DOCUMENT RESUME

ED 206 725 TM 810 637

AUTHOR Hambleton, Ronald K.; Eignor, Daniel R.
TITLE Competency Test Development, Validation, and Standard-Setting.
INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 20 Oct 78
NOTE 52p.; Paper presented at the Minimum Competency Testing Conference of the American Education Research Association (Washington, DC, October 12-14, 1978).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Competence; Criterion Referenced Tests; *Cutting Scores; Elementary Secondary Education; *Minimum Competency Testing; *Models; *Test Construction; Test Format; Test Reliability; *Test Validity

ABSTRACT

In light of the widespread use of competency testing, the authors consider that it is important to determine ways of developing and using competency testing to insure that it achieves its full potential. The paper, in three parts, introduces a model for the development and validation of competency tests, reviews several methods for setting standards or competency levels, and makes suggestions for future research and development. Firstly, definitions of competency testing, criterion referenced tests, and standards are provided. The twelve step development and validation model introduced incorporates: competency selection; test specification; writing and editing test items; determining content validity; further editing; test assembly; standard setting; test administration; collection of reliability, validity, and norm data; preparation of users and technical manuals; periodic collection of additional information. The standard setting models considered are continuum models, of which the major assumption is that mastery is a continuously distributed ability. These models are further subdivided, for descriptive and comparative purposes, into judgmental, empirical, and combination models. The characteristics of the nineteen models thus categorized are then discussed. The development of guidelines for competency test development and further work on the moral and technical issues involved in standard setting are recommended. (AEF)

October 20, 1978

Competency Test Development, Validation, and Standard-Setting[1,2,3]

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

The establishment of minimum competency testing programs in elementary and secondary schools, and for many professions, has reached immense proportions (or epidemic proportions, if you view the trend negatively). For example, well over half (33 to be exact) of our states have passed legislation requiring assessment of the "competence" of their elementary and high school students (Pipho, 1978). Further, many of these states require that students demonstrate at least a minimum level of performance on a set of competencies in order to receive a high school graduation diploma. Why are so many state legislatures mandating minimum competency testing? It appears that it is to discourage schools from the practice of promoting all students and awarding high school graduation diplomas based on school attendance only. It is common for legislators and

---

[1]Preparation of this paper was supported, in part, by a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

[2]Laboratory of Psychometric and Evaluative Research Report No. 84. Amherst, MA: School of Education, University of Massachusetts, 1978.

[3]A paper presented at the AERA Minimum Competency Testing Conference, Washington, October 12-14, 1978.

parents to say that minimum requirements in the "basic skills" must be set for students to graduate with a diploma which has some meaning. Perhaps it is not surprising to observe that participating states are approaching the task of establishing minimum competency testing programs differently. Some states are emphasizing "life skills," others "school skills," and yet other states have incorporated both types of skills into their competency testing programs. Also, the school years in which testing is done varies from one state to the next. Finally, there are variations in the ways competencies are identified and measured, and standards set (Haney and Madaus, 1978).

The rapidity of change in school, district, and statewide testing programs and the demand for high quality tests has dictated that substantial research and development work be undertaken. Included among the more important research and development topics are: Identification and definition of competencies, management of competency testing programs, development and validation of competency tests, methods of determining standards, and uses and interpretations of competency test scores (Brickell, 1978).

Other speakers at this AERA Competency Testing Conference have considered the philosophy and assumptions of competency testing programs, as well as their potential (and in some cases, demonstrated) effects on student performance and school curricula. Our contribution to the conference will be to consider some ways for developing and using competency tests to insure that competency testing programs achieve their full potential, whatever that potential may be. Specifically, this paper was prepared to accomplish three purposes:

1. To introduce a model for developing and validating competency tests.

2. To provide a review of several promising methods of determining "standards" or "minimum performance levels."

3. To offer several suggestions for future research and development.

We will not debate the merits of competency testing in this paper. Others are far more informed about the issues and capable of articulating them to those who have an interest. Our work will begin at the point where (1) a decision has been made to initiate a competency testing program, and (2) a set of competencies has been identified and tests to measure individual performance on the competencies are required. Three other points concerning our work should also be mentioned:

1. Attention is focused on the use of competency tests for making decisions about individuals. When groups of examinees are of primary interest (as in program evaluation studies or many statewide testing programs), approaches to competency test development and test score usage are somewhat different. (For example, individuals and test items can be sampled—i.e., matrix sampling is used—and "standards" are set for group performance.)

2. Many of our examples will be from elementary and secondary school settings although most of the testing technology discussed applies equally well to the development of competency tests in other content areas.

3. We will focus on the construction of paper and pencil tests. Steps for constructing performance tests are basically the same but special attention must be given to topics such as the design and use of behavioral checklists, and inter-rater reliability.

The remainder of the paper is divided into three sections: Development and Validation of Competency Tests, Methods of Standard Settings, and Suggestions for Future Research and Development.

Development and Validation of Competency Tests

A Competency Test

Perhaps we should begin with a definition of a competency test:

> A competency test is designed to determine an examinee's level of performance relative to each competency being measured. Each competency is described by a well-defined behavior domain.

The definition makes clear that the purpose of a competency test is to provide information about an individual examinee's level of performance on each competency which is measured by a test. There will be as many test scores as there are competencies measured by a test. Also, competencies are clearly written so that there will be a high level of agreement among users of the test about the content (behaviors) defining the competency. This desirable goal can be accomplished through the use of "domain specifications" (Popham, 1978a). This term will be described in more detail later. There is one other point. There is nothing inherent in the definition of a competency test which requires test scores to be compared to "standards." In fact, the percentage scores (reported by competency) provide excellent descriptive information about examinee performance. Since it is common, however, to interpret examinee test performance relative to standards (an examinee who scores equal to or above a standard set at 70% [say] on the set of test items included in a competency test is described as a "master" or "competent"), it is necessary to introduce a new term, "minimum competency testing."

> A minimum competency test is designed to determine whether an examinee has reached a prespecified level of performance relative to each competency being

> measured. The "prespecified level" or "standard" may vary from one competency to the next. Also, each competency is described by a well-defined behavior domain.

A "standard" (sometimes it is called a "cut-off score" or a "minimum proficiency level") is a point on a test score scale which is used to separate examinees into two categories, each reflecting a different level of proficiency relative to the competency measured by the test under consideration. It is common to assign labels such as "master" or "competent" to those persons in the higher-scoring category and "non-master" or "incompetent" to those persons in the lower-scoring category. Note that if a test measures more than a single competency and if examinees are to be classified into competency categories based on their performance on <u>each</u> set of items measuring a competency, as is often the case, a standard is set for <u>each</u> competency measured by the test. There will be as many competency decisions as there are competencies measured by the test.

