DOCUMENT RESUME

ED 065 533 TM 001 454

AUTHOR	Fremer, John
TITLE	Criterion-Referenced Interpretations of Survey Achievement Tests.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-TDM-72-1
PUB DATE	Jan 72
NOTE	35p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Academic Achievement; Achievement Tests; Behavioral Objectives; *Behavioral Science Research; Behavior Theories; *Criterion Referenced Tests; Educational Development; Group Norms; *Measurement Techniques; Skill Development; Test Validity

ABSTRACT

The paper approaches criterion-referencing as a problem of validating tests for particular inferences about human behavior. Some definitions of the term "criterion-referenced" are reviewed, and the position is taken that direct inferences about what a test-taker can or cannot do--criterion-referenced inferences, that is--need not be restricted to tests that are composed of actual samples of the behaviors of interest. Primary attention is given to the criterion of minimal competency in some significant educational area, but other applications are also discussed. Several methods are suggested for validating tests for making inferences to a particular criterion or to several criteria of interest. Some of the limitations of the methods are discussed, and the suggestion is made that more than one method be used to validate any desired criterion-referenced inference. (Author)

TEST DEVELOPMENT MEMORANDUM
TDM-72-1 JANUARY 1972

CRITERION-REFERENCED INTERPRETATIONS OF SURVEY ACHIEVEMENT TESTS

John Fremer
Test Development Division, ETS

Abstract

The paper approaches criterion-referencing as a problem of validating tests for particular inferences about human behavior. Some definitions of the term "criterion-referenced" are reviewed, and the position is taken that direct inferences about what a test-taker can or cannot do--criterion-referenced inferences, that is--need not be restricted to tests that are composed of actual samples of the behaviors of interest. Primary attention is given to the criterion of minimal competency in some significant educational area, but other applications are also discussed. Several methods are suggested for validating tests for making inferences to a particular criterion or to several criteria of interest. Some of the limitations of the methods are discussed, and the suggestion is made that more than one method be used to validate any desired criterion-referenced inference.

Criterion-Referenced Interpretations of Survey Achievement Tests

Meaning of Term "Criterion-Referenced"

This paper[1] takes the position that it is meaningful to relate performance on survey achievement tests to significant real-life criteria such as minimal competency in a basic skills area. A number of ways of establishing relationships between test scores and criterion performance are discussed, but all of the approaches have as their goal the development of criterion-referenced interpretations of test scores. Before possible ways to determine interpretive links are described, it may be useful to review the general idea of criterion-referencing.

Glaser in one of his more recent writings on the topic has offered the following definition of a criterion-referenced test:

> "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." (Glaser & Nitko, 1971, p. 653)

Glaser goes on to suggest that criterion-referenced tests can be differentiated from norm-referenced tests in that they do not focus on the problem of individual differences and are not aimed at the task of determining an individual's relative standing in some norms group. Rather, they tell you what an individual can or cannot do. Glaser talks about the need to construct a criterion-referenced test by defining a population of tasks. Some samples of populations of tasks are all possible pairs of two-digit numbers that might be added or a list of words all of which would have to be spelled.

Many of the articles on the subject of criterion-referenced tests have made use of the Glaser definition, but it is not the only one available. Ebel (1971),

---

[1]The author acknowledges the contribution of Rex Jackson to this paper through his ETS memoranda on the use of teacher judgments to establish criterion levels and through his expression and review of ideas regarding criterion-referencing.

for example, has characterized criterion-referenced measurements as follows:

> "The essential difference between norm-referenced and criterion-referenced measurements is in the quantitative scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle, on some average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performances above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of those abilities. The scale units consist of subdivisions of these total score ranges." (Ebel, 1971, p. 282)

Both the Glaser and the Ebel statements contribute perspectives on the term "criterion-referenced." Their definitions contrast criterion-referenced and norm-referenced tests. For a further discussion of this distinction see Appendix A.

Still another view of criterion-referencing is provided by Popham and Husek (1969, p. 2):

> "Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion; i.e., performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced. The meaningfulness of an individual score is not dependent on comparison with other testees. We want to know what the individual can do, not how he stands in comparison with others."

It is interesting to note that these various definitions agree in that they emphasize the direct interpretability of scores on criterion-referenced tests, but differ in the extent to which they make reference to the method by which the test is constructed. Ebel emphasizes the scale from which interpretations are to be made. Other writers have taken the Glaser position that the method of construction is central; Jackson (1970, p. 3), for example, states:

> "...the term 'criterion-referenced' will be used here to apply only to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents."

The definition of a criterion-referenced test as one that yields direct criterion-referenced interpretations by virtue of the method by which it was constructed

leads to the development of tests by defining populations of tasks and then choosing representative samples from these populations. The narrower the definition of a population of tasks, the more homogeneous the population will be and the greater the degree of confidence one will be able to have about an inference from performance on a sample of such tasks to the total population of tasks. Because of the dependence of this method of criterion-referencing on the ability of the test constructor to specify a limited population of tasks, it seems most appropriate to situations wherein the number of tasks is delimited by the nature of the subject matter -- i.e., identification of the letters of the alphabet -- or where the domain can be specified with reference to particular instructional materials -- i.e., the content of subunit ten of the text used by a particular class. Criterion-referencing by sampling from a fixed population seems most clearly appropriate to classroom developed tests or to special situations that have clearly defined limits.

