This article describes some of the problems of criterion-referenced tests

Criterion-referenced reading test: Stop, look and listen

by Leo M. Schell

Ten years ago hardly any educators knew what a criterion-referenced test (CRT) was; today there are dozens of commercial ones and hundreds of teacher-made ones. But the problem is that there has been little discussion within the reading community of the pros and cons of these tests. Indeed, James Popham of UCLA, one of the original and most ardent proponents of criterion-referenced tests, has become so disenchanted with the quality of some of the tests he so strongly favors that he recently lamented that some of these tests "are less fit for schools than they are for paper shredders." (6)

Educators should not be cynics, skeptics nor "againers" of something new. But they should be knowledgeable, evaluative, cautious and professional. They need to avoid the poorest of these tests and exercise great caution in constructing their own. Thus, this article describes some of the common problems of many CRTs and suggests some guidelines by which they may be appraised.

CRTs—Part of a System

CRTs are intended to be an integral part of an instructional system. Given as pretests, they indicate which students need which skills. Given as post-tests, they indicate who learned how much of what was taught and indirectly prescribe future instruction. In fact, some CRTs are integral parts of instructional systems that provide materials and recommendations for such instruction.

This system seems based on four fundamental assumptions:

1. Reading can be divided into small, discrete entities.
2. These entities can be written as objectives.
3. These objectives can be measured via specially constructed test items.
4. Standards for mastery can be set.

These assumptions are extraordinarily important because they depart to some degree from common instructional and testing beliefs of the past—and even many current ones. For one thing, they define to a great degree what this thing called reading is and how its achievement and growth should be measured. One problem is that not everybody can agree with one or more of these assumptions. Psycholinguists such as Kenneth Goodman (3) or Frank Smith (6) might easily reject the first assumption. Educators who agree with the psycholinguistic point of view would have a difficult time accepting the first premise upon which CRTs are based.

Some measurement specialists (as will be explained later) may disagree substantially with the fourth assumption, arguing that the problems of setting standards is so complex, so fraught with unresolved problems, that the assumption is actually dangerous and that tests based on that assumption should be labeled "Potentially hazardous." Therefore, these assumptions need to be examined carefully by educators and not taken lightly.

Validity

Whether CRTs measure what they say they measure should not be a problem since there is supposed to be a close correspondence between test items and corresponding objectives. This is called content validity which is judgmental and logical. A person should be able to inspect an objective and its corresponding test item(s) and decide with a reasonable degree of confidence whether the item generally measures its objective.

However, the objectives for numerous CRTs are unavailable. Not only does this violate one of the assumptions on which CRTs are based but it makes it difficult if not impossible to determine the validity of the test, to know how well a test item measures its objective. Without objectives, few of us are capable of determining a test's validity, and, therefore, we remain ignorant. Ignorance may be blissful but it's also unprofessional and potentially dangerous since we will or will not assign instruction to children on the basis of test results. Invalid tests give potentially invalid test results which in turn may lead to either unneeded instruction—or even lack of needed instruction. Validity is not irrelevant.

McClung (4) points out that CRTs should have **instructional validity,** a variation of curricular validity. He argues that there must be some way of knowing whether or not the stated objectives were actually taught in the classroom. He states that instructional validity should be a central concern to educators because if test items are not representative of the instruction then test results—and subsequent use of them—will be inappropriate. Instructional validity could be particularly troublesome with CRTs that are independent of the instructional program, e.g., a commercial CRT from one publisher used with a basal reader program from another publisher. In such cases, the test could easily measure something that wasn't taught or not measure something important that was. Thus a rigorous comparison of the test, curriculum and instruction is crucial.

Another aspect of validity is that some tests include mislabeled items. One subtest of critical reading requires that statements be numbered as to their order of occurrence. To this author's knowledge, sequence of events is not mentioned by any reading authority as a skill in critical reading. Can we assume that a child doing well on this test is really a good critical reader? Another example

of questionable validity is found on a widely used phonics subtest which claims to measure sound-letter associations. The audio tape says both the stimulus word, e.g., **put,** and several response words, e.g., **pet, gate, pony.** The examinee is to choose which of these response words ends with the same sound and letter as the stimulus word. But since the stimulus word is shown **in print,** it seems as if this test merely measures the ability to match final letters rather than the ability to associate a sound with its corresponding letter. What does a child really know who does well on this test? And can we validly assume that children doing poorly on it need sound-letter instruction?

Another example of questionable validity is found in one CRT from one of education's largest publishers which claims to measure over 15 separate comprehension skills, e.g., Equivalent Sentences, Main Idea: Unstated, Author's Purpose, etc. For over 35 years we've known that current testing procedures are inadequate to validly divide comprehension into more than 2-3 categories. Drahozal and Hanna (1) report on the latest such failure. Are all these subtests really measuring what their title says they are? If they are, they are valid and we can have some degree of confidence in them. But if not, they are invalid to some unknown degree and our confidence in them is diminished to the same degree. We are not interested in validity merely for its own sake; we are interested in it because the test results direct our subsequent instruction, they determine who will receive further teaching and who won't. This requires valid, not questionable, information.