It is important at this point to separate three types of <u>standards</u>. Consider the following statement:

> School district A has set the following target—It desires to have 85% or more of its students in the second grade achieve 90% of the reading objectives at a standard of performance equal to or better than 80%.

Three types of standards are involved in the example:

1. The 80% standard is used to interpret examinee performance on each of the objectives measured by a test.
2. The 90% standard is used to interpret examinee performance across all of the objectives measured by a test.
3. The 85% standard is applied to the performance of second graders on the set of objectives measured by a test.

Only the first use of standards will be of interest in this paper.

From the definitions above, it is clear that minimal competency tests are a special type of competency test (tests where standards are introduced to interpret examinee performance) and as we shall see later, competency tests are a special type of criterion-referenced test (i.e., those tests which are used usually in certification and licensing situations).

Finally, there is nothing inherent in the definition of competency testing (or minimum competency testing) which precludes the measurement of school skills (for example, arithmetic, spelling, and reading) or life skills (for example, balancing a check book, following directions, or answering a job advertisement).

Competency Tests and Criterion-Referenced Tests

The competency testing technology would be in an embryonic stage were it not for the work done in developing a criterion-referenced testing technology since the late 1960's. A competency test is simply a particular kind of criterion-referenced test and therefore, like a criterion-referenced test, it must be developed and used in ways somewhat different to better-known norm-referenced tests. Glaser (1963) and Popham and Husek (1969) introduced the notion of criterion-referenced testing so that test score information of the type needed to make a variety of individual and programmatic decisions would be available. Norm-referenced tests are designed, principally, to facilitate the use of scores derived from the tests to make comparative statements about individuals. This is not the primary type of information required by individuals who implement competency-based testing programs. They require information about the level of individual performance relative to well-defined content domains (referred to as "domain specifications").

A considerable amount of progress has been made during the last ten years toward the establishment of a practical and usable criterion-referenced testing technology. The existence of this technology (see, for example, Hambleton & Eignor, 1978; Hambleton, Swaminathan, Algina & Coulson, 1978; Millman, 1974; Popham, 1978a) makes it possible, among other things, to develop criterion-referenced tests for use in diagnosing student learning deficiencies, monitoring student progress, and evaluating school programs. The same basic technology is useful also for individuals who must develop and validate minimum competency tests for (say) high school graduation, although matters such as the selection of competencies for inclusion in a test and approaches for developing and validating tests will be handled somewhat differently.

At what stage of development is a competency testing technology? There would be considerable agreement among measurement specialists on the statements offered below:

1. Definitional problems have been sorted out (for example, distinctions among norm-referenced, criterion-referenced, competency-based, domain-referenced, and objectives-referenced tests are clear).
2. The need for "domain specifications" is clear and adequate methods for developing them do exist.
3. There is at least an adequate technology available for developing and validating competency tests.
4. The problem of test score reliability has been articulated clearly and approaches now exist for determining reliability of scores for various intended uses.
5. Methods for using and reporting competency test score information are available.

The interested reader is referred to Hambleton et al. (1978) and Popham (1978a) for further discussion of the points above.

Of course, there remains a considerable amount of work to be done. The four topics below are especially important:

1. Improved guidelines for preparing domain specifications,
2. Guidelines for evaluating competency tests and test manuals,
3. Research on the relationships among test length, test score reliability and test score validity,
4. Further consideration of issues and methods of determining standards, and of guidelines for implementing each of the methods.

How should a competency test be developed and validated? This problem is addressed in the next section of the paper.

Steps in Test Development and Validation

A twelve step model for developing and validating competency tests is presented in Figure 1. The importance of each step in the model depends upon the size and scope of the test development and validation project. An agency with the responsibility of producing a state-wide competency test will proceed through the steps in a rather different way from a small consulting firm or a school district.

In brief, the twelve steps are as follows:

Step 1--Competencies must be prepared or selected before the test development process can begin.

Step 2--Test specifications are needed to clarify the test's purposes, desirable item formats, number of test items, instructions to item writers, etc.

1. Preparation and/or Selection of Competencies

2. Preparation of Test Specifications (for example, Specification of Item Formats, Appropriate Vocabulary, and Number of Test Items/Competency)

3. Writing Test Items "Matched" to Competencies

4. Editing Test Items

5. Determining Content Validity of the Test Items

    a. Involvement of Content Specialists
    b. Collection of Student Response Data

6. Additional Editing of Test Items

7. Test Assembly

    a. Determination of Test Length
    b. Test Item Selection
    c. Preparation of Directions
    d. Layout and Test Booklet Preparation
    e. Preparation of Scoring Keys
    f. Preparation of Answer Sheets

8. Setting Standards for Interpreting Examinee Performance

9. Test Administrations

10. Collection of Reliability, Validity and Norms Information

11. Preparation of a User's Manual and a Technical Manual

12. Periodic Collection of Additional Technical Information

Figure 1. Steps for Developing and Validating Competency Tests.

Step 3--Items are prepared to measure competencies included in the test (or tests, if there are going to be parallel-forms, or levels of a test varying in difficulty).

Step 4--Initial editing of items is completed by the individuals writing them.

Step 5--A systematic assessment of items prepared in steps 2 and 3 is conducted to determine item validities. Essentially, the task is to determine the content validity of the test items.

Step 6--Based on the data from step 5, it is possible to do further item editing, and in some instances, discard items that do not at least adequately measure the competencies they were written to measure.

Step 7--The test (or tests) can be assembled.

Step 8--A method for setting standards to interpret examinee performance is selected, and implemented.

Step 9--The test (or tests) can be administered.

Step 10-Data addressing reliability, validity, and norms can be collected and analyzed.

Step 11-A user's manual and a technical manual should be prepared.

Step 12-This step is included to reinforce the point that it is necessary, in an on-going way, to be compiling technical data on the test items and tests as they are used in different situations with different examinee populations.

Whether a competency test or a minimum competency test is being developed, steps one to six will be the same. At step seven, it is possible (although not essential) that different methods will be used to select test items. Step eight is unique to minimum competency testing. Remaining steps in the model (steps 9 to 12) are essentially the same for the two types of tests. About the only differences are those concerning approaches to validating test scores. Clearly, since the two types of tests are intended to accomplish different purposes, approaches for validating test scores will, in general, be different.

Four of the steps (1, 3, 5, and 7) in developing a competency test will be discussed next. Useful references for an expanded discussion of the other steps are Hambleton and Eignor (1978); Hambleton, Swaminathan, Algina, and Coulson (1978); Millman (1974); and Popham (1978a).

