

F I N L A N C E

The Finnish Journal of Language Learning and Language Teaching

Vol. III 1984

Edited by

Liisa Korpimies

Language Centre for Finnish Universities

University of Jyväskylä · Finland

ISSN 0359-0933

Sauli Takala
Kasvatustieteen tutkimuslaitos
Jyväskylän yliopisto

SOME PERSPECTIVES ON CRITERION-REFERENCED MEASUREMENT

Introduction

There are several reasons for the recent interest in a direction in measurement and evaluation which frequently is referred to as "criterion-referenced" measurement. By criterion-referenced measurement is usually meant a type of measurement that is deliberately constructed to yield scores that are directly interpretable in terms of specified performance standards related to specific classes or domains of tasks (Glaser, 1963; Glaser and Nitko, 1971; Popham, 1978).

Classical test theory, which formed the basis of the psychological study of individual differences, abilities, etc., has had only tenuous links with learning theory (Cronbach, 1957; Glaser and Nitko, 1971). It has not, therefore, been much interested in aptitude-treatment interaction. Serious practical work on truly individualized instruction on a scientifically sound basis is of relatively recent origin. When interest in adaptive instructional systems grew in the 1960's, it became evident that there was a need for tests that are very sensitive to the content of individualized programs (Glaser, 1963).

Another reason why criterion-referenced measurement became a topic of growing interest is that when increasingly large sums of money became available through national budgets for experimental educational programs, it became a standard practice to require a research-based evaluation of the effectiveness of the programs. Since the contents of such programs were often based on new ideas about content and treatment, it was to be expected that standardized tests would not be considered very suitable to measure the effects obtained. More program-specific tests were needed.

A third reason for increased interest in criterion-referenced measurement derives from the growing demands for proof that national educational

systems are working in a satisfactory way and that the money allocated to cover educational costs is well spent. When the performance of national systems are assessed, there is a growing interest to make sure that tests measure what has been taught and that test results tell the general public what students can do and what they cannot do.

A fourth reason is more closely related to decisions concerning individual students. When almost automatic promotion from grade to grade has become a pattern with the introduction of comprehensive-type educational systems, there has been growing concern that students may be promoted without having learned the knowledge and skills needed in the subsequent grade ("social promotion"). If school systems decide to adopt a stricter promotion policy, it is important that the amount of the risk of making false decisions is minimized. Program-specific tests are a useful tool in administering such a promotion policy.

This paper will first review some major sources of criterion-referenced measurement and describe briefly some alternative conceptualizations. Some comparisons are made between criterion-referenced and norm-referenced measurement. After that, stages in CRM are described with major emphasis on methods of content specification and the construction and selection of items. The paper then moves to discuss standard setting as an issue in CRM. This is followed by a review of how validity and reliability are treated in CRM. The paper concludes with a brief account of the uses of CRM and of current problems and issues in CRM.

1. Criterion- and Norm-Referenced Measurement

It was estimated that there were some 600 references on criterion-referenced measurement towards the end of the 1970's. Practically all of them were published during that decade. Yet, criterion-referenced measurement is not such a new idea.

E.L. Thorndike wrote about the difference between absolute and relative measurement some seventy years ago. Around 1950 Vahervuo in Finland carried out several studies on absolute and relative grading and on their theoretical basis. Still, it was in an article by Robert Glaser in 1963 that the term "criterion-referenced test" was introduced. The idea was favorably received but it did not lead to further work until in 1969 when Popham and Husek took up the concept and explicated further some of its implications.

Programmed learning and the behavioral objectives movement (e.g. Mager, 1962) were a major source in the emergence of criterion-referenced measurement. Carefully outlined teaching programs will not lead to a normal distribution of scores if the programs are, indeed, effective. There should be a high percentage of high scores and a decrease in variance. The latter is problematic for classical test theory, because most of its indices rely heavily on variance. Thus, it seemed necessary to conclude that variance-based estimates of test reliability are less appropriate in mastery-type instructional programs since they would unjustifiably label criterion-referenced tests as being of low reliability. New approaches were clearly needed (Popham and Husek, 1969).