Criterion-Referencing Through Validation for Specific Criteria

Direct inferences about what a test-taker can or cannot do -- criterion-referenced inferences, that is -- need not be restricted to tests that are composed of actual samples of the behaviors of interest. Considerable use can be made of the very high relationships that have been observed among many apparently diverse tasks within such global areas as "reading," "language usage," or "mathematics." Although some writers have argued that only a sample of tasks directly associated with a particular objective can permit generalization to that objective, the author suggests that other tasks that are not samples of that objective may provide just as good a basis for such a generalization, once the basis for interpretation has been established. More generally, a sample of tasks covering a number of objectives can permit sound inferences to whole classes of objectives, including many not represented in the sample. Given that the number of objectives that can be identified for any significant content area is limited primarily by the patience of the

objectives generator, the use of a survey test as a basis for making criterion-referenced inferences permits considerable efficiency in testing.

The idea of validating a survey test for a specific criterion-referenced interpretation can perhaps be best explained by reviewing a sample use. Consider the possibility that evidence was sought regarding the ability of an individual or some defined group of individuals to read newspaper editorials and to determine the main idea. The direct task-sampling approach to criterion-referencing would require that a population of such editorials be identified and that a systematic method be specified for developing questions to test for understanding of the main idea. Because of the considerable variability from newspaper to newspaper in the reading difficulty of editorials, it would probably be necessary to limit the population to specific types of newspapers. For discussion purposes, it will be useful to assume that a specific newspaper was identified as the target for a criterion-referenced inference, e.g., the Los Angeles Times, or the New York Times, or the Detroit Free Press. The population of editorials could then be specified as all editorials in one section of the paper during a given time period, e.g., one year. The sampling rule might take a form such as "select every 26th editorial by a procedure that provides 14 editorials, 2 for each day of the week." Considerable thought and effort would need to be devoted to the task of analyzing the type of knowledge about editorial reading ability that was desired so that an appropriate method could be developed for selecting editorials of specified quality, type, and difficulty.

The position taken in this paper is that any measurement procedure that would permit a sound inference regarding an individual's ability to read a particular set of newspaper editorials could reasonably be considered a criterion-referenced test of the ability to read these editorials. A criterion-referenced test of this

nature might include some questions similar to the behavior about which an inference was to be made. An example would be a reading test that included passages that were not editorials, with questions about the main idea of these passages, or their mood and tone, or the logicalness of the argument. Use of a survey reading test to make inferences about editorial reading would probably be viewed as a logical step by most individuals who were interested in reading abilities.

A criterion-referenced test might, on the other hand, consist solely of questions quite different from the behavior of interest. Consider a vocabulary test that requires the test-taker to identify the correct synonym for a word. Given the often demonstrated homogeneity of reading "subskills," it seems likely likely that performance on such a test would be so closely related to editorial reading ability that a vocabulary test could be used to make direct statements about editorial reading ability. If a vocabulary test was used for this purpose, its validity would have to be demonstrated for determining editorial reading ability. The process of validation might include identification of the levels of editorial reading ability about which inferences would be made. As a result of the process, a table for score interpretation would be created that might look like the following:

Vocabulary Test Score	Detroit Free Press Editorial Reading Ability
20, 19, 18, 17, 16	Can read editorials and grasp main idea without difficulty. Also, can make inferences beyond what is read and can often find flaws in the writer's argument.

Vocabulary Test Score	Detroit Free Press Editorial Reading Ability
15, 14, 13, 12	Can usually read editorials and grasp main idea, but occasionally has difficulty with the more complicated editorials.
11, 10, 9, 8, 7	Can read some editorials and grasp main idea, but has much difficulty with most editorials. Often concludes that main idea is something quite different from that intended by the writer.
6, 5, 4, 3, 2, 1, 0	Cannot read editorials and obtain main idea. Some individuals earning these scores have almost no reading ability.

A number of observations about the hypothetical criterion-referenced table may be in order. First of all, the usefulness of the interpretive table would depend heavily on the method used to establish the relationship between vocabulary test scores and editorial reading ability. One essential aspect of a good method would be the use of a large and broad enough sample of editorials to permit sound generalization to the many editorials, most not yet written, that would presumably constitute the criterion of interest. By using a broad sample of editorials and, possibly, by using several methods of questioning to determine whether an individual

had understood the main idea, an accurate table of interpretation could be created.

A second facet of the hypothetical criterion-referenced table that bears mentioning is the fact that the levels of editorial reading ability which have been identified are not the only possible ones. The model table attempts to expand the "yes, he can--no, he can't" dichotomy that is often proposed as the model for criterion-referenced or mastery testing. One reason for avoiding a blanket yes or no statement in this instance is that the criterion of reading an editorial to get the main idea is a complex one, as is almost any educational criterion of interest. Consider the apparently homogeneous criterion, "can add two one-digit numbers." It is clear that this can be divided into "adding a one-digit number to itself" -- i.e., 1 + 1, 2 + 2, etc., and "adding two dissimilar one-digit numbers" -- i.e., 1 + 2, 6 + 7. This latter set of tasks is harder than the former as could be demonstrated by asking students at various age levels to tackle samples of the two sets of tasks. Thus, the notion of an absolute dichotomy of mastery versus nonmastery will seldom be meaningful.