1. Statement of Competencies.—It is popular to write competencies in "behavioral terms." However, while behavioral statements have some desirable features (for example, they are relatively easy to produce), they often lack the clarity necessary to permit a clear determination of the domain of test items measuring the behaviors defined by a

competency. If the proper domain of test items measuring a competency is not clear, the task of preparing valid test items is more difficult. Also, it is impossible to select a representative sample of test items from that domain if the domain is not clearly specified. Since it is often desired to interpret examinee performance on a sample of test items measuring a particular competency as an estimate of that examinee's level of performance in the larger domain of items, it is essential to have the domain of test items specified clearly, and to choose a representative sample of test items.

Domain specifications are an important new development in competency testing (Popham, 1978a). Domain specifications clarify the intended content specified by a competency. Such information is invaluable to teachers (they must teach the competencies defined in the domain specifications), to parents (they often wish to have information about the competencies), and to item writers (they must produce "valid" test items, i.e., test items that are representative of the domain of items measuring each competency). There are at least four steps outlined by Popham for the development of domain specifications. The first involves the preparation of a general description. The general description could be a behavioral objective, a detailed description of the competency, or a short cryptic descriptor. Next, a sample test item is prepared. This will help to clarify the domain of test items and to specify item format. The third step is perhaps the most difficult. It is necessary to indicate the content included in the domain. In the final step, characteristics of response alternatives or response limits are specified. An example of a domain specification is shown in Figure 2.

SKILL: The student will identify the tone or emotion expressed in a paragraph.

SAMPLE ITEM:

Directions: Read the paragraph. Underline the best word to complete the sentence.

Jimmy had been playing at the beach all day. It was time to go home. Jimmy sat down in the back seat of the car. He could ha dly keep his eyes open.

Jimmy felt _____.

A. afraid B. friendly C. tired D. kind

CONTENT:

1. The paragraph will contain situations which are familiar to the students being tested.

2. The paragraph will contain no less than three and no more than six sentences. The readability level will be no higher than Second Reader.

3. The emotions expressed will be from the following list:

sad	mad	angry
tired	scared	friendly
happy	lucky	smart
kind	excited	proud

RESPONSE MODE:

1. Responses will be one word in length.

2. The items will contain one correct and three incorrect responses.

3. Distractors are to be words describing a feeling and may be taken from the list above.

4. Avoid having distractors as possible answers. (i.e., in the sample item, "mad" would not be a good choice for a distractor. Jimmy could feel mad about leaving the beach.)

Figure 2. An example of a domain specification from the reading area. (The authors are grateful to Marlene Teichert for the example.)

The important aspect of implementing the steps is that they lead to specified item domains; it is not necessary, however, that homogeneous content domains be produced. Specificity and homogeneity are different concepts. Millman (1974) makes this point, "The domain being referenced by a [criterion]-referenced test may be extensive or a single, narrow objective, but it must be well defined, which means that content and format limits must be well specified" (p. 314).

3. Generation of Test Items.—Once domain specifications are defined, the test constructor must generate test items. If the domains are defined in a perfectly precise manner, then the items themselves would not need to be generated. The items would simply be a logical consequence of the domain definitions (for example, see Hively, Patterson, & Page, 1968). Unfortunately, however, such precision will seldom be achieved in practice and so test items must be produced and procedures, like those described in step five, used to check the adequacy of the test items.

Principles of item writing used in norm-referenced achievement test construction apply to competency tests as well. It is necessary though, for item writers to attend closely to the domain specifications. Test items should be written to "tap" behaviors in the domain of behaviors defined by the domain specifications. After editing of the test items, the next step is to determine the item validities.

5. Determination of Content Validity.—Generally speaking, the quality of competency test items can be determined by the extent to which they reflect, in terms of their content, the domains from which they were derived. The problem here is one of item validation; unless one can say with a high

degree of confidence that the items in a competency test measure the intended competencies, any use of the test score information is questionable. When domain specifications are utilized, the domain definition is never really precise enough to assume a priori that the items are valid. Thus the quality of the items must be determined in a context independent from the process by which the items were generated. This is an a posteriori approach to item validation. Some procedures have been designed to assess whether or not a direct relationship between an item and a domain or objective exists through analysis of data collected after the item is written (Hambleton & Eignor, 1978; Hambleton & Fitzpatrick, in preparation; Popham, 1978a).

There are two approaches which may be used to establish the (content) validity of test items. The first approach, and the approach we feel holds the most merit, involves the judgment of test items by content specialists. The judgments that are made concern the extent of "match" between the test items and the domain they are designed to measure. Questions asked of content specialists about content validity of test items can be reduced to two important ones:

1. Is the format and content of an item appropriate to measure some part of the domain specification?

2. Does the available set of test items adequately sample a particular domain?

A second approach is to apply empirical techniques to examinee response data in much the same way empirical techniques are applied in norm-referenced test development. In fact, along with some recently developed empirical procedures for competency tests, several norm-referenced test item statistics can (and should) be used. The problem is to ensure that these statistics are used and interpreted correctly in the context of

competency test development. Item statistics should be used to detect aberrant items that need to be reworked, and not to make final decisions about which items are to be included in a competency test. An excellent review of item statistics for use with competency tests has been prepared by Berk (1978).

7. Test Assembly.—The length of a competency test (or more importantly, the number of test items measuring each competency in a test) is directly related to the usefulness of the test scores obtained from the test. Short tests typically produce imprecise competency score estimates, and lead to competency decisions which prove to be inconsistent across parallel-form administrations (or retest administrations). (An examinee competency score is the proportion of items in the pool of items defined by a domain specification that the examinee can answer correctly. A competency score estimate is obtained by administering a sample of items to the examinee and calculating his/her proportion-correct score.)

Three factors should be considered in making decisions about the number of items:

1. the relationship between number of test items and the importance placed upon the particular competency,
2. the relationship between the number of test items and the minimum acceptable level of test score reliability,
3. the relationship between the number of items and available testing time.

In terms of factor one, it may be the case that some competencies are more important relative to the goals of the competency testing program than others. If the test developer plans for the test to cover multiple competencies, he/she should then plan, when drawing samples of items from each domain of items "keyed" to a competency, to more heavily sample the

most important competencies.

