Maximizing Equity and Access in Test Construction

Carolyn Emrick Massad, Ph.D.

July 2003 (Revision 1, December 2003)
Maximizing Equity and Access in Test Construction

This paper was prepared for an invited presentation at the Gallaudet University National Conference on “High Stakes Testing: Are Deaf and Hard of Hearing Children Being Left Behind,” November 15–16, 2002. The paper is to appear in a book based on presentations at the Gallaudet Conference and is available here by permission of the publisher, Pearson Educational Measurement.

President Bush’s education agenda (The White House, 2001) reinforced what some states were already addressing in their policies aimed at improving education. With the signing of the No Child Left Behind Act of 2001 (NCLB) in January 2002 (Public Law 107–110), accountability provisions require that states set content standards and have assessments linked to those standards. Some states already had developed standards and had assessments in some content areas. However, in many cases, the standards and the means by which they will be measured are now being selected or developed. Even in those states that had standards, the standards are being reviewed and revised (Linn, 2000; Linn, Baker & Betebenner, 2001; Heubert, 2001 & 2003; Olson, 2002; Popham, 2002).

Content standards refers to what educators previously called curriculum goals or instructional objectives (Popham, 2002). A standard alone does not define the expected achievement. It is the measure of that standard that provides the definition of the standard. Today as educators work to select and create tests to measure standards that they have written, they are refining and defining the standards.

The intent underlying the development of standards-based measures is to provide information to support instruction and optimal student learning. The idea is that student achievement will be reflected in the assessment results. Yet this may or may not be the case since the assessment is merely one instance or way of finding out what students may know and are capable of doing. Further, it will not indicate whether or not the students have had access to the knowledge and skills for which they are to be tested. Nevertheless, the assessment results should be used to provide means to enable students to learn what they need to know to be successful and to facilitate access to that knowledge. Otherwise, there is little value in the assessment.

Because of the limitations of a single measure of achievement given at one point in time on the continuum of student learning, it is important that the measure be the best it can possibly be and that the diversity of the population be recognized in developing and/or selecting that measure so that equity and access are maximized. Specifically, the range of cognitive styles that impact on how people learn must be recognized. Also, the language of the testing environment needs to be made as accessible as possible, particularly to the
many students for whom English is not the first language or for whom English is not fully accessible and for whom it may not be the sole language for learning. Essentially, English or any auditorily-based language is never fully accessible to a person who is not fully hearing growing up. For example, hearing people have the advantage of unhampered access to English and thus have easier access to incidental learning. They are also likely to be more familiar with the multiple meanings of words and idiomatic lexicon and expressions, as well as with complex English constructions (Karchmer and Mitchell, 2002; Marschark, 2001; Mounty, 2002; Randall, McAnally, Rittenhouse, Russell, and Sorensen, 2000). These are things that must be kept in mind by test developers and those who select tests, particularly if the results of the assessment are to reflect what all students need to know and be able to do.

A prerequisite for developing and/or selecting measures of students’ knowledge and skills is having a complete understanding of the students to be tested and their learning experiences. For example, it is critical to know the age and academic grade level of the students, what their curriculum and learning experiences are like, as well as whether English was accessible to them growing up and/or the language of instruction. This information is essential to determine the level of the content and skills to be measured, as well as the language to be used in the measure (Martin, 2001; Massad, 1995; Mounty, 2001; Rosenfeld, Mounty and Ehringhaus, 1995). It is most important for developing and/or selecting measures that are appropriate for the students being tested and for providing the means by which all those being tested can demonstrate what they know.