Another major source, which is related to programmed learning and individualized learning programs, is the work done to discover learning hierarchies and curriculum (task) hierarchies (Gagne et al, 1962; Resnick, 1967). This work revealed that the testing of learning outcomes requires a thorough analysis of the subject matter as a preliminary step to item construction.

Criterion-referenced testing has been defined in a number of ways. According to Berk (1980), at least fifty different definitions have been proposed since Glaser's first paper. Perhaps the most concise definition has been suggested by Popham (1978, p. 93): "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain." This means that the interpretability of the test result is of primary concern. Whereas in norm-referenced measurement an individual's test score derives its meaning mainly from its relationship to the scores of other examinees (relative interpretation), the scores on a criterion-referenced test derive their meaning from the scores' relationship to a class or domain of tasks (absolute interpretation). Thus a domain score can be interpreted in terms of what an individual can do and what he cannot do and it also indicates what proportion of all possible tasks (items) of the whole item universe the individual could have solved if they were administered to him rather than only a sample of them. A domain score lends itself to absolute interpretations and can be used both for qualitative and quantitative descriptions (what is mastered and how much is mastered).

Several terms for this kind of testing have been proposed within the criterion-referenced movement. Ebel (1962) proposed a term "content-standard test" to describe a test which produces test scores which indicate what percentage of a systematic sample of defined tasks a person has solved correctly. Osburn (1968) used the term "universe-defined test" to refer to a test which produces an unbiased estimate of his score in an explicitly defined item content universe. Hively (1962) prefers the term "domain-referenced test" as a less ambitious term than universe-defined test. Carver (1974) has advocated the use of edumetric (rather than traditional psychometric) tests to measure within-individual growth (competence) instead of between-individual differences (ability, intelligence).

The term "objectives-based test" has sometimes been used as a near-synonym for criterion-referenced tests. If the items are simply derived from behavioral objectives without a strictly predetermined procedure, however, objective-based tests do not lend themselves to criterion-referenced interpretation.

The term "mastery test" has been derived mainly from the mastery learning system developed by Bloom (1968, 1971), largely on the basis of the model of school learning proposed by Carroll (1963). The main purpose of mastery tests is to help in the classification of students as masters or nonmasters of an objective in order to facilitate the management of an individualized teaching program.

If one were shown a test which only contained the instructions to students and the test items, it would be difficult to say whether the test is a criterion-referenced test or a norm-referenced test. In order to be able to make that decision it is necessary to know how the test was produced. It is in the work prior to the assembly of a test that most of the effort needs to be spent in producing a criterion-referenced test. Differences between two forms of criterion-referenced testing (domain-referenced and mastery tests) and norm-referenced testing are summarized in Table 1. The first five stages in the development of tests refer to the planning stage and the rest to the technical aspects of tests and their uses.

TABLE 1. Characteristics of Two Types of Criterion-Referenced Tests and of Norm-Referenced Tests (adapted from Millman, 1974, and Berk, 1980).

Stages of Development	Alternative Conceptualizations		
	Criterion-Referenced Testing		Norm-Referenced Testing
	Domain-Referenced	Mastery	
1. Specification of Content Domain	Maximum specification of content limits <u>Methods:</u> 1. Item transformations 2. Mapping sentences 3. Algorithms 4. Item forms 5. Amplified objectives 6. Test specifications	Content limits only partially specified <u>Methods:</u> Instructional and behavioral objectives	Content limits only partially specified <u>Methods:</u> Instructional and behavioral objectives
2. Item Construction	Generation rules	Traditional rules	Traditional rules
3. Specification of Item Domain	Infinite or finite item universe	Infinite ?	Infinite ?
4. Item Analysis	Purpose to detect flawed items <u>Methods:</u> 1. A priori judgement of item-objective congruence by subject matter experts 2. A posteriori computation of item statistics	Purpose to detect flawed items <u>Methods:</u> ?	Purpose to select items <u>Methods:</u> A posteriori computation of item statistics
5. Item Selection from Item Universe	Random	Nonrandom (?)	Nonrandom

- Table 1 (cont.).