The fact that the tasks that define a criterion vary in difficulty contributes one type of uncertainty to any attempt to identify those individuals who have achieved mastery and those who have not. It is possible to identify those who seem to complete just about every sample task we set for them and say with confidence that they have achieved mastery. (Since lapses of attention, or errors of marking paper and pencil tests, will reduce the possibility of perfect performance, some provision for such errors will have to be made. A typical approach is to use a level such as 85% correct as a standard for mastery thus allowing up to 15% of erroneous responses.) It is also possible to identify those individuals that are unable to handle any or more than a small number of the sample of tasks set for them and say, again with considerable confidence, that this group of people has not achieved mastery. Given the complexity of most criteria, though, many

people will fall in a gray area where classification is uncertain. The more closely a test is focused on the particular age or grade level of the group examined, the larger the gray area is likely to be. The vocabulary test for establishing editorial reading ability, for example, would probably classify almost all first graders as nonmasterers and most twelfth graders as masterers. The results for sixth graders would be much less clear-cut.

Minimal Competency as a Criterion

Earlier discussion in this paper has emphasized criterion-referenced interpretation, but has been consistent with the idea that a criterion-referenced test is a test that indicates what a student can or cannot do. In some situations, this definition needs to be expanded to include the possibility that the criterion of interest may be sufficiently complex that it would be better to think of the task of the criterion-referenced test as one of categorizing individuals as members or nonmembers of some defined group. Consider the criterion of minimal competency in some significant educational area; although we might be prepared to give an extensive definition of what constitutes minimum competency, we would probably be hard pressed to assert that our definition depended on whether or not an individual could do one simple task. It would be much more likely that minimal competency would encompass a variety of behaviors, some of which would be more important than others. Careful review of what constitutes a reasonable approach to minimal competency is essential before attempting to determine the extent to which school children of the various grade levels in a school, school district, or state have achieved minimum competency in the basic skills. This is precisely the task that has been attempted by a number of local school districts and state departments of education. Although it is clearly not the intention of the educators who adopt this approach to focus all of their resources and attention on bringing everyone up to some minimum standard, to the exclusion of helping most children

achieve the considerably higher levels of performance that they are capable of, it is of crucial importance to help as many individuals as possible achieve the skills necessary to function effectively in our society. Therefore, the allocation of resources to the main task of helping each individual realize his fullest potential involves placing special weight on those educational objectives that have high survival value. Thus, important issues of social values and priorities affect judgments based on measurements of achievement, whether standardized tests or other procedures, when attention is focused on determining how many students have failed to meet minimal competency standards.

In many situations, we will be interested in determining what proportion of students in a subgroup of interest have failed to achieve minimal competency. This goal can be distinguished from that of identifying which individuals have failed to meet this standard. This distinction between interest in individuals and interest in groups is a crucial one, as summary statistics for groups can be much more accurate than the determinations for individuals within these groups. Thus, it will be much easier to develop and implement procedures for validating criterion-referenced interpretations for groups than for individuals. Even where criterion-referenced interpretations are desired only for groups, the difficulty of the validation task will vary depending on whether or not the same criteria will be set for different groups. Similarity of criteria will depend in turn upon the similarity of educational objectives for different subgroups of students. If the same educational objectives are held for all the students in a heterogeneous population, then the aggregation of data on the proportion of students meeting specified criteria will be considerably more accurate than the determination that has been made for any single individual.

Methods of Criterion-Referencing

1. Using Nontest Information to Set a Minimal Competency Standard --

One method of relating test performance to a minimal competency standard

would involve a review of the information available to a school district or state on the proportion of students at a grade level who are labeled as failures in a particular subject. The types of information that could be considered would include school records regarding the percentage of students who receive failing grades or who are referred to remedial classes in a subject, and research and evaluation reports that estimate the proportion of students who complete a grade level without the minimal skills in a specified subject (e.g., mathematics, reading, other language skills) necessary for profiting from instruction in that subject at the next grade level.

By soliciting this type of information from many sources, it may well be possible to identify some percentage level to serve as a rough estimate of the proportion of students failing to achieve minimal competency. Perhaps most sources would agree that the minimal competency level at a specified grade level for most subject-matter areas for any large group, such as the school children in a city or a state, is probably not as high as the twentieth percentile and yet probably not lower than the tenth percentile. It may be possible to use a variety of kinds of data to establish fairly narrow bounds within which minimal competency would lie.

After an estimate was made of the proportion of students at a particular grade level in a school district or state who failed to achieve minimal competency in a subject, it would then be possible to apply this proportion or percentile to the score distribution for the appropriate test in a survey achievement battery. If, for example, it was estimated that 12% of the fifth graders in a school district or state failed to achieve minimal competency in reading, that reading test score would be identified which corresponded to the twelfth percentile for the same population. It would then be possible to determine, for every school in the district or state, the proportion of fifth-grade students achieving that score or higher. Because wide differences in performance would be likely among the

schools in a district or state, there would be considerable variation from school to school in the number of students achieving this level of performance. At some schools there would be very few or no students who failed to achieve this minimal level. At other schools, substantial proportions of students would receive scores falling below the twelfth percentile.

Clearly, this is a normative approach and it would cause many thoughtful people to challenge, among other things, whether the content of many survey tests was closely enough related to the criterion of interest to warrant the use of test scores in the manner described. It is easy to find loud and angry critics of the normative approach, yet a strong case can be made for its use as one of the methods of assessing the effectiveness of schools. Each of the tests in a well-constructed survey achievement test battery is composed of test questions which sample important behaviors in the area being tested. The balance of questions typically represents the outcome of a careful and conscientious review of a number of subject-matter and psychometric considerations. Each question included in the test battery has very likely been subjected to a series of intensive critical reviews, and the questions have probably been pretested on a sample of school children at the grade level for which the final test is administered. Using a percentile as a basis for evaluating schools is a recognition of the fact that what we can expect students to do can be reasonably based on what we have observed that students can do. The nature of our competitive society is such that each student will eventually have to measure himself against other individuals who are now students with him. No matter how great his absolute level of competence, he will be judged by whether or not he can perform as well as or better than other individuals seeking the same job.

