test Format; *Testing Problems; Test Items
IDENTIFIERS Test Appropriateness

ABSTRACT

Traditional tests fall into two categories, both of which have several advantages and disadvantages that need to be considered when determining the type of test to use. Constructed-response tests, such as essay tests, ask students to construct their own responses. Thus, students are required not only to recall but to organize and often apply knowledge. On the other hand, selected-response tests, such as multiple choice tests, ask students to select an answer between or among alternatives. While questions for constructed-response tests are relatively easy to prepare, they are much more difficult to grade and often contain relatively few questions. One of the advantages to constructed-response tests is that responses are less affected by guessing, and clues about students' thought processes can be provided. Selected-response tests require much more time to create, but scoring is much easier. One major advantage of these tests is for measuring knowledge of specific facts. Essay and written retellings are the most common of the constructed-response item types. Other types of constructed-response test are the cloze, completion, and short answer items. Special caution should be taken when using cloze tests to measure reading ability, since the reading act itself seems to be disrupted by cloze testing. Selected-response items include true/false or alternate response, matching, and multiple choice. While there are several basic problems and limitations surrounding all types of assessments, many problems can be attributed not just to the test itself, but to misuse of the test. (Twenty references are attached.) (RS)

ASSESSMENT: ALL TESTS ARE NOT CREATED EQUALLY

by

Janet L. Powell

California State University, San Marcos

and

Cindy Gillespie

Ball State University

ASSESSMENT: ALL TESTS ARE NOT CREATED EQUALLY

Testing continues to be an important, yet controversial topic to most educators. The public and private sectors are demanding more accountability through standardized testing from the public schools. At the same time, experts are still arguing about issues such as text bias, ambiguity, and even the very validity of tests. Standardized test usage continues to grow despite these debates. Teacher-made tests also remain an integral piece of the assessment of student's abilities in most classrooms.

While new and hopefully better assessments are being developed, the "traditional" forms of tests are still being used by a majority of classroom teachers. This article reviews some fundamental elements of traditional tests in an effort to clarify some of the issues surrounding them, so that teachers may select and create tests that are appropriate to their goals and the knowledge that they want to measure. First, an overview will be presented of the two major categories of tests. Next, specific types of test items will be examined in more detail.

Constructed-response vs. Selected-response Tests

Traditional tests fall into two major catagories. Both have several advantages and disadvantages that need to be considered when determining which type of test to use. Constructed-response tests, such as essay tests, ask individuals to *construct* their own responses. Thus, students are required not only to recall, but to organize and often apply knowledge. On the other hand,

selected-response tests, such as multiple choice tests, ask individuals to *select* an answer between or among alternatives.

There are many things to consider when choosing between constructed-response tests and selected-response tests. While questions for a constructed-response test are relatively easy to prepare, they are much more difficult to grade. A considerable amount of time must be spent in creating clear criteria, such as scoring rubrics, for assessing the answers. Likewise, scoring the tests takes considerable time. The scoring of constructed-response test items involves at least some subjectivity, even when criteria have been carefully established. Another disadvantage is that these tests contain relatively few questions, which in some cases prevents adequate sampling of the subject matter.

A cumulative listing from historic and contemporary test and measurement specialists (Ahmann & Glock, 1975; Cook, 1950; Cunningham, 1986; Ebel & Frisbie, 1986; Gronlund, 1982; Mehrens & Lehmann, 1984; Payne, 1974; Popham, 1978; Roid & Haladyna, 1982; Thorndike & Hagen, 1969; Wesman, 1971) suggests advantages and disadvantages to the constructed-response test items. The first advantage is that students do construct their own answers. Responses are less affected by guessing, and clues about students' thought processes can be provided. There is another important factor to consider which can be an advantage, or a disadvantage, depending on the purpose of giving the test. The scores given on constructed-response tests are directly related to how well the student can write, adding one

more factor into what is actually being measured.

Despite the complexities of scoring, the use of the constructed-response test is rising. Many feel that the advantages far outweigh the disadvantages. With the focus on process over product, and the push for more-and-more writing in the classroom, test developers are certain to continue the pursuit of refining and redesigning constructed-response tests.

There are also trade-offs when a selective-response test is used. These tests require much more time to create, but scoring them is relatively quick. Many people favor selective-response tests because they believe they are completely objective, but this may be erroneous. Many people favor selected-response tests on the assumption that they are totally objective. However, the scores on a selected-response test can also be considered as subjective since "right" and "wrong" answers are pre-determined by the test developer (Ebel, 1979, pp.100-101). This a weakness of selected-response that is often ignored.