In reference to factor two, the relationship of the number of test items to minimum reliability requirements, guidelines are not readily available. The Spearman-Brown formula, which relates test length to reliability, is reasonable to use only with norm-referenced tests. Similar relationships need to be developed for competency tests. The following procedure should be helpful to those determining test length when competency score estimation is the problem of interest. The solution is a conservative one, i.e., test lengths determined by this method will be a little longer than they need to be to obtain the degree of precision required by the test developer. The formula[1] is:

$$\text{Test Length} = \frac{.25}{(\text{degree of precision})^2}$$

Ask yourself (or interested others): What degree of precision is required of the competency score estimates? Discuss the degree of precision question in the same way you would the standard error of measurement. A primary difference between the two is that competency score estimates are defined on a scale (0, 1).

At present we are working on tables relating test length to reliability when the test is used for making competent/incompetent decisions about examinees. The research is just beginning; thus, we are unable to report

[1]The formula can be derived from the binomial test model.

any results at this time. However, two points can be made. One, it is unlikely that fewer than five or six items measuring a competency will produce desired levels of reliability. Two, while no tables or formulas exist to connect test length to reliability (or consistency) of decision-making, reliability can be studied empirically after the administration of a pool of test items to a group of examinees (step 5b). "Post-hoc" test forms of varying lengths can be constructed and reliability estimates may be calculated, on the assumption that examinees would have responded in the same way had they been presented with the "parallel-forms" rather than a single large pool of test items. By varying the length of the forms and the formation of parallel-forms (i.e., which items are placed in which forms), the relationship between test length and reliability for a specified sample of examinees for a pool of test items measuring a particular domain specification can be studied.

The item selection process is straightforward provided the competency test developer has been careful in defining competencies and in constructing test items. That is, the test developer has to have been careful to define the size of his/her domain to be consonant with the test's purpose. If the purpose of testing is to make decisions on, for instance, broad school competencies, large domain sizes can be tolerated. If, however, the purpose of testing is to provide information for remedial instruction, a smaller domain size is needed. Popham (1978a) has offered some suggestions for ascertaining domain size. The critical point for item selection is that the domain be a reasonable size so that proper sampling from the domain can occur. If the domain is so large that it is difficult to see how to generate a set of items from the domain for the test, then the domain must be broken up into sub-domains and items generated for

those sub-domains. The sampling process should be clear for these sub-domains. Thus, it is critical that the domain be of a size that a set of items can be clearly constructed from the domain, and then the sampling process can be carried out without complications.

Having defined a domain size that is manageable for sampling is not enough; the test developer must also be careful to ascertain that all the items constructed for the domain do indeed "tap" the behavior specified. The items must adhere to the restrictions imposed on the domain specifications.

If the size of the domain is manageable for the sampling process and the test developer is sure that the items generated "tap" the specified behaviors, then the item selection process is straightforward. The test is constructed by taking either a random or stratified random sample of items from the domain.

One advantage of choosing representative sets of test items is that examinee test scores (or proportion-correct scores) provide "unbiased" estimates of their "true" competency scores. It is possible also to set standards and interpret examinee test performance relative to these standards. Unfortunately, when the number of test items is small (as is frequently the case), the consistency of decisions (competent/incompetent) across a retest administration or across a parallel-form administration of a test may be distressingly low. Increasing the number of test items measuring each competency is helpful but often it is not feasible to do so. One answer to the dilemma is as follows: When the primary purpose of the testing program is to make dichotomous decisions about examinees, a more effective test can be produced if test items from the available pool of test items measuring each competency are

selected based on their statistical properties. Specifically, if (say) a standard is set at 80%, it would be best to select test items which have p-values (item difficulty levels) in the region of .80 and which have the highest discrimination indices. A test constructed in this way will have maximum discriminating power in the region where decisions are being made and therefore more reliable and valid decisions will result. One possible drawback is that scores derived from the test cannot be used to make descriptive statements about examinee levels of performance on the competencies measured in the test. This is because test items measuring each competency are not necessarily a representative sample. In theory, there is at least one way to make descriptive statements about examinee levels of performance on the competencies measured by a test when non-random or non-representative samples of test items are chosen. It can be done by introducing concepts and models from the field of latent trait theory. The feasibility, however, of such an approach has not been tested.

Methods of Standard-Setting

Numerous researchers have catalogued many of the available standard setting methods (Glass, 1978a; Hambleton & Eignor, 1978; Hambleton et al., 1978; Jaeger, 1976; Millman, 1973; Meskauskas, 1976; Popham, 1978b; Shepard, 1976). If one fact is clear it is that all standard setting methods are arbitrary and this point has been acknowledged by nearly every contributor to the area. All of the methods are arbitrary because they involve judgments of one kind or another (for example, raters may be asked to identify test items which a minimally competent examinee

should be able to answer) and choices (for example, a choice of standard-setting methods must be made). But the "arbitrariness" of standard-setting methods is not a satisfactory reason for rejecting the methods. A quote from Popham (1978a) is especially appropriate here:

> Unable to avoid reliance on human judgment as the chief ingredient in standard-setting, some individuals have thrown up their hands in dismay and cast aside all efforts to set performance standards as arbitrary, hence unacceptable.
>
> But Webster's Dictionary offers us two definitions of arbitrary. The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is, "determinable by a judge or tribunal." The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is, "selected at random and without reason." In my estimate, when people start knocking the standard-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition.
>
> But the first definition is more accurately reflective of serious standard-setting efforts. They represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd. (p. 168)

In a recent review of the standard-setting literature, Hambleton and Eignor (1978) discussed six different sets of methods for setting standards. This review was an expansion of some earlier work by Millman (1973) and Meskauskas (1976). A discussion of the same sort, adding some standard-setting methods recently advanced (i.e., Jaeger, 1978; Zieky and Livingston, 1977), would perhaps prove helpful, if only to identify the more than twenty methods advanced to date. Such a discussion of methods will not be presented here, however, because a large number of these methods do not appear to be useful for setting standards in minimum competency testing programs. Those methods that appear to us to be applicable will be discussed in some detail. Also, a number of comparisons will be made in this "sifting out" of relevant methods, the first being the useful distinction made by Meskauskas (1976) between continuum and state models.

Continuum and State Models

The basic difference between continuum and state models has to do with the underlying assumption made about ability. According to Meskauskas, two characteristics of continuum models are:

1. Mastery is viewed as a continuously distributed ability or set of abilities.
2. An area is identified at the upper end of this continuum, and if an individual equals or exceeds the lower bound of this area, he/she is termed a master.

State models, rather than being based on a continuum of mastery, view mastery as an all-or-none proposition (i.e., either you can do something or you cannot). Three characteristics of state models are:

1. Test true-score performance is viewed as an all-or-nothing state.
2. The standard is set at 100%.
3. After a consideration of measurement errors, standards are often set at values less than 100%.