Stages of Development	Alternative Conceptualizations		
	Criterion-Referenced Testing		Norm-Referenced Testing
	Domain-Referenced	Mastery	
6. Cut-off Score Selection	Optional	Required	Required (?)
7. Validity	Content	Content	Criterion-related
	Construct	Criterion-related	
	Decision	Construct Decision	
8. Reliability	1) Consistency of decisions (\hat{p}_0, k)	Consistency of decisions (\hat{p}_0, k)	Traditional procedures (based on correlation)
	2) Dependability ($\phi(\lambda)$)		
	3) Error of measurement or estimate around domain score		
9. Score Interpretation	Performance in relation to domain (level of functioning)	Performance in relation to required level of mastery	Performance in relation to other examinees
	Performance in relation to required level of mastery		
10. Item and Test Variance	Not required	Not required	Required

2. Stages in Test Construction

2.1. Specification of Content

It is in the specification of the content domain that the greatest challenge and also the greatest merit of criterion-referenced testing lies. In traditional norm-referenced tests the content limits are only partially specified. Short instructional and behavioral objectives are used as the basis for item generation. As Bormuth (1970) and Anderson (1972), among others, have shown, there is so much room left for interpretation that the items may reflect the characteristics of the test constructor more than those of the instructional program. Too much room is left for creativity, which according to Popham (1978, 1980), is not as desirable as strict adherence to the content limits. Several methods have been proposed for making domain specification more adequate. These will be discussed below in some detail, since this is a crucial part of all criterion-referenced measurement.

Item Transformations

Bormuth (1970) has suggested that linguistic analysis based on transformational grammar could be used to make explicit the methods by which items are derived from statements of instructional objectives. Bormuth advocates operationalism as a way of introducing rigor into item construction and sees syntactic operations as a promising way to do this. His method is illustrated below. It shows some item transformations that have been performed on a sentence "The older sister put out the fire." Using syntactic transformations several comprehension questions could be asked about the sentence.

It seems obvious that Bormuth's method is a useful tool for generating items testing the comprehension of written and spoken discourse. Anderson (1972) provides some other examples of ways of generating questions to test discourse comprehension. One weakness of these methods is, however, that the emphasis is on sentence level operations rather than discourse level units. Recent work on discourse analysis by Halliday and Hasan, van Dijk, Meyer and others will be of use in moving from sentence to discourse-level testing.

Transformation Name	Question
Echo	The older sister put out the fire?
Tag	The older sister put out the fire, didn't she?
Yes-No	Did the older sister put out the fire?
Noun deletion	Who put out the fire? What did the older sister put out?
Noun modifier deletion	Which sister put out the fire?

Using these examples of item transformation, supply answers for Problem Set 2.

Problem Set 2
Item Transformations

The following statement appears as part of a paragraph in a science unit on balance scales: The heavier object is closer to the ground. Only items formed by the "yes-no" and "noun modifier deletion" transformations are to be used in a test to measure comprehension of this statement. What questions can be used?

1. Yes-No: _____
 2. Noun modifier deletion: _____
-

Answers:

1. Is the heavier object closer to the ground?
2. Which object is closer to the ground?

(Source: Millman, 1974)

Mapping Sentence

Mapping sentences are used in facet analysis developed by Guttman (1969). Facet analysis can be used to describe the boundaries and structure of a domain of testing conditions. Facets are those dimensions or characteristics on which items in a given domain can differ. Facet analysis was used by the present writer in 1980 in an attempt to conceptualize the domain of written composition for the IEA International Study of Written Composition. The first attempt is illustrated below. (For a later version, see Takala, 1982).

Millman (1978) also used facet analysis in his study of how the form and content of items are related to item difficulty.