Certainly one major advantage of the selected-response tests is for measuring knowledge of specific facts. Selected-response tests allow a broad sampling of subject matter in a highly-structured testing situation. The questions can be constructed to measure knowledge in any area. The scoring is simple, primarily objective, and reliable (Cunningham, 1986; Mehrens & Lehmann, 1984; Nunnally, 1967; Payne, 1974; Roid & Haladyna, 1982). However, this very advantage can also be considered a disadvantage. Many believe that these tests do not require much "real" thinking since there can only be one correct answer to questions. These critics believe such tests encourage little

more than rote memorization (Bracey, 1990; Haney & Madaus, 1989; Neill & Medina, 1989; Valencia & Pearson, 1987). However, when the objective of the assessment is to measure knowledge of facts, these tests can provide a relatively accurate assessment of such knowledge.

Thorndike and Hagen (1969, pp. 67-72) state there are theoretical issues to consider when choosing what to include in a test. One consideration deals with the adequacy of the test in eliciting student response. Choosing whether to develop a constructed-response test or a selected-response test should coincide with the purpose of the test. Popham (1978, pp. 44-45) states that for measuring knowledge of factual information, the selected-response test is more efficient. The selected- response test is also useful when a high degree of specificity is needed, such as tests designed to see if reteaching of facts is necessary. However, for measuring originality, the ability to synthesize ideas, write effectively, or to solve problems, constructed-response tests are obviously better.

Test Item Choice

Constructed-response Test Items

The types of items associated with constructed-response tests include essays, written retellings, cloze, completion, and short answer items.

Essay & Written Retellings

The most common of the constructed-response item types are the essay and written retellings. As can already be inferred, answering a well developed

essay question can require application of knowledge, and other forms of higher-level thinking, rather than simple recall. Therefore, essay tests, when written and scored with care, can provide some evidence of the student's ability to apply knowledge. However, a written retelling, though it requires construction of an answer like the essay, requires simple recall for the most part. Consequently, the differences between responses to critical essay questions and written retellings are enormous. It must be remembered that success on essay and written retelling tests in particular are tied to the student's ability to write. Again, this can be considered an advantage or a disadvantage, but it must always be remembered when interpreting the results of the tests.

Cloze (fill-in), Completion and Short Answer Tests.

Other types of constructed-response tests are the cloze, completion, and short answer items. While these tests do not rely as heavily on the student's ability to write as do the written retelling and essay, it still must be considered somewhat of a factor. The amount of information that is required to answer these types of questions can vary significantly. They can require little more than simple recall if not written with care.

A special word of caution is needed for using cloze tests to measure *reading* ability. Powell (1988) and Ashby-Davis (1985) agree that cloze tests require quite different thinking processes than other traditional forms of assessments. While taking a cloze test, students read slower and reread more often. Powell (1988) had students "think-aloud" as they completed reading tests. In · verbal protocols, the students did not tie in their background

knowledge to the passage during a cloze test as much as they did when taking multiple choice tests or giving retellings The student's attempts to understand the text appeared to be limited to the sentence level rather than the passage level. This research suggests that cloze tests may not be a valid measure of overall reading performance, since the reading act itself seems to be disrupted by cloze testing. However, cloze tests may be useful in determining a student's ability to use context clues.

Selected-response Test Items

The types of items associated with selected-response tests include true/false or alternate-response, matching and multiple choice.

True/False Items.

True/false items require the examinees to determine the truth or falsity of a statement. Advantages and disadvantages of true/false items have been cited by authorities in the field of test and measurement (Ahmann & Glock, 1975; Cook, 1950; Cunningham, 1986; Ebel & Frisbie, 1986; Mehrens & Lehmann, 1984; Payne, 1974; Roid & Haladyna, 1982; Swezey, 1981; Thorndike & Hagen, 1969; Wesman, 1971). Advantages of the true/false item include speed in scoring, ease of construction, inclusion of a larger number of items and measurement of factual knowledge. There are several disadvantages to true/false items. It is very difficult to write good true/false test items. For example, items about controversial material are difficult to write. There are also many instances where an answer is not unequivocally true or false; there are degrees of correctness. Finally, the fifty-fifty percent chance of getting a

question correct by guessing must be acknowledged when interpreting the scores.

Matching Items.