Another aspect of validity is how an objective is measured. One test measures the characteristics of a given literary form by having the examinee write **myth, legend, fairy tale,** or **tall tale** by a definition such as "This type of story takes place in a 'never-never land' and often features fairies." Another test measures the same general objective by asking the test taker to read a passage typical of a kind of literature and asks the examinee to select which of four genres it is probably from. Are both items equally valid to appraise the same objective? They claim to be. I doubt it.

Numerous other examples could also be cited of tests and test items whose validity should be questioned or challenged. Educators should select only those tests whose items best mirror the objective being measured; they should be skeptical of any which are questionable.

Reliability

Conventional procedures for determining reliability are not appropriate for nor applicable to mastery CRTs. These procedures require variability in scores, a range of scores so it can be seen whether the low scores are consistently low and the high scores consistently high. But most CRTs are deliberately constructed to produce low variability because typically 80 percent or more of the examinees are expected to answer nearly all the items correctly. But even though traditional reliability assessment methods are inappropriate for indicating the reliability of most CRTs, we do know some general things about what makes a test reliable.

One is test length or the number of items measuring an objective. The longer a test or the more items that measure an objective, the more reliable the test tends to be. Yet many commercially published CRTs that I examined used only two items to measure an objective and several used only one. In multiple-choice tests where guessing is possible, so few items as this may not unequivocally indicate whether or not an examinee possesses the stated competence. Popham (5) states that it is "technically impossible to get a decent fix on an examinee's status with respect to a particular skill by using only a handful of items." Furthermore, he warns that in situations where the stakes are high "such as when a student's graduation from high school hinges on mastering the skills represented by a test, then attempting to squeeze by with a paucity of items is both professionally and ethically irresponsible."

Related to the number of items is the matter of guessing. Some tests use only three responses, which gives a 33⅓ percent chance of getting the answer correct by guessing. And several I examined provide only two responses, thereby giving the examinee a 50 percent chance of guessing the right answer. Did the student **know** an answer or did he/she **guess** it? This is what reliability data helps us determine. In the absence of such numerical information, educators wishing to select the best CRT need to determine how many items measure each objective and what the examinees' chances of guessing the right answer are.

Cut-Off Scores

Cut-off scores are probably the single most perplexing, troublesome and unresolved aspect of CRTs. A fundamental concept of CRTS is that a standard is set and if the examinee meets or exceeds it, then we can assume he/she probably needs no more instruction at this time in that skill. How standards are set is therefore of unparalleled importance.

The interested educator searches test manuals in vain for an answer, for a rationale for the standards. Was it a consensus of experts or the arbitrary judgment of one person? How does anyone **know** that correctly answering 70 percent of the items on a test indicates proficiency, competency or mastery? Glass (2) has written compellingly and movingly on this topic. He concludes, "I have examined a half dozen classes of methods for establishing mastery levels, standards or cut-off scores; each has proved to yield arbitrary and potentially dangerous results."

This is an enormously complicated topic but one of extraordinary cruciality. If the cut-off score is too easy, students will be passed who would merit from further instruction; yet if the standard is too difficult, students who shouldn't be will be given unnecessary instruction. Educators should be wary of tests that provide no information on how standards were set and which imply "Trust me." Popham (5) says that one characteristic of a well-constructed CRT is "the availabilty of normative data that will permit educators to answer more sensibly the question: 'How good is good enough?' " Currently, hardly any commercial CRTs provide such data and obviously it is not available for the superabundance of teacher-constructed ones that fill reading "methods" textbooks and others for which advertisements flood our daily mail.

Conclusion

This article in no way is an attempt to halt the current move toward using more and more criterion-referenced tests in reading instruction. Properly constructed CRTs can definitely help teachers improve both their teaching and children's learning. But we should be aware that merely because a measuring device is labeled "criterion-referenced" does not make it an adequate or worthwhile test. Consumer advocates have recently begun to demand that canned foods plainly state in writing what the contents inside the can are so that potential buyers will have more to rely on than the enticing photo on the can's label. Educators wanting the best for their students would be well advised to look for and demand precisely the same thing from tests labeled "criterion-referenced."

References

1. Drahozal, Edward C. and Gerald Hanna. "Reading Comprehension Subscores: Pretty Bottles for Ordinary Wine." **Journal of Reading,** 21 (Feb. 1978), 416-420.
2. Glass, Gene V. "Minimum Competence and Incompetence in Florida." **Phi Delta Kappan,** 60 (May 1978), 602-605.
3. Goodman, Kenneth, ed. **The Psycholinguistic Nature of the Reading Process.** Detroit, Mich.: Wayne State University Press, 1968.
4. McClung, Merle S. "Are Competency Testing Programs Fair? Legal?" **Phi Delta Kappan,** 60 (Feb. 1978), 397-400.
5. Popham, W. James. "Well-Crafted Criterion-Referenced Tests." **Educational Leadership,** (Nov. 1978), 91-95.
6. Smith, Frank. **Understanding Reading.** New York: Holt, Rinehart and Winston, Inc., 1971.