There are at least three methods for setting standards that are built on a state model conceptualization of mastery. The models take into account measurement error, deficiencies of the examination, etc., in "tempering" the standard from 100%. These methods have been referred to by Glass (1978a) in his review of methods for setting standards as "counting backwards from 100%." State model methods advanced to date include the mastery testing evaluation model of Emrick (1971), the true-score model of Roudabush (1974), and some recently advanced statistical models of Macready and Dayton (1977). However, since state models are somewhat less usefulness than continuum models in elementary and secondary school minimum competency testing programs, they will not be considered further in this paper. Our failure to consider them further in this paper, however, should not be interpreted as a criticism of this general approach to standard-setting. The approach seems to be especially applicable with many performance tests.

Traditional and Normative Procedures

Before discussing further the various continuum models of standard setting, two other models for standard-setting should be mentioned. These methods, which seem to have limited value in setting minimum competency standards, have been referred to by a variety of names. We will call them "traditional standards" and "normative standards."

Traditional standards are standards that have gained acceptance because of their frequent use. Classroom examples include the 90 to 100 percent is an A, 80 to 89 percent is a B, etc. It appears that from time to time such methods have been used in setting standards for minimum competency tests.

"Normative" standards refer to any of three different uses of normative data, two of which are, at best, questionable. In the first method, use is made of the normative performance of some external "criterion" group. As an example, Jaeger (1978) cites the use of the Adult Performance Level (APL) tests by Palm Beach, Florida schools. Test performance of groups of "successful" adults were used to set competency standards for high school students. The notion is that the test performance of "successful" adults provides a basis for setting standards for high school students. Such a procedure can be criticized on a number of grounds. Jaeger (1978) points out that society changes, and that standards should also change. Standards based on adult performance may not be relevant to high school students. Shepard (1976) points out that any normatively-determined standard will

immediately result in a multitude of counterexamples. Further, Burton (1978) points out that relationships between skills in school subjects and later success in life is not readily determinable, hence, observing the degree of achievement on the test of some "successful" norm group makes little sense. Jaeger (1978) goes on to say: "There are no empirically tenable "survival" standards on school-based skills that can be justified through external means."

A second way of proceeding with normative data is to make a decision about a standard based solely on the distribution of scores of examinees who take the test. Such a procedure circumvents the "minimum test score for success in life" problem, but the procedure is still not useful for setting standards. For instance, Glass (1978a) cites the California High School Proficiency Examination, where the 50th percentile of graduating seniors served as the standard. What can be said of a procedure where whether or not an individual passes or fails a minimum competency test depends upon the other individuals taking the test? In the California situation, the standard was set with no reference at all to the content of the test or the difficulty of the test items.

The third use of normative data discussed in the literature concerns the supplemental use of normative data in setting a standard. Shepard (1976), Jaeger (1978), and Conaway (1976, 1977) all favor such a procedure. Recently Jaeger (1978) advanced a standard setting method which requires judges to make judgments on item content. In his method, Jaeger calls for incorporation of some tryout test data

to aid judges in reconsidering their initial assessments. Shepard (1976) makes the following point:

> Expert judges ought to be provided with normative data in their deliberations. Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. . .of course, the norms are not automically the standards. Experts still have to decide what "ought" to be, but they can establish more reasonable expectations if they know what current performance is then if they deliberate in a vacuum.

We agree with Jaeger, Conaway, and Shepard about the usefulness of normative data when used in conjunction with a standard setting method.

Consideration of Several Promising Standard Setting Methods

Other methods for setting standards to be discussed in this paper are either built on a continuum model of ability or some other unexpressed model. For convenience, the methods under discussion were organized into three categories or models. These models and methods are presented in Figure 3. The models are labelled "judgmental," "empirical," and "combination." By judgmental is meant that data are collected from judges for setting standards, or a judgment is made about the presence or lack of a variable (for instance, guessing) that would effect the standard. Empirical methods require the collection of examinee response data to aid in the standard-setting process.

Figure 3. A classification of models and methods for setting standards.

Judgmental Models		Combination Models		Empirical Models[1]	
Item Content	Guessing	Judgmental-Empirical	Educational Consequences	Data—Two Groups	Data-Criterion Measure
Nedelsky (1954)	Millman (1973)	Contrasting Groups (Zieky and Livingston, 1977)	Block (1972)	Berk (1976)	Livingston (1975)
Modified Nedelsky (Nassif, 1978)		Borderline Groups (Zieky and Livingston, 1977)			Livingston (1976)
Angoff (1971)					Huynh (1976)
Modified Angoff (ETS, 1976)					Van der Linden and Mellenbergh (1977)
Ebel (1972)					
Jaeger (1978)					
		Bayesian Methods		Decision-theoretic[2]	
		Hambleton and Novick (1973)		Kriewall (1972)	
		Novick, Lewis, Jackson (1973)			
		Schoon, Gullion Ferrara (1978)			

[1]Involve the use of examinee response data.

[2]In addition, there are a number of decision-theoretic models that deal with test length considerations. These are also applicable to cut-off score determination (see, for example, Millman, 1974).

Empirical Methods

A number of methods have been developed that require a criterion measure, performance measure, or true ability continuum. Livingston (1975) has presented a procedure based on linear or semi-linear utility functions in which he looks at the use of these functions in viewing the effects of decision-making accuracy based upon a particular performance standard. Livingston (1976) presented a method for choosing standards by stochastic approximation techniques. Once again, the procedure depends upon a performance measure, and a standard set on that measure. Huynh (1976) bases a standard-setting method for a competency test to an external criterion. Finally, the work of Van der Linden and Mellenbergh (1977) depends upon the existence of a latent ability variable that can be dichotomized into two categories, labeled "competent" and "incompetent." The standard is then set based upon a risk or expected loss function.

These methods have only been briefly mentioned because they all are difficult to apply in practice since they require a criterion variable upon which success and failure (or probability of success and failure) can be defined. External criterion variables which would be appropriate for validating high school certification tests are going to be difficult to gain agreement about and probably very difficult to measure. For example, how would you go about defining "life success" and measuring it? Reading experts, for instance, are not going to have the same idea about what the minimally competent person can read. Should he/she be able to read at 12th grade level, or the 8th grade level? For example, Jaeger (1978) has noted, "Educators would no

sooner agree on the proportion of New York Times front page passages eleventh-graders should be able to comprehend and explain, then they would the proportion of multiple-choice test items those eleventh-graders should answer correctly, so as to be labeled "minimally competent." Thus, the jist of this reasoning is that if agreement can't first be reached on the criterion measure, then this isn't going to aid in setting standards on the test. Given the situation, one may want to go ahead and try to set the standards on the test without considering criterion-measures. Such a recommendation seems especially relevant for promotion and high school certification examinations.