Matching items require students to match items placed in two or more columns. Historical and current literature (Ahmann & Block, 1975; Cook, 1950; Cunningham, 1986; Ebel & Frisbie, 1986; Mehrens & Lehmann, 1984; Payne, 1974; Popham, 1978; Roid & Haladyna, 1982; Swezey, 1981; Thomdike & Hagen, 1969; Wesman, 1971) cite the advantages and disadvantages of matching items. A matching format offers several advantages. Items are easy to construct and are more efficient than multiple-choice. Items are economical of space and time and are written in a compact form. Questions written as matching items are reasonably free from guessing. Disadvantages of matching items are that they are suitable for measuring association only, and they are susceptible to clues. Good matching items are also difficult to write.

Multiple-Choice Items.

Multiple-choice items require pupils to select a response from a specified number of options. Each multiple-choice item consists of two parts: the stem and suggested responses. Test and measurement authorities (Ahmann & Glock, 19785; Cook, 1950; Cunningham, 1986; Ebel & Frisbie, 1986; Mehrens & Lehmann, 1984; Payne, 1974; Popham, 1978; Roid & Haladyna, 1982; Swezey, 1981; Thomdike & Hagen, 1969; Wesman, 1971) state thst there are advantages and disadvantages of multiple-choice items. Multiple-choice items can be adopted to a wide variety of material and can

measure understanding, discrimination and judgment. They can be scored quickly and can provide diagnostic information if the response patterns are analyzed. One limitation of the multiple-choice item is that an extra amount of time and skill is required to construct good items. It is difficult to provide three or four plausible incorrect responses, and there is a tendency to write only recall questions.

Conclusions

While there are several basic problems and limitations surrounding all types of assessments, many of the problems surrounding them can be attributed not just to the test itself, but to the misuse of the test. For example, information about *process*, or how students came to certain conclusions, can only be inferred from all types of tests. In order to really understand where a student's thinking went wrong, one must literally ask the student to explain how they came up with an answer. Informal assessments such as this are extremely important to the overall assessment of all students.

We need to be more aware of what different types of tests measure, and the valid conclusions we can make from the test scores. Too often tests are used to measure something that cannot be measured by that test, and then make decisions about curriculum and placement based on invalid information. Tests in and of themselves cannot give educators all the answers. Literally *all tests* can only be considered as one sample of a student's ability, and must be considered along with other factors for a valid assessment of student progress. It would be difficult to find any educator who wouldn't agree that we must find

better assessment methods. Testing has not kept up with advances in educational theory. Portfolio assessment and authentic assessment are two of the ways that leaders in the field are making strides in improving assessment. However, as we are developing new ways to assess students, we must be mindful of how we use the ones we already have.

References

Ahmann, J., & Glock, M. (1975). Evaluating pupil growth (5th ed). Boston: Allyn and Bacon.

Ashby-Davis, C. (1985). "Cloze and comprehension: A qualitative analysis and critique." Journal of Reading, 28.

Bracey, G. W. (1990). "Teachers, thinking, and testing." Phi Delta Kappan, 71, 404-7.

Cook, W. (1950). Achievement testing. In W. Monroe (Ed.), Encyclopedia of Educational Research (pp. 1461-1477). New York: Macmillan.

Cunningham, G. (1986). Educational and psychological measurement. New York: Macmillan.

Ebel, R. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R., & Frisbie, D. (1986). Essentials of educational measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. (1982). Constructing achievement tests (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Haney, W. & George Madaus. (1989). "Searching for alternative to standardized tests: Whys, whats, and whithers." Phi Delta Kappan, 70 , 683-87.

Mehrens, W., & Lehmann, I. (1984). Measurement and evaluation in education and psychology (3rd ed.). New York: Holt, Rinehart & Winston.

Neill, D., & Medina, M. "Standardized testing: Harmful to educational health." Phi Delta Kappan, 70, 688-97.

Nunnally, J. (1967). Psychometric theory. New York: McGraw-Hill.

Payne, D. (1974). The assessment of learning. Lexington, MA: DC Heath.

Popham, W. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Powell, J. (1988). An examination of comprehension processes used by readers as they engage in different forms of assessment. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

Roid, G., & Haladyna, T. (1982). A technology for test item writing. New York: Academic.

Swezey, R. (1981). Individual performance assessment: An approach to criterion-referenced test development. Reston, VA: Reston Publishing.

Thorndike, R., & Hagen, E. (1969). Measurement and evaluation in psychology and education (3rd ed.). New York: J. Wiley and Sons.

Valencia, S., & Pearson, D. (1987). "Reading assessment: Time for a change." The Reading Teacher,40, 726-33.

Wesman, A. (1971). "Writing the test item." In R. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 81-130.