The pattern of schools doing better or worse than the average school, when the proportions of students reaching minimal competency is examined, would closely parallel the results that would be obtained if we were to compare the mean performance of students on the same test. The idea of focusing attention on a minimal competency percentile would merely take into account the fact that we have particular interest in the extent to which students have achieved these minimal levels.

The idea of using a particular percentile as an indication of attainment of a desired criterion level need not be limited to the concept of minimal competency. There might well be other circumstances under which we would wish to determine how many students in various schools have achieved much higher levels of competency. For example, we might be interested in the extent to which the students in a particular grade were capable of performing work usually thought of as being the province of students two or more years advanced in grade. Again, using a variety of sources of information we might determine that for fourth-grade students the 80th percentile on a reading test generally corresponds to a level of reading skill usually thought of as characteristic of sixth-grade students. In that case, we might want to compare all the schools in the state to see what proportion of their fourth-grade students were scoring above that score equivalent to the 80th percentile on state-wide norms. In making this type of analysis, we would have to pay careful attention to the types of items that were contributing to the scores of the fourth-grade students. We could not ignore the content of items and simply treat as equivalent all scores based on the same number of correct answers. It might well be the case, for example, that the 80th percentile fourth-grade student was indeed not performing exceptionally well on sixth-grade material, but was doing a particularly good job on fourth-grade material which he had learned better

than most fourth-grade students. This problem is much less likely to present itself in the area of reading than it is in the area of mathematics where there are much more marked sequence effects in instruction, so that the sixth-grade student who is performing poorly is likely to have mastered some of the work offered at the sixth-grade level while missing a number of things that were introduced at the fourth-grade level. The high-scoring fourth-grade student, on the other hand, is much more likely to be handling all the fourth-grade material very effectively while not having particular success with those concepts and approaches to which he has not yet been introduced.

The idea of referencing tests to criterion levels other than minimal competency is explored in some detail in Appendix B.

2. Teacher Judgments of Individual Test Questions --

The approach described above is essentially normative and empirical. Another possible way to reference survey achievement tests to the criterion of minimal competency is to turn to teachers to ask for their judgments on the extent to which individual questions could be answered correctly by students who are performing in class at a minimal competency level. This method requires judges to examine each question in a test and to estimate what fraction of a group of "barely passing" students would answer that question correctly. Each estimate can be considered as an estimate of the probability that a student exactly at the minimum performance level would succeed on the question. One can sum these estimates across questions to obtain an estimate of the expected score of a student at this minimum level. The results for a number of judges are then averaged to obtain the minimum criterion level score.

A few words about the limitations of this procedure may be in order. It requires judges, who have firsthand experience with students of the type tested, to form a clear conception of the abilities of students at the minimum level and

to make a number of judgments about the likely performance of such students on specific questions. For certain questions the judgments may be difficult ones to make. In no case can one be completely confident that his estimate corresponds exactly to some supposed "true" value. Because these judgments are pooled across questions, however, and then averaged over a number of judges, no single estimate has a very heavy weight in determining the outcome. If judges are able, in general, to make realistic judgments about the performance of minimally competent students on individual questions, the procedure should yield results that are useful and accurate within certain tolerances. It is best to think of the method not as a totally objective statistical procedure, but as a structured way of bringing the wide experience of panels of judges to bear on a specific test score interpretation problem.

It may be useful at this point to review the application that has been made of this teacher judgment approach for the Michigan Assessment Program, a state assessment program for which ETS has been providing measurement assistance since it was initiated in 1969. In February of 1971, two representatives of ETS and representatives of the assessment met with panels of teachers and subject-matter specialists to try this approach for the 1970-1971 form of the fourth-grade and seventh-grade tests in Reading, Mechanics of Written Expression, and Mathematics. For each panel an attempt was made, (within the constraints of panel size,) to obtain reasonable diversity with respect to region and type of community. The following were the major components of each of the panel meetings:

(a) Description of the procedure to be used by the panel members.

(b) Discussion of the criterion level and characteristics of students above and below the level.

(c) Practice on sample questions for which item-analysis data, (i.e., question difficulty) were available.

(d) Rating of questions.

The most difficult problem faced by each of the panels was definition of the criterion. Ideally, each panel would have been able to specify a number of behaviors that discriminate students above and below the minimum level, but limited time precluded a very thorough treatment of this subject. In fact, this objective was achieved only by the mathematics panel which discussed in some detail the specific mathematical skills that characterized students who were making satisfactory progress in the fourth and seventh grades. The nature of the subject-matter of mathematics clearly made it more amenable than reading or writing to this treatment. The panels for reading and writing were unable to come to similar agreement in the time available, and more general ways of characterizing students making satisfactory progress were adopted by these panels.

In rating questions, panelists were asked to judge how likely the barely passing student was to succeed on the question for any reason. Because the questions are multiple choice, a student might succeed on some questions by guessing. On other questions, he might be led astray by plausible incorrect choices. Essentially, the panelists were asked to judge whether the student would arrive at the correct answer in the presence of the given alternative.