One example of a decision-theoretic procedure is due to Kriewall (1972). This procedure is based upon the definition of (usually) two mastery states. The standard on the test is then selected as the point that minimizes "false positive" and "false negative" errors in the classifying of individuals into the defined mastery states. Once again, the problem with this method is evident. The mastery categories would in this case be "competent" and "incompetent," and they are essentially undefined. Until people can agree on a definition of "competence" in a given situation, it is not possible to use the method. You cannot minimize errors of prediction if the categories to be predicted can't be established. Jaeger (1978) has noted that many of the methods allow for different utilities to be associated with false positive and false negative errors, in this case passing the "minimally incompetent" person or failing the "minimally competent" person. However, there are no guidelines for establishing these utility values, so another problem exists with the methods.

Finally, Berk (1976) has presented a method that is very similar to the decision-theoretic methods just discussed. Rather than setting the mastery states arbitrarily and observing the probabilities of false positive and false negative errors on the criterion, Berk suggests the optimal standard be based on response data from samples of instructed and uninstructed students. Berk offers a number of procedures to be used in conjunction with his method. We feel that the procedure holds great merit for classroom instructional settings, and have devoted a great deal of time to a discussion of it in our recent review (Hambleton & Eignor, 1978). The problem involved with using the procedure for setting standards on minimum competency tests is immediately evident. There is no simple way of establishing groups of students instructed on the competencies included in the test and groups which have not had instruction. Other extreme groups might be formed (for example, "successful" adults and "unsuccessful" adults) and their performances compared on the test for the purpose of setting an optimum standard. Clearly though, results from such comparisons can be explained in numerous ways and therefore results of this sort have limited practical value.

Block (1972) introduced a method referred to as "educational consequences." In this method one looks at the effect the setting of a standard of proficiency has on future learning or other related cognitive or affective success criteria. Block conducted an experimental study to consider the effect of different standards on several outcome measures. The standard for which the valued outcome is maximal (it could be a combination of valued outcomes) becomes the standard the next time the test is used.

Glass (1978a) has likened this approach to the general approach of operations research and the concern for maximizing a valued commodity by finding an optimum point on a mathematical curve. Glass has pointed out the need for non-monotonic curves relating performances to the valued outcomes, which are not likely to be the case, in order to locate a maximum. Glass also talks about the problem of how to weight individual outcomes to form a composite outcome. There is yet another problem, perhaps even more serious than the non-monotonicity problem. One can't maximize a valued outcome if the outcome can't be defined in any reasonable manner. In sum, to utilize Block's method, there would have to be concensual agreement on what a valued outcome of being competent is. This would seem to be as difficult a task as trying to get people to define behaviors associated with minimum competency.

Finally, Millman (1973) has suggested that standards be adjusted for the effects of guessing. A systematic error is introduced when the test item format allows a student to answer items correctly by guessing. Millman suggests raising the standard to take into account the expected contribution attributed to pure guessing. Educational Testing Service has corrected the standards on the NTE exams and the Insurance Licensing Exams to take care of guessing. The problem here is that for minimum competency tests, pure random guessing rarely occurs and because of this, the effects of raising the standards as if it had, is unknown. Clearly, more work in this area is needed.

Bayesian methods will not be discussed because they allow standard setters to augment the setting of standards with prior information and/or group information on the examinees in question. Bayesian

methods also provide a statement of probability concerning an examinee's true level of competency exceeding the standard. To use the Bayesian methods, however, a standard must first exist. Any one of the methods to be discussed next could be used to set the standard.

Judgmental Models

What follows is a brief discussion of several judgmental methods. Comments, comparisons and recomendations for use will be offered also. Table 1 provides a summary of some of the similarities and differences among the methods.

i. Nedelsky's Method

In Nedelsky's method, judges are asked to view each question in a test with a particular criterion in mind. The criterion for each question is, which of the response options should the minimally competent student (Nedelsky calls them D-F students) be able to eliminate as incorrect. The minimum passing level (MPL) for that question then becomes the reciprocal of the remaining alternatives. For instance, if on a 5 alternative multiple choice question, a judge feels that a minimally competent person could eliminate two of the options, then for that question, MPL = $\frac{1}{3}$. The judges proceed with each question in a like fashion, and upon completion of the judging process, sum the values for each question to obtain a standard on the total set of test items. Next, the individual judge's standards are averaged. The average i denoted $\#_o$.

Table 1

A Comparison of Several Standard Setting Methods

Question	Judgmental						Combination	
	Nedelsky	Modified Nedelsky	Angoff	Modified Angoff	Ebel	Jaeger	Contrasting Groups	Borderline Group
1. Is a definition of the minimally competent individual necessary?	Yes	Yes	Yes	Yes	Yes	No	No	Yes
2. What is the nature of the rating task—or items, or individuals?	Items	Items	Items	Items	Items	Items	Individuals	Individuals
3. Are examinee data needed?	No	No	No	No	No	No	Yes	Yes
4. Do judges have access to the items?	Yes	Yes	Yes	Yes	Yes	Yes	Usually, but don't need to	Usually
5. Are the judgments made in a group setting or individual setting?	Both	Both	Both	Both	Both	Both	Individual	Individual
Choices of methods to use for setting standards on minimum competency tests.				✓		✓	✓	

Nedelsky felt that if one were to compute the standard deviation of individual judge's standards, that this distribution would be synonomous with the (hypothesized or theoretical) distribution of the scores of the borderline students. This standard deviation, σ, could then be multiplied by a constant K, decided upon by the test users, to regulate how many (as a percent) of the borderline students pass or fail. The final formula then becomes:

$$\hat{\pi}_o = \hat{\pi}_o + K\sigma .$$

How does the $K\sigma$ term work? Assuming an underlying normal distribution, if one sets K=1, then 84% of the borderline examinees will fail. If K=2, then 98% of these examinees will fail. If K=0, then 50% of the examinees on the borderline should fail. The value for K is set by (say) a committee prior to the examination.