Mapping Sentence for the Domain of Writing
Following Guttman's Facet Analysis Scheme

<p>A. <u>Activity</u></p> <p>1. Receive 2. Send</p>	<p>a/an</p>	<p>B. <u>Channel</u></p> <p>1. auditive 2. visual</p>	<p>message which deals with</p>	<p>C. <u>Content/topic</u></p> <p>1. self 2. school 3. home town 4. hobbies 5. 6.</p>	<p>and whose</p>	<p>D. <u>Communication Partner</u></p> <p>1. addressor 2. addressee</p>
<p>has/is</p>	<p>E. <u>Role relationship between addressor and addressee</u></p> <p>1. a higher social status 2. an equal social status 3. a lower social status 4. identical with addressor</p>	<p>and which is</p>	<p>F. <u>Degree of publicity/ formality</u></p> <p>1. private 2. semi-public 3. public</p>			
<p>consisting of</p>	<p>G. <u>Input-output relationship (stimulus-response)</u></p> <p>1. repetition of input 2. modification of input 3. internal input</p>	<p>and whose purpose is</p>	<p>H. <u>Function</u></p> <p>1. to preserve the message (documentative) 2. to inform (referential) 3. to persuade (emotive) 4. to describe (descriptive) 5. 6.</p>			

Different configurations of variables lead to different rhetorical modes (narrative, exposition, argumentation, etc.)

Examples:

A2 + B2 + C2 + D2 + E1 + F3 + G2 + H1 = a personal letter to a friend
A2 + B2 + C2 + D1 + E3 + F2 + G4 + H2 = a letter of application

Algorithms

The use of algorithms is closely related to facet analysis and mapping sentences. It also uses listing technique. The following example from Millman (1980) illustrates the use of algorithms in generating items.

The item generating process follows rules to ask for the cube of an integer between 4 and 7 and to produce four possible answers. The command MULTCHOICE letters the alternatives and randomizes them. The algorithm on line 40 generates the correct answer.

```

10 A=RANDOM (4,7)
20 A, SUPER (3), "EQUALS:"
30 FROM
40 A*A*A
50 A*3
60 A+3
70 A+30
80 A*10+3
90 A*A
100 (A=1)*A*A
110 MULTCHOICE
120 RIGHT 1
130 WRONG CHOOSE AT RANDOM

```

5^3 equals:
A. 53
B. 125
C. 8
D. 150 Answer: B

(Source: Millman, 1980)

Item Form

Perhaps the most sophisticated method of content specification is the so-called item form suggested by Hively (1968). Item forms serve two purposes (Hively et al, 1973): 1) they obviate the necessity to store individual items by substituting a set of written rules through which items can be generated when needed; and 2) they enable the relationship among items to be traced by giving clear specifications of relevant item characteristics. Thus there are two major parts in all item forms - one which tells how the items should be generated and another which describes their salient characteristics. One item form used by Hively and his colleagues in the Minnesota Mathematics and Science Teaching Project (MINNEMAST) is shown below.

ITEM FORM 16.14*

Comparing two objects on equal-arm balance and choosing a symbol to complete a statement of the weight relation.

GENERAL DESCRIPTION

The child is asked to compare the weights of two objects that may be (1) indistinguishable by feeling but easily distinguished on the balance, (2) indistinguishable even on the balance. In each of these situations, size varies as an irrelevant dimension. An equal-arm balance is available but instructions for its use are non-derivative. The child is asked to predict one of the three symbols ($>$, $<$, and $=$) and place it in the blank space provided between the two weight symbols.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for all Cells

The equal-arm balance is of similar construction to that used in MINIREAST Unit 16, made of Timberly, cardboard, string, a metal weight, and a foot ruler.

The objects are opaque, cylindrical bottles, identical except for weight (either 23 gm. or 25 gm.) and size (either 2" x 4" or 2 1/2" x 1 1/2"). Each is identified by a lower-case letter assigned at random.

The child is asked to complete a symbolic statement, corresponding to the weight relation, by choosing the correct relation symbol.

Distinguishing among Cells

Three weight relations (detectable by balance only, not by feeling or "feel") defined in terms of the location of the objects when placed in front of the child:

left > right; left < right; left = right.

Three size relations:

left > right; left < right; left = right.