Following the panel meetings, the ratings were tallied and averaged and after an adjustment that was required because two panelists did not take the possibility of chance success into account when making judgments, the results in the table below were obtained (the score level given is the minimum "passing" score).

	Number of Raters	Total Number of Test Questions	Minimum Criterion Level	Percent of Students Statewide Scoring Below Minimum Level
Reading - 4th grade	5	50	19	16.5
Reading - 7th grade	5	50	18	11.3
Expression - 4th grade	4	55	23	24.5
Expression - 7th grade	4	65	25	19.7
Math - 4th grade	6	40	14	14.6
Math - 7th grade	4	40	13	14.0

The conclusion reached at the time that this pilot study was carried out was that the critical scores which were identified should be considered only as tentative guidelines. This conclusion was reached despite the fact that the obtained minimal score levels are in the regions one would expect. It is generally advisable to exercise caution in interpreting test results with respect to cutting scores because of the errors of measurement present in any set of scores. When the cutting scores are themselves determined by an essentially judgmental process, the location of the cutting score may be subject to an appreciable amount of error. The notion that results of the pilot study needed to be supplemented by further analyses was supported by the facts that two of the panels failed to define the criterion to their satisfaction and that a number of panelists felt uncertain about their ability to make the judgments required.

The initial use of the teacher judgment method clearly suggested that alternative methods would have to be used in conjunction with this approach. Judges had problems with their task and the level of agreement among them was not very high. Problems with agreement among raters were undoubtedly related to the inability of the groups to define the criterion of minimal competency to their satisfaction. The situation resembles that often encountered in connection with the scoring of essays or other free-response exercises. It is essential to achieve agreement about guidelines for making judgments. Provision needs to be made for raters to develop preliminary standards, to apply these standards on a sample of materials, and then to discuss differences in ratings. Given adequate attention to the definition of the criterion, the approach may be a useful addition to a set of methods aimed at the goal of defining minimum competency.

3. Teacher Judgments Regarding Which Students Are Performing at Minimum Competency Levels --

The method just described required teachers to estimate the performance on individual test questions of students whose performance in class was just at a

minimal competency level. In the Michigan Assessment Program study, the mathematics panel was the only one that succeeded in characterizing the behaviors associated with the term "minimal competency." It may be that a more concerted effort to define minimal competency in terms of student behavior would facilitate the task of referring test scores to a minimum competency criterion. Two closely related approaches that attempt to achieve this goal will be described. The "global judgments" approach merely requires teachers to indicate which of their students are performing at a minimal competency level. The "analysis of classroom performance" approach requires teachers to rate their students with respect to mastery of a number of objectives judged to be central to the achievement of minimal competency in any basic skill area.

a. Global Judgments Approach

Under the global judgments approach it would be necessary for a sample of teachers throughout a school district or state to identify for each of the basic skills areas the students in their classes who are performing at minimal competency. Although some teachers might be willing to so identify their students without additional instructions regarding the definition of the term minimal competency, most teachers will ask for considerably more information before undertaking this task. Minimal competency could be defined generally for each area as that degree of mastery essential for satisfactory further progress in the fourth or seventh grade. The steps that could be taken to determine what degree of specificity of language would be necessary for teachers to undertake the student rating task with confidence would probably best include the convening of panels of teacher and curriculum specialists to review the concept of minimum competency with respect to each basic skill area. It seems clear that standards would vary from school to school and region to region depending on the experience of the teachers and other educators involved. Perhaps several possible definitions might be developed

and arrangements made for pilot testing each of these definitions to see whether they prove to be workable when teachers are asked to use them as a basis for rating students. The definitions might be tested both by calling for a feedback from teachers who served as raters, and by analyzing the relationship between the ratings obtained and other data available for students in the sample.

Once workable definitions of minimal competency in each of the basic skills areas have been obtained, teachers would be asked to indicate which of their students have or have not achieved minimal competency using the global definitions that were derived. The selection of the sample of teachers to carry out this method and the resolution of such questions as whether or not teachers would be asked to rate each of their students on all three basic skills areas or whether each teacher would be asked to rate his or her students on only one basic skills area will not be explored in this paper. This latter question, though, is an interesting one because it offers the possibility of investigating the extent to which minimal competency in one basic skill area is related to minimal competency in another basic skill area. That such a relationship exists can be inferred without any experimental testing, but the closeness of the relationship warrants investigation.

After the results of student ratings have been collected, the next step under the global judgment approach will be that of relating the judgment of teachers to the performance of students on the survey achievement test battery. It is easy to predict that the students who were judged to be above a minimal competency level of performance will obtain higher scores, on the average, than students who were not judged to have reached minimal competency. How close will this relationship be, however, and how much agreement will there be among the judgments made by teachers who were working with students from quite different backgrounds and general achievement levels? One way of looking at the data which will be obtained

will be that of identifying the band of test scores that would be most closely related to the dividing line between students above and below minimal competency for a variety of student groups. Looking at the reading test, for example, it would be possible to identify for every classroom included in the rating sample that test score on the reading test which seems to be most closely related to the dividing line between students above and below minimal competency in reading as judged by the teachers. Consider the following possible distributions of such cutoff scores for a fourth-grade reading test as determined in 100 classrooms:

Hypothetical Distributions of Fourth-Grade Reading Cutoff Scores

Score	A	B
26-30	2	10
21-25	13	20
16-20	35	20
11-15	35	20
6-10	13	20
1-5	2	10
Total	100	100

It seems clear that hypothetical distribution A in which 70 classrooms have cutoffs in the 11-20 range would lead us to be much more confident that our global judgment approach was providing the kind of information that we could use to categorize students as minimally competent or not minimally competent on the basis of scores on a fourth-grade reading achievement test than would distribution B.