The final result of the applications of Nedelsky's method will be an absolute standard. This is because the standard is arrived at in a manner independent of the score distributions of any reference group. In fact, the standard is arrived at prior to application of the test to the group one is concerned about testing. However, while the standard can be called absolute, there is a great deal of judgment involved in applying the method.

ii. Modified Nedelsky

Nassif (1978,, in setting standards on the competency-based teachers education and licensing systems in Georgia, utilized a modified Nedelsky procedure to set standards. A modification of the Nedelsky method was needed to handle effectively the volume of items in the program. In the modified Nedelsky task, the entire item (rather than each distractor) is examined and classified in terms of two levels of examinee competence. The following question was asked about each item: "Should a person with minimum competence in the teaching field be able to answer this item correctly?" Possible answers were "yes," "no," and "I don't know." Agreement among judges can be studied by a simple comparison of the ratings by judges to each item. A standard may be obtained by averaging the number of "yes" responses given by judges to the set of test items.

iii. Ebel's Method

Ebel (1972) goes about arriving at a standard in a somewhat different manner, but his procedure is also based upon the test questions rather than an "outside" distribution of scores. Judges are asked to rate items along two dimensions: Relevance and difficulty. Ebel uses four categories of relevance: Essential, important, acceptable and questionable. He uses three difficulty levels: Easy, medium and hard. These categories then form (in this case) a 3 x 4 grid. The judges are next asked to do two things:

1. Locate each of the test questions in the proper cell, based upon relevance and difficulty,
2. Assign a percentage to each cell; that percentage being the percentage of items in the cell that the minimally-qualified examinee should be able to answer.

Then the number of questions in each cell is multiplied by the appropriate percentage (agreed upon by the judges), and the sum of all the cells, when divided by the total number of questions, yields the standard.

Three comments can be made about Ebel's method that should be sufficient to convince people to be careful in using it. One, Ebel offers no prescription as to what the number or type of descriptions should be along the two dimensions. This is left up to the judgment of the individuals judging the items. It could likely be the case that a different set of dimensions applied to the same test could yield a different standard. Two, the process is based upon the decisions of judges, and while the standard could be called absolute in that it is referenced to no other distribution, it can't be called an "objec-

tive" standard. Three, a point about Ebel's method has been offered by Meskauskas (1976):

> In Ebel's method, the judge must simulate the decision process of the examinee to obtain an accurate judgment and thus set an appropriate standard. Since the judge is more knowledgeable than the minimally-qualified individual, and since he is not forced to make a decision about each of the alternatives, it seems likely that the judge would tend to systematically over-simplify the examinees task . . . Even if this occurs only occasionally, it appears likely that, in contrast to the Nedelsky method, the Ebel method would allow the raters to ignore some of the finer discriminations that an examinee needs to make and would result in a standard that is more difficult to reach. (p. 138)

iv. Angoff's Method

When using Angoff's technique, judges are asked to assign a probability to each test item directly, thus circumventing the analysis of a grid or the analysis of response alternatives. Angoff (1971) states:

> . . .ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (p. 515)

v. Modified Angoff

ETS (1976) utilized a modification of Angoff's method for setting standards. Based on the rationale that the task of assigning probabilities may be overly difficult for the items to be assessed (National Teacher Exams) Educational Testing Service instead supplied a seven point scale on which certain percentages were

The following scale was offered:

5 20 40 60 75 90 95 DNK

where "DNK" stands for "Do Not Know."

ETS has also used scales with the fixed points at somewhat different values; the scales a - consistent though in that seven points are given to choose from. The National Teacher Exam program specified 60 as the center point since the average of percent correct on past exams centered around 60%. The other options were then spaced on either side of 60.

vi. Jaeger's Method

Jaeger (1978) recently presented a method for standard-setting on the North Carolina High School Competency Test. Jaeger's method encorporates a number of suggestions made by participants at a 1976 NCME annual meeting symposium presented in San Francisco by Stoker, Jaeger, Shepard, Conaway, and Haladyna; it is iterative, based on judges from a variety of backgrounds, and employs normative data. Further, rather than asking a

question involving "minimal competence," a term which is hard to operationalize, and conceptualize, Jaeger questions are instead:

> "Should every high school graduate be able to answer this item correctly?" "____Yes, ____No." and
> "If a student does not answer this item correctly, should he/she be denied a high school diploma?"
> "____Yes, ____No."

After a series of iterative processes involving judges from various areas of expertise, and after the presentation of some normative data, standards determined by all groups of judges of the same type are pooled, and a median computed. The minimum median across all groups is selected as the standard.

Comparisons Among Judgmental Models

We are aware of two studies that compare judgmental methods of setting standards; one study was done in 1976, the other is presently underway at ETS.

In 1976, Andrew and Hecht carried out an empirical comparison of the Nedelsky and Ebel methods. In the study, judges met on two separate occasions to set standards for a 180 item, four options per item, exam to certify professional workers. On one occasion the Nedelsky method was used. On a second occasion the Ebel method was used. The percentage of test items that should be answered correctly by the minimally competent examinee was 69% by the Ebel method and 46% by the Nedelsky method.

Glass (1978a) described the observed difference as a "startling finding." Our view is that since directions to the judges were different, and procedures differed, we would not expect the results from these two methods to be similar. The authors themselves report:

> It is perhaps not surprising that two procedures which involve different approaches to the evaluation of test items would result in different examination standards. Such examination standards will always be subjective to some extent and will involve different philosophical assumptions and varying conceptualizations. (p. 49)

Ebel (1972) makes a similar point:

> . . . it is clear that a variety of approaches can be used to solve the problem of defining the passing score. Unfortunately, different approaches are likely to give different results. (p. 496)

Possibly the most important result of the Andrew-Hecht study (and this result was not reported in the Glass paper) was the high level of agreement in the determination of a standard using the same method across two teams of judges. The difference was not more than 3.4% with each method. Data of this kind addresses a concern raised by Glass (1978a) about whether judges can make determinations of standards consistently and reliably. In at least this one study, it appears that they could. From our interactions with staff at ETS who conduct teacher workshops on setting standards, we have learned that teams of teachers working with a common method obtain results that are quite similar. And this result holds across tests in different subject matter areas and at different grade levels. We have observed the same result in my work. Of course, there are conditions which must be met if agreement between judges is to be obtained.

Donald Rock at ETS is presently pursuing research on the use of the Nedelsky and Angoff methods for standard setting on Real Estate Certification Examinations. The results of this study, which have not been released, should shed some light on the comparability of the two judgmental procedures used most frequently to date.

Combination Models

Two very attractive methods which we will refer to as combination methods will be considered next. They were first proposed by Zieky and Livingston (1977). In these methods, judges are asked to make judgments of the mastery levels of students, rather than about test items. Teachers would be the most reasonable choice as the judgements to be made concern a student's level of mastery of the area being tested. They must identify students as "adequate," "inadequate," or "borderline" relative to the content area of interest. The task of imagining a minimally competent student or group of students is circumvented, and for this reason alone, these methods are in favor. What follows is a very brief description of the two methods. Readers interested in a more thorough discussion, along with helpful hints for applying the methods should refer to Zieky and Livingston (1977).