CELL MATRIX

Weight Relations
(Detectable by Balance Only)

Size Relations	$W_1 > W_2$	$W_1 < W_2$	$W_1 = W_2$
$S_1 > S_2$	(1)	(4)	(7)
$S_1 < S_2$	(2)	(5)	(8)
$S_1 = S_2$	(3)	(6)	(9)

* Originally developed by Willis Hively.

ITEM FORM SHEET

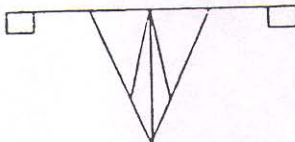
MATERIALS Beam Balance Objects 1 and 2 Item F.O. 16.14.0 Stimulus-response sheet (attached) Pencil	SCRIPT Here are two objects. They have symbols attached to them. Compare them by weight and write one of these three signs (point) in the blank (point) to form the comparison sentence. You may use this balance if you need it.
DIRECTIONS TO E Place materials in front of child (two areas of objects given above.)  Balance + = objects □ = S-R sheet Subject	SCRIPT Here are two objects. They have symbols attached to them. Compare them by weight and write one of these three signs (point) in the blank (point) to form the comparison sentence. You may use this balance if you need it.

RECORDING

Attach Stimulus-Response sheet to this page.

Describe what child did.

If balance was used, insert object symbols in schematic drawing of the balance given below, and mark the position of the plumb-line at the time of child's judgment.



DESCRIPTION OF MATERIALS

Pencil (F.O. 16.1.1)
 Beam Balance (F.O. 16.13.1) Equal-arm beam balance made from Timberly materials as described in MINIREAST Unit 16.
 Set of Weight Comparison Objects (F.O. 16.14.0). Set of opaque plastic cylindrical bottles with Army filling lids. Two sizes of bottles have been chosen. The small bottle has a length of 2" and a diameter of 1/2". The large bottle has a length of 2 1/2" and a diameter of 1 1/2". Two weight values have been chosen so that the objects cannot typically be distinguished by feeling but can be distinguished on the balance. Each object is designated by a randomly chosen, lower-case letter.

Size	Weight	
	23 gm	25 gm
small	a	m
large	b	n

Stimulus-Response sheet (attached to Item F.O. 16.14.1). A sheet of paper approximately 6" x 4" with the following display:

Write $>$, $<$, or $=$ in the blank

W _____ W.

where 1 and 2 are the appropriate subscripts from Replacement Schemes.

REPLACEMENT SCHEME

(F) Objects

Cell 1:	(a, b)	
Cell 2: <td>(m, n)</td> <td></td>	(m, n)	
Cell 3: <td>Choose</td> <td>from R.S. 16.13</td>	Choose	from R.S. 16.13
Cell 4: <td>(7, m)</td> <td></td>	(7, m)	
Cell 5: <td>(8, n)</td> <td></td>	(8, n)	
Cell 6: <td>Choose</td> <td>from R.S. 16.14</td>	Choose	from R.S. 16.14
Cell 7: <td>Choose</td> <td>from R.S. 16.15</td>	Choose	from R.S. 16.15
Cell 8: <td>Choose</td> <td>from R.S. 16.15</td>	Choose	from R.S. 16.15
Cell 9: <td>Choose</td> <td>from R.S. 16.17</td>	Choose	from R.S. 16.17

REPLACEMENT SETS

R.S. 16.13	Ordered pairs (m, n):	(8, 9)
R.S. 16.14	Ordered pairs (a, m):	(8, 9)
R.S. 16.15	Ordered pairs (b, n):	(8, 9)
R.S. 16.16	Ordered pairs (a, b):	(8, 9)
R.S. 16.17	Ordered pairs (m, n):	(8, 9)

SCORING SPECIFICATIONS

A correct response is made by writing the correct symbol ($>$, $<$, or $=$) in the blank space to complete the comparison sentence. This should be: in Cells 1, 2, and 3; $<$ in Cells 6, 7, and 8; $=$ in Cells 7, 8, and 9.