In addition to examining the extent to which different schools did or did not agree on the score level associated with minimal competency levels, it would be possible to analyze the extent to which high and low scoring students within a school were or were not categorized accurately by the score which best divided them into minimal competency and nonminimal competency groups. Such an analysis would permit a determination of the degree to which proper classification of students

varies with the overall level of performance of students on a test. Since factors such as lack of motivation can lead to low test scores, it may be that low scoring students would be less accurately categorized by test scores than higher scoring students.

b. Analysis of Classroom Performance Approach

A second technique for using teacher judgment of students as an approach to criterion-referenced interpretations of survey achievement tests would involve asking teachers to identify that set of student performances in the classroom which would define minimal competency in the area covered by a test. Again, this task might best be carried out by assembling educators to review the expected outcomes of instruction for students at the grade levels tested in an assessment program. The performances that would be described by appropriate panels or committees of educators would describe valued educational outcomes at whatever level of specificity could reasonably obtain given the nature of the outcome. Not all such statements of outcomes would be statements of discrete and easily observable behaviors. Any attempt to reduce all valued educational outcomes to discrete bits of easily observable behavior will tend to subvert legitimate educational goals in the interest of ease of measurement.

It is likely that the identified educational objectives will vary greatly in the degree of importance attached to them by any educational group. One step toward refining the educational objectives so that they could serve as a basis for teachers evaluating their student's progress would be the classification of objectives as follows:

(1) Crucial and universal: essential to attain for every child; mastery is essential to further progress

(2) Highly important and universal: if educational system is functioning properly, nearly every child should be able to do by age X

(3) Desirable but not universal: expect to attain for limited percentage of children

(4) Optional, local option, side effect

(5) Irrelevant

(6) Harmful, counterproductive

These categories are included as a reminder that some outcomes of instruction are neither intended nor desired.

The mechanism for sorting these objectives might be a survey of educators and other interested groups. One such group might be users of the educational product. If interest were centered on secondary school students, employers would be a natural resource. In any event, teachers of the grades above those taking the tests should participate in the process. Various elaborations of the survey approach might be used -- perhaps the Delphi technique could be applied to raise the level of consensus in the ratings. Statements with a large rating variance might be identified and refined further.

After educators had tackled the difficult task of assigning priorities to the educational objectives in each of the areas of interest, some set of objectives for each area would be used as a basis for teachers to rate their students with respect to attainment of each of the objectives. In those cases where the educational objectives can be stated as specific behaviors, we can probably expect that teachers will be able to make such judgments with a fair degree of competence. Where the objectives require inferences beyond easily observable behaviors, much less confidence in the ratings will be possible. The pilot testing of the objectives that have been identified might provide a basis for eliminating some objectives on the grounds that they cannot be reliably rated. It would be necessary to design pilot testing so that judgments of reliability could be obtained. Another outcome of the pilot testing of rating procedures might be that some objectives that appear to be quite distinct can indeed be collapsed into a single objective because students develop in such a way that attainment of one of the objectives almost always signals attainment of the other.

The refinement of the rating procedure on the basis of pretesting would result in the development of appropriate materials and procedures that could be employed on a large sample of students by their classroom teachers. For each of the students so rated information on the student's performance on the survey achievement tests would be collected also. The relationship among the various ratings and performance on the appropriate instrument would be studied following some of the same procedures indicated for the global judgment approach. The outcome of whatever statistical analyses were performed would be the identification of bands of test scores that would be associated with the following judgments:

(1) Below minimum competency level

(2) Uncertain

(3) Above minimum competency level

The width of the uncertain band under any of the methods of developing criterion-referenced interpretations would vary with the validity of the test for the intended discrimination. The width of the uncertain band would be chosen so as to keep the probability of error when classifying a student as above or below the minimum level within some specified boundary. Alternatively, scores could be reported in terms of the probability that a student with a particular score is above or below a minimal level.

4. Supplemental Work Sample Tests --

Each of the methods and variations on methods of criterion-referencing that have been discussed so far have involved the interpretation of survey achievement tests rather than the development of new tests. It is possible to develop new tests with a much narrower focus than the typical survey achievement tests. (This approach was implied in the earlier discussion of the idea of criterion-referenced interpretations.) These would cover smaller areas of content and could also contain items with a narrower spread of difficulty than the questions in typical

survey achievement tests. Looking first at the issue of range of content, it is clear that any significant educational area and particularly a basic skills area could be dissected finely enough to produce lists of educational objectives so exhaustive that they could only be tested by devoting a substantial portion of the school year to the project, if every student were tested separately on each objective. It would, therefore, clearly not be practical to undertake the task of preparing sets of test questions to cover every possible objective that was of interest in each area. It would be possible, though, to identify certain educational objectives in each area that were of such crucial importance that it would be reasonable to determine the extent to which students had achieved these particular objectives. The prospect of testing a significant number of these objectives seems within the realm of possibility if the techniques of item and people sampling were to be employed. If, for example, it proved possible to identify for a particular area some small set of educational objectives that were to be assessed in great detail, it would be possible to put together sets of test questions focused particularly on those objectives and to administer these sets of test questions to every nth student at a particular grade level who takes a survey achievement test. It would then be possible to analyze the relationship between the survey achievement test and each of these focused subtests. It may well prove to be the case that performance on some of the focused subtests would be so highly related to performance on the relevant survey test that there would be no increase in measurement accuracy associated with using the subtest rather than the survey test to make an inference about the proportion of students having achieved mastery of that educational objective. There could remain, however, some focused subtest that would indeed provide measurement that was sufficiently better than that provided by the total test in an area to warrant the inclusion of that subtest in a future assessment program.