There are many ways in which content and skills may be measured. However, there are excellent resources available to guide test developers and users. Some general guidelines for developing and using assessment measures are the focus of the Code of Fair Testing Practices in Éducation (1988) that was developed by the Joint Committee on Testing Practices of the American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education. The Code includes guidelines for developing and/or selecting tests, interpreting scores, and striving for fairness. The Joint Committee also produced the Joint Standards for Educational and Psychological Testing (1985, 1999), which provides the foundational guidance for all testing. In 2001, five national associations—the American Association of School Administrators, the National Association of Elementary School Principals, the National Association of Secondary School Principals, the National Education Association, and the National Middle School Association—convened the Commission on Instructionally Supportive Assessment to address the concern that federally-mandated and state-administered tests have instructional utility. The Commission produced Building Tests to Support Instruction and Accountability: A Guide for Policymakers (2001), containing nine requirements that must be met to ensure a responsible educational assessment system for the improvement of learning. Further, most test developers have their own guidelines for their test development work, including test and testing program specific guidelines.

Essentially, language is the primary medium of learning and testing for all students. To develop tests that will provide good measures of what deaf and hard of hearing students
know requires an understanding of the students’ communication processes and skills and an awareness of how English generally functions for them. Especially for the deaf and hard of hearing students, the way in which language is used in the classroom and in testing can create insurmountable hurdles and impede learning as well as prevent an accurate assessment of what they know and can do (Mounty, 2002; Marschark, 2001, Martin, 2001; Randall, McAnally, Rittenhouse, Russell, and Sorensen, 2000). Just as educators seek ways to maximize learning for all students in the classroom, test developers must seek ways to maximize all students’ ability to demonstrate what they know on tests.

**Guidelines for Test Construction**

What follows are general guidelines that would be appropriate for constructing test questions for any population but that have particular relevance for testing deaf and hard of hearing students.

Each type of test question is efficient for measuring some learning outcomes and inefficient or inappropriate for measuring others. The types of selected response questions include the (1) true-false or alternative response, (2) matching, and (3) multiple-choice.

The true-false or alternative response question is most useful where the desired outcomes are determining the truth or falsity of a statement, distinguishing between fact and opinion, or discriminating between appropriate and inappropriate responses, such as “yes/no,” “right/wrong,” “agree/disagree.” The true-false or alternative response question is generally inadequate for measuring the more complex learning outcomes.

The matching type question is similarly restricted and is limited almost entirely to learning outcomes that call for the identification of simple relationships and the ability to classify things into given categories.

The multiple-choice question is the most generally adaptable and can be used effectively to measure a variety of learning outcomes, from the simple to the complex. However, even the multiple-choice question is inefficient or inappropriate for measuring some of the more complex learning outcomes.

Types of constructed response questions include (1) the short answer and (2) the essay or other types of open-ended questions.

The short-answer type question, calling for a few words or sentences, effectively measures the recall of specific facts but is generally inappropriate for measuring understanding, application, interpretation, and other complex learning outcomes.

The essay or other types of open-ended questions are the most effective for measuring the ability to organize data, for presenting original ideas, and for problem solving. Furthermore, if ability to write is being measured, the constructed response is the most direct measure.
Additionally, there are other measures of learning outcomes, such as performance observations, interviews and in-depth observations, and long-term activities, such as projects, extended written assignments, and portfolios.

Regardless of the type of questions used in a measure, the fundamental principles in the construction of the questions include that they:

1. measure only educationally important content and skills and be based on a representative sample of the content and skills,
2. elicit only the outcomes that are to be measured, and
3. do not prevent or inhibit the students’ demonstration of the desired outcomes.

Since the standards-based measures consist primarily of multiple-choice and constructed-response questions, the remainder of this paper will focus on those two types of questions. Because the language of instruction is typically English, the standards-based measures are usually encoded in English. Therefore, attention also will be given to the importance of ensuring that the language of the measures does not prevent or inhibit students’ ability to demonstrate what they know and can do. To maximize equity and access requires careful examination of the questions and whether or not the English structure, usage, and vocabulary are appropriate for all students.