(Source: Hively et al, 1973)

As will be seen from the item form, any item form has the following characteristics (Osborn, 1968): 1) it generates items with a fixed syntactic structure, 2) it contains one or more variables (variable elements), 3) it defines a class of item sentences by specifying the replacement sets for the variables.

Such elaborate schemes as item forms guarantee that the domain is well defined and the population (universe) of items can be precisely described. It is, however, immediately obvious that to produce item forms must be very laborious and time consuming. It is also questionable whether similar levels of specificity can be reached in any other field than the formal languages of mathematics, logic and science.

Amplified Objectives

After finding out that item generation on the basis of traditional behavioral objectives was subject to too much interpretation and that using item forms was too demanding and led to "hyperspecificity", Popham (1980) worked with the so-called amplified objectives. As the name suggests, these are more detailed forms of behavioral objectives. They include 1) a brief statement of the objective, 2) a sample item, and 3) an amplified objective which specifies (a) the testing situation, (b) response alternative, and (c) criteria of correctness. The following example illustrates amplified objectives.

While amplified objectives clearly define the measured domain and specify item generation in greater detail than simple behavioral objectives, Popham (1980) observes that this attempt to "shoot for just the right balance between clarity and conciseness" failed. There was still too much room left for the personal interpretation of item writers.

Objective: Given a sentence with a noun or verb omitted, the student will select from two alternatives the word which most specifically or concretely completes the sentence.

Sample Item

Directions: Mark an "X" through one of the words in parentheses which makes the sentence describe a clearer picture.

Example: The racer (tumbled, went) down the hill.

Amplified Objective

Testing Situation

1. The student will be given simple sentences with the noun or verb omitted and will be asked to mark an "X" through the one word of a given pair of alternative words which more specifically or concretely completes the sentence.

2. Each test will omit nouns and verbs in approximately equal numbers.

3. Vocabulary will be familiar to a third- or fourth-grade pupil.

Response Alternatives

1. The student will be given pairs of nouns or pairs of verbs with distinctly varied degrees of descriptive power.

2. In pairs of verbs, one verb will either be a linking verb or an action verb descriptive of general action (e.g., is, goes), and one verb will be an action verb descriptive of the manner of movement involved (e.g., scrambled, skipped).

3. In pairs of nouns, one noun will be abstract or vague (e.g., man, thing), and one noun will be concrete or specific (e.g., carpenter, computer).

Criterion of Correctness

The correct answer will be an "X" marked through the more concrete, specific noun or through the more descriptive action verb in each given pair.

(Source: Millman, 1974)

Test Specifications

Experience with amplified objectives led Popham and his colleagues to believe that a so-called limited focus strategy was desirable. This means that the strategy is to focus measurement and to limit it to "a smaller number of assessed behaviors, but to conceptualize these behaviors so that they were large scale, important behaviors that subsumed lesser, en route behaviors" (Popham, 1980, p. 21).

The test specification consists of 1) a short general description, and 2) a sample item, which give the reader a general idea of what the test might contain. These are followed by 3) a detailed specification of the stimulus attributes and 4) response attributes including

specification of the correct answer and, in the case of multiple choice items, of the reasons for various distractors. The test specification is illustrated below.

An Illustrative Set of Criterion-Referenced Test Specifications
for a High School Minimum Competency Test in Reading

DETERMINING MAIN IDEAS

General Description

The student will be presented with a factual selection such as a newspaper or magazine article or a passage from a consumer guide or general-interest book. After reading that selection, the student will determine which one of four choices contains the best statement of the main idea of the selection. This statement will be entirely accurate as well as the most comprehensive of the choices given.

Sample Item

Directions. Read the selections in the boxes below. Answer the questions about their main ideas.

THE COLD FACTS

Had you lived in ancient Rome you might have relieved the symptoms of a common cold by sipping a broth made from soaking an onion in warm water. In Colonial America you might have relied on an herbal concoction made from sage, buckthorn, goldenseal, or bloodroot plants. In Grandma's time, lemon and honey was a favorite cold remedy, or in extreme cases, a hot toddy laced with rum. Today, if you don't have an old reliable remedy to fall back on, you might take one of thousands of drug preparations available without prescription. Some contain ingredients much like the folk medicines of the past; others are made with complex chemical creations. Old or new, simple or complex, many of these products will relieve some cold symptoms, such as a stopped-up nose or a hacking cough. But not a single one of them will prevent, cure, or even shorten the course of the common cold.