The fact that a particular subtest did not result in improved measurement of its particular objective would not mean that nothing of value was gained from trying out the subtest in an experimental setting. By establishing through experimental pretesting the relationship between the survey achievement test and the subtest of interest, it would be possible to make inferences to the educational objective of interest from the survey achievement test. In some ways this seems to be the central problem facing school districts and state departments of education; i.e., what inferences can be made from performance on survey achievement tests to specific competencies of students.

In those instances wherein it proves useful to continue using a focused subtest for accuracy of measurement of a particular educational objective, it may still prove useful to include the survey achievement test performance in an area in any equation for making an inference to the proportion of students who have achieved a particular objective. The degree of correlation among subtest and survey test would be of course, the determining factor. Let us assume, for example, that a special subtest was developed in the area of mathematics to test the performance of students on fractions. The use of a subtest on fractions as one of the instruments given to every eighth student along with the total survey battery would permit a determination for each school of how well every eighth student performs in this area. Previous administrations of this subtest with the survey mathematics test would have established the precise relationship between performance on this subtest on fractions and the total mathematics test at the grade level of interest. In making an inference regarding mastery of fractions for a particular school, we could assign a weight in our inference-making equation to the performance of the one-eighth group of students who took the fractions test, but we could also include in our equation the performance of the total group of

students on the total mathematics battery. It may be that when scores on the total mathematics battery were high enough so that good discrimination was possible at the minimal competency level it would be to our advantage to assign greater weight to the total mathematics battery in making an inference regarding the ability to solve fractions than we should give to performance on the subtest specifically devoted to fractions. The crucial issue here is the degree of error associated with predictions from the larger test to the smaller test as opposed to the degree of error associated with the fact that we will have tested only a sample of students in a particular school on the subtest.

The issue previously raised regarding the degree of error associated with two possible approaches to making inferences to ability with respect to a particular educational objective would relate in good part to the spread of item difficulties in the total test as opposed to the focused test. If the collection of test items regarding fractions that were included in a special subtest were all at a level of difficulty most appropriate for making discriminations between minimally competent students and those who were not minimally competent, it appears highly likely that the fractional subtest would be more useful as a basis for making judgments about a student's mastery of fractions than would be the total test with its broad range of difficulties. It is a maxim of test score theory that a test discriminates best if it contains items of about middle difficulty for the particular ability level about which you want to make inferences.

5. Stand-alone Work Sample Tests --

In addition to sampling with focused tests so that the relationship between the total tests and the focused tests can be determined, there may also be some need in an assessment program to administer some focused tests on a sampling basis so that results can be reported directly for groups of students. It may be, for

example, that there are some educational objectives that are of such great interest to people that they should be measured directly even though indirect measurement would give as good or even better measurement, because of greater efficiency, than any direct approach. Earlier in this paper, for example, the possibility of assessing a student's ability to read newspaper editorials through the use of a vocabulary test was discussed. Although a vocabulary test or perhaps a survey reading comprehension test may be just as good a predictor of editorial reading ability as a test made up solely of editorials, it may be that the greater face validity of a test made up of editorials would make it a desirable component of an assessment battery on a sampling basis. Similarly, a survey mathematics test might predict quite well the ability to add up the figures on a menu or to compute an invoice for a sale in a store, yet a survey test would lack the face validity of a collection of such tasks. Wherever there is a need to report the results of testing to individuals who are uncomfortable or even suspicious of statistical evidence of relationships, the use of the direct measure may be a desirable move. It seems clear that many of the results of the National Assessment could have been obtained much more easily through the administration of survey tests that are now available in any of the areas assessed in the National Assessment. By focusing attention, however, on particular test questions, the National Assessment has succeeded in capturing public interest in what students can do in particular areas. This same gain might accrue to any assessment program if a similar strategy of using questions with high public interest and clear implications were followed. We must exercise caution in this area, however, since it is easy to mislead people into believing that the results on one question somehow provide a direct insight into student competencies. It is more likely that the score that a student earns on a particular question is as much a function of the design and format of that question as it is of the

difficulty of the concept involved; another question that looks very much like the original may result in quite different student performance.

Summary

The idea has been advanced that criterion-referencing may profitably be approached as a problem of validating tests for particular inferences about human behavior. In taking this position, the writer recognizes that he is merely advocating a way of using tests that has been the goal of test developers at least since the time of Binet, who developed his intelligence tests to identify students who would have difficulty functioning in normal classrooms. Several methods have been suggested for validating tests for making inferences to a particular criterion or to several criteria of interest. In each instance, the method suggested draws on well-established psychometric procedures, but each method carries with it the certainty of some degree of error that is associated with all measurement. It is suggested, therefore, that more than one method be used to validate any desired criterion-referenced inference.

References

Ebel, Robert L. Criterion-referenced measurements: limitations. School Review, 1971, 79, 282-288.