1. Borderline-Group Method

Once teachers have identified a group of students whose achievement is judged to be borderline in the area being tested, the test is administered and the median test score for this group becomes an estimate of the standard.

ii. Contrasting-Group Method

Once teachers have identified groups of students they are sure are definite masters or non-masters of the skills being measured by the test, the test is given, and score distributions plotted for each group. The intersection of the score distributions becomes the first estimate of the standard. This can then be adjusted up or down to obtain the required balance between "false-positive" and "false-negative" errors.

The Contrasting-Groups Method is very similar to a method offered independently by Berk (1976). Berk assumes that the students being assessed are masters or non-masters on the basis of whether or not they have been instructed on the content measured by the test. On the other hand, Zieky and Livingston ask teachers to judge the students on the skills in the test. The major point to be made is that procedures offered by Berk for analysis of the data (a validity coefficient, utility analysis) are also applicable with the Contrasting-Groups Method.

Some Final Remarks

Our review of the literature identified a variety of methods for setting standards. However, when one tries to apply these methods to minimum competency tests, problems arise. The empirical methods require an external criterion measure which often is very hard to obtain. When external criterion measures can be obtained, methods proposed by Livingston (1975, 1976), Huynh (1976), Van der Linden and

Mellenbergh (1976), Hambleton and Novick (1973), Kriewall (1972), and Berk (1976) will be very useful. At the present time, the best methods for setting standards on elementary and secondary school minimum competency tests are those that deal directly with the test. These methods do require judgments, and arbitrary standards are obtained. Given the state of affairs in the area of standard settings, however, we can only suggest that any method be carefully used, and that the expressed concerns and recommendations of researchers on this topic (for example, Conaway, 1976, 1977; Glass, 1978a, 1978b; Haladyna, 1976; Jaeger, 1976; Shepard, 1976) be carefully considered.

Suggestions for Future Research and Development

In our paper we have introduced a model for developing and validating competency tests and we have considered several methods of setting standards. In this final section, several suggestions for future research and development will be offered. The suggestions are organized by the two major topics of the paper:

Competency Test Development and Validation

1. Technical guidelines are needed for the evaluation of competency tests and test manuals. The AERA/APA/NCME Test Standards have some value for this purpose but are incomplete and what relevant material there is in the Test Standards is scattered throughout a 75-page document.

2. Usable guidelines for determining test lengths (number of test items/competency) are not available. There are several technical contributions on the problem in the literature but the contributions are rather complex mathematically and therefore not readily usable by practitioners.

3. More needs to be learned about the development and validation of performance tests since many of the competencies being discussed by designers of competency testing programs can be measured best by performance tests.

4. Considerable attention should be given to the development of guidelines for writing domain specifications. Also their use in developing competency tests and in facilitating proper test score interpretations should be evaluated. Finally, the merits of domain specifications in comparison with other approaches for describing item pools (for example, algorithmic transformation of sentences from written instruction into test items, facet designs and others) should be considered.

5. Latent trait models are being used in the development of some norm-referenced tests and in the interpretation of norm-referenced test scores. The models appear to have potential also for use with competency tests. Equating of scores from one form of a competency test to another is one of the more promising applications. Clearly, more research on the feasibility of using latent trait models with competency tests is called for.

Standard-Setting Methods

1. There is a need for considerably more work on both the moral and technical issues involved in standard-setting.

2. There needs to be considerably more study of the term, "minimally competent" because if the term is better understood, it may be possible to link existing standard-setting methods to the intended meaning or meanings of the term and thereby greatly facilitate the selection of a standard-setting method (or the development of new methods).

3. For "acceptable" standard-setting methods, implementation strategies need to be developed, evaluated, and made ready for wide use. At present there are few guidelines or procedural steps available for applying any of the standard-setting methods. (An exception to this is the excellent work by Popham [1978b] and Zieky and Livingston [1977].)

The purposes of competency testing programs can only be accomplished (1) if quality competency tests are constructed and (2) if scores derived from the tests are interpreted and used correctly. We hope our paper will facilitate the accomplishment of both objectives.

References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Berk, R. A. Criterion-referenced test item analysis and validation. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, 1978.

Block, J. H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-190.

Brickell, H. M. Seven key notes on minimum competency testing. Phi Delta Kappan, 1978, 59, No. 9 (May), 589-592.

Burton, N. Societal standards. Journal of Educational Measurement, 1978, 15, in press.

Conaway, L. E. Discussant comments: Setting performance standards based on limited research. Florida Journal of Educational Research, 1976, 18, 35-36.

Conaway, L. E. Setting standards in competency-based education: Some current practices and concerns. Paper presented at the annual meeting of NCME, New York, 1977.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Educational Testing Service. Report on a study of the use of the National Teachers Examination by the State of South Carolina. Princeton, NJ: Educational Testing Service, 1976.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, in press. (a)

Glass, G. V. Minimum competence and incompetence in Florida. Phi Delta Kappan, 1978, 59, No. 9 (May), 602-605. (b)

Haladyna, T. Comments: Measurement issues related to performance standards. Florida Journal of Educational Research, 1976, 18, 33-34.

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, in press.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1978.

Hambleton, R. K., & Fitzpatrick, A. Review techniques for criterion-referenced test items. Manuscript in preparation.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Haney, W., & Madaus, G. Making sense of the competency testing movement. National Consortium on Testing, Staff Circular No. 2. Cambridge, MA: The Huron Institute, 1978.

Hively, W., Patterson, H. L., & Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Jaeger, R. M. Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 1976, 18, 22-27.

Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the 1978 spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of AERA, Chicago, 1972.

Livingston, S. A. A utility-based approach to the evaluation of pass/fail testing decision procedures. Report No. COPA-75-01. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1975.

Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. Report No. COPA-76-02. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1976.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Co., 1974.

Nassif, P. M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of NCME, Toronto, 1978.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Pipho, C. Minimum competency testing in 1978: A look at state standards. Phi Delta Kappan, 1978, 59, No. 9 (May), 585-587.

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

Popham, W. J. Setting performance standards. Los Angeles: Instructional Objectives Exchange, 1978. (b)

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

Roudabush, G. E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of NCME, Chicago, 1974.

Schoon, C. G., Cullion, C. M., & Ferrara, P. Credentialing examinations, Bayesian statistics, and the determination of passing points. Paper presented at the annual meeting of APA, Toronto, 1978.

Shepard, L. A. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 23-32.

Van der Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1977, 1, 593-599.

Zieky, M. J., & Livingston, S. A. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, NJ: Educational Testing Service, 1977.