Consequently, in measuring learning outcomes other than reading skills, it is important also to keep in mind that although the language of instruction was English, the measure should not be primarily a measure of reading: the focus must be on the content and skills being measured. We want the student to be able to obtain correct answers to questions that measure the attainment of the learning outcome being tested. Our measure would not be considered valid if the student answered such questions incorrectly merely because of readability factors, such as inordinately complex sentence structures or exceptionally difficult vocabulary, or because the type of response called for was not clearly specified. These factors, which are extraneous to the purpose of the measure, limit and modify the student’s responses and prevent the student from demonstrating the true level of achievement attained. Such factors are as unfair as determining a person’s running ability when he or she has a sprained ankle. Although a measure of running would be obtained, it would be restricted by a factor not intended to be included in the measure.
Good measurement requires that those who develop or review and select the questions that are to be used in tests, such as the standards-based measures, must establish criteria for determining whether questions are appropriate and valid for the purpose of the test. To discuss the criteria, it would be helpful to consider the vocabulary of measurement. To this end, consider the following diagram of a multiple-choice question:

Stem------------------- What is the setting for this passage?
Correct Answer------
(A) In a Vietnamese village *
(B) On a Chinese farm
(C) Under a huge tree
(D) On a small river

Note that A, B, C, and D are the options for the stem. B, C, and D are distracters or incorrect options while A is the key or correct answer.

In addition, here is an example of a constructed-response question with its associated terminology:

Prompt---- Carefully read the passage above. Based on this information, describe in detail the three most common causes of indigestion. Use examples from the passage to support your answer. This is not a test of your writing ability, but your answer should be written as clearly as possible. You may use any one or combination of formats (e.g., lists, outlines, prose paragraphs, diagrams) that is appropriate for your response.

Now, let us turn to the criteria for determining whether questions are appropriate and valid for the purpose of the test. The first, second, and third criteria below would apply to both multiple-choice and constructed-response questions. The fourth criterion focuses only on multiple-choice questions, while the fifth criterion focuses only on constructed-response questions.

(1) The question should be about a concept, idea, or skill that is important for the examinee to know or to understand as opposed to some obscure or trivial fact. In addition, for questions measuring skills, such as reading, the question should be as interesting as possible to students at the targeted grade level.
(2) The question should be structured around one central idea or problem that is clearly and completely presented in the stem of the question. Those whose primary language is not English and those who have grown up without full accessibility to English understand a complete question better than an incomplete stem that is completed by the options.

- The stem should be meaningful by itself.

- The stem should include as much as is needed to answer the question and should be free of irrelevant material. When one is calling for a judgment or a conclusion, the criterion or authority that is the basis for the correct answer should be specified.

- The stem should not pose a form of the question “What do you think…” or “What would you do…” because the correctness of any of the options could then be defended.

- A negatively stated stem should be used only when significant learning outcomes require it. For example, knowing the “least important…” is seldom related to important learning outcomes.

(3) The language of the question is most important. It should be clear and concise.

- The vocabulary should be as simple as possible unless vocabulary that is complex or difficult is being tested.

- The stem and options, if there are options, should be simple, short and in straightforward context.
  
  - The information should be presented in the active voice and present tense whenever possible.

- The question should avoid complex structures that are unnecessary to measure what is intended.
  
  - Whenever possible, the subject and its verb should be together.

  - Sentences with multiple embedded clauses should be avoided, as they are particularly difficult for students who have not fully mastered English.

  - All parts of prepositional phrases should be kept together.

  - When using a pronoun, especially at the beginning of a sentence, its referent should be clear.
• The question should not contain double negatives. For example, it is inappropriate to ask the student to answer by identifying an option or writing about something that is not true or that is false or incorrect. As an extreme example, a student would be confused by having to recognize that it is not true that John Milton did not write Paradise Lost.

• When several questions are based on the same referent or on stimulus material such as a passage, graphs, or charts, each question should be independent of the other questions in the set. Moreover, it is imperative that the student be able to arrive at the correct answer from the stimulus material. For example, arriving at the correct answer should not be based on having answered correctly a previous question in the set. In addition, for sets of multiple-choice questions, distracters in one question should not provide clues for answering another question in the set.

(4) The options of multiple-choice questions are as important as the stem. The difficulty of a question depends largely on the options. The finer the distinctions that must be made to select the correct answer from the distracters, the more difficult the question.

• The correct answer to the question should be meaningful and related to important learning outcomes. Recognizing that no option is a correct response is seldom related to important learning outcomes; therefore, “none of the above” should not be an option.