Glaser, Robert & Nitko, Anthony J. Measurement in learning and instruction. In Robert L. Thorndike (Ed.), Educational Measurement, Washington, D. C.: American Council on Education, 1971. Pp. 625-670.

Jackson, Rex Developing criterion-referenced tests. TM Report No. 1. Princeton, New Jersey: ERIC Clearing House on Tests, Measurement, & Evaluation, 1970.

Popham, James W., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

APPENDIX A

Contrasting Criterion-Referenced and Norm-Referenced Tests

The major tool for interpreting almost all educational tests is the norms distribution. Such distributions are used to arrive at norms tables, percentile scores, stanines, and other such descriptive tools. By using a norms table one can determine the relative position of an individual or group in a norms population. If, for example, for seventh-grade students at a particular junior high school that participated in a state assessment program, the mean mathematics score on a survey achievement test is 50.2, one can refer to the norms table for that state and determine the percentile rank of that score. The percentile rank indicates the proportion of the schools in the state that obtained lower seventh-grade mathematics scores. When looking at the score for an individual, a norms table can be used to determine what percentage of the students in some defined population obtained a lower or a higher score than that individual. The often voiced criticism, though, with respect to the use of tests with individuals is that the information that a mathematics test should provide is how much mathematics does each student know. Educators indicate that they are interested in how well a student does relative to other students, but that they recognize the fact that they do not really know how well the other students did either. A somewhat similar argument applies to the use of tests to obtain information about groups of students, such as all students at a particular grade level in a school. Questions are asked about the number of these students who have achieved enough mastery of any subject to have satisfied the minimum objectives of instruction. Or, if one looks at the other end of the accomplishment spectrum, how many have mastered so many of the objectives of that grade level as to be ready to undertake advanced work that is usually thought of as part of the content of higher grades.

The distinction between absolute and relative standards implied by the terms "criterion-referenced" and "norm-referenced" has a counterpart in areas other than testing. Consider, for example, the world of wages, a real-life area that is of interest to almost everyone. What most people want to know when they apply for a job is how much money they will be making. They will surely be interested in whether or not they are making more money than other people with similar titles, in the same organization or in the same field, but that information alone is insufficient. The dollar figure is meaningful to them even without any reference to the salaries of any other individuals. This is not to say that comparisons are not of interest and are not valuable but the score itself, or in this case the salary figure, has meaning. Its meaning derives from the individual's awareness of what such a salary can buy.

APPENDIX B

Criterion-Levels Other Than Minimum Competency

This report has focused on methods of relating assessment battery performance to minimal competency levels. It is clear, though, that much of education is directed at helping students realize their fullest potential rather than bringing them up to some minimum. A fully individualized measurement and evaluation program within any classroom reviews the performance of each child with respect to his or her capacities. Many students enter any grade with developed competencies well above any realistic minimums that could be set for most students or even for the average student to reach at the end of that grade. Although the completely individualized assessment that can be done at the classroom level is not practical for assessment at the school district or state level, it may be possible to identify other criterion levels of interest. The bulk of this report has been devoted to a discussion of possible methods of relating assessment battery performance to minimal competency. In this section, some speculations will be made about criterion levels other than minimal competency that might be of interest to educators in each of the basic skills areas of reading, mathematics, and language arts.

1. Reading -- The highest criterion level of interest in the area of reading might be defined as "having the ability to read independently, without teacher assistance, literature of a specific level of difficulty." By varying the specified level of difficulty of the literature that is to be read without assistance, various levels of competency could be specified. It might be useful to have another criterion that related to the ability to read literature at a specified level of difficulty without discomfort, but only with teacher assistance. What is suggested here is the often made distinction between a student's independent reading level, his instructional level, and his frustration level.

It would also be possible to establish criterion levels within the field of reading according to one's predictions of how well the student is able to handle the reading demands in the other subject-matter areas in the school. This consideration may have influenced the educators who helped identify minimum competency levels for the reading tests in the Michigan Assessment Battery, but it seems likely that the levels they set were too low to encompass all the students who would have great difficulty functioning in their social studies classes, for example, because of their low reading skill.

2. Mathematics -- In the field of mathematics the highest criterion level might relate to the ability of students to take higher level mathematics. At grade seven, for example, it may or may not be customary in a particular school district to make predictions about whether or not a student would be capable of handling geometry at a higher grade, but most seventh-grade teachers could make some comparable judgment about the students in their course. It might be hard here to differentiate between a teacher's estimate of how bright a student is and his estimate of how much the student had profited by instruction, but it is doubtful whether this distinction can ever be a clear one.

For an intermediate level of competency in the mathematics area, it might be possible to identify that group of students capable of answering all problems with which they have already had experience, but with no apparent ability to translate their knowledge to new situations using the basic principles that they should have acquired.

3. Language Arts -- Within the language arts area, levels of competency might be established, as was suggested in the reading area, that would relate to the students perceived ability to communicate the knowledge that he has obtained in areas such as science or social studies through the written word. Some small percentage

of students can be identified who would not be likely to lose credit in their other subjects because of problems of writing. Other real-world criteria might also be established, using the logic that has been employed in the National Assessment area. In that setting, for example, a number of measurement people have discussed the possibility that a letter written to the personnel office of a company might or might not receive favorable attention merely because of the quality of the language contained within it. It would be possible to have such letters read by personnel managers so that they could classify them appropriately.