• If the question is asking for only one correct answer, there should be only one correct answer in the options available. “All of the above” should never be used as an option as it is inconsistent with there being only one correct answer.

• The distracters should be selected and formulated with care—incorrectness should not be the sole criterion. They should be plausible to students at the targeted grade level. Nor should the distracters be tricky.

Sources of good distracters include the following:

- Common misconceptions and common errors in technique

- A statement which is itself true, but which does not satisfy the requirements of the problem

- An incorrect statement that may sound plausible to those who have not attained the desired learning outcome

• The options should be phrased so that they all maintain a similar relationship to the thought in the stem. The options should be stated clearly and concisely.
There should be no clues given in the correct answer, such as similarity between the stem and the correct answer in wording, phraseology, or grammatical construction. If this similarity is not maintained between the stem and other options, the student can respond based on the similarity alone. In addition, “text bookish” wording in the correct answer alone is another clue.

(5) Constructed-response questions provide a statement of the situation, setting, or problem along with response instructions.

- The question or prompt specifies to what the student is to respond. The greater the clarity and detail in the prompt of what is expected, the more likely the question is to measure what is intended to be measured.
  - The question should be as brief as clarity allows. Restatement may sometimes be necessary to prevent misunderstandings, but, then, perhaps the restatement alone may be all that is needed.
  - Questions should not be able to be answered simply by “yes” or “no.”

- There are basically two types of constructed-response questions:
  - restricted-response questions, which provide highly structured directions and are typically scored by a point method, using a model answer as a guide
  - extended-response questions, which do not describe a task too rigidly but indicate scoring criteria. This type of question is scored by the rating method, using defined criteria as a guide.

- Specific directions for responding may appear before or be included within the prompt. The greater the clarity and detail provided the student, the more likely the question is to measure what it is intended to measure. The response instructions include such things as:
  - minimum or maximum length of response, and
  - specific points to be covered and/or performance required, such as explaining a solution in addition to describing one, stating a position and providing a rationale for it, etc.

- Criteria for scoring, or rubrics, are an important part of the question. The criteria for scoring the question should accompany the question when constructed-response questions are reviewed, as they help to determine whether or not the question has been clearly and concisely stated. There are generally three types of criteria:
Content criteria, which deal only with the presence of knowledge and its accuracy and not with the use of knowledge for application, analysis, synthesis, or evaluation; responses are judged solely on whether the content of the response leads to the conclusion that the student is knowledgeable in the content area.

Organization criteria, which deal with whether the response is logically presented.

Process criteria, which deal with whether the process(es) required in the prompt and/or the response instructions were carried out in the response. For example, they may be concerned with the problem-solving steps of:

- defining the problem,
- generating alternative solutions,
- weighing the alternative solutions and choosing among them,
- implementing the choice,
- evaluating the outcome

or, they may be concerned with predicting what a character in a passage would do in a specified situation and supporting the prediction with details in the passage.

Not all types of criteria need to be included with each question. However, the criteria for a question should be clear to the student, whether implied or explicit. The student should not be asked to read the mind of the person who wrote the question.

- If the question is to measure content knowledge and skills other than writing, the response should not be evaluated on writing skills. If the response is to be scored on writing skills as well as scored on content and/or other skills, this should be made clear in the specific directions for responding, and the two scores should be kept separate to distinguish what is measured by each.

- If all students are to be compared based on the assessment of a learning outcome measured by constructed-response questions, optional questions or prompts should not be offered as alternatives from which students may choose. Alternative prompts eliminate the ability to make comparisons.

When a test consists solely of constructed-response questions, there should be a larger number of questions that require relatively short answers rather than a few questions requiring long answers. Such an approach provides a broader sample of the students' knowledge and skills.
• The time limit, if any, and point value of each constructed-response question should be specified whether they are in a test consisting of mixed types of questions or in one composed solely of constructed-response questions.

In conclusion, it is to be hoped that the standards-based assessments will be developed and used appropriately for the purpose for which they are intended: to support instruction and optimal student learning.

References