Reproduced with permission from *Test Specifications, IOX Basic Skill Tests: Secondary Level, Reading* (Los Angeles: The Instructional Objectives Exchange, 1978), pp. 21-24.

1. Which one of the following is the best statement of the main idea of the article you just read?
 - a. Old-fashioned herbal remedies are more effective than modern medicines.
 - b. There are many kinds of relief, but no real cures, for the common cold.
 - c. Some of today's cold preparations contain ingredients much like those found in folk remedies of the past.
 - d. Americans spend millions of dollars a year on cold remedies.

Stimulus Attributes

1. Each item will consist of a reading selection followed by the question "Which one of the following is the best statement of the main idea of the (article selection) you just read?" Eligible reading selections include adaptations of passages from factual texts such as general-interest books and consumer guides and pamphlets. Care should be taken to pick selections of particular interest to young adults and to avoid selections which may in the near future appear dated. Each reading selection will be titled, will be at least one paragraph long, and will contain from 125-250 words. Not more than 1,000 words of reading material can be tested in any set of five items. At least two of the five items in any set of five items must contain reading selections that are more than one paragraph long.

2. If necessary, the following modifications may be made to a selection used for testing:
 - a. A title may be added if the selection does not have one, or if the selection represents a section of a longer piece whose title would not be applicable to the excerpt. If a title is added, it should be composed of a brief, interest-getting and/or summarizing group of words.
 - b. A selection may be shortened, but only if the segment which is to be used for testing makes sense and stands as a complete unit of thought without the parts which have been omitted. If necessary, minor editing can be done to a reading selection which represents a shortening of a longer piece, but this editing should be for the purposes of clarity and continuity only, and not for the purposes of increasing or decreasing the difficulty level, or changing the content, of the text.
3. Reading selections used for testing should not exceed a 9th grade reading level, as judged by the Fry readability formula.

Response Attributes

1. A set of four single-sentence response alternatives will follow each reading selection and its accompanying question. All of these statements must plausibly relate to the content of the reading selection, either by reiterating or paraphrasing portions of that selection or by building upon a word or idea contained in the selection.
2. The three incorrect response alternatives will each be based upon a lack of one of the two characteristics needed by a correct main idea statement: *accuracy* and *appropriate scope*. A correct main idea statement must be accurate in that everything it states can be verified in the text it describes. It must have appropriate scope in that it encompasses all of the most important points discussed in the text that it describes.
3. A distractor exemplifies a *lack of accuracy* when it does any one or more of three things:
 - a. Makes a statement contradicted by information in the text.
 - b. Makes a statement unsupported by information in the text. (Such a statement would be capable of verification or contradiction if the appropriate information were available.)
 - c. Makes a statement incapable of verification or contradiction; that is, a statement of opinion. (Such statements include value judgments on the importance or worth of anything mentioned in the text.)
4. A distractor exemplifies a *lack of appropriate scope* when it does one of two things:
 - a. Makes a statement that is too narrow in its scope. That is, the statement does not account for all of the important details contained in the text.
 - b. Makes a statement that is too broad in its scope. That is, the statement is more general than it needs to be in order to account for all of the important details contained in the text.
5. The important points which must be included in a main idea statement are those details which are emphasized in the text by structural, semantical, and rhetorical means such as placement in a position of emphasis, repetition, synonymous rephrasing, and elaboration. Whether any given main idea statement contains all of the important points that it should is always debatable rather than indisputable. The nature of the question asked on this test, i.e., select the *best* main idea statement from among those given, attempts to account for this quality of relative rather than absolute correctness.
6. The distractors for any one item must include at least one statement that lacks accuracy and one statement that lacks appropriate scope. On a given test, between 10 and 20 percent of the distractors should be sentences taken directly from the text.

7. The correct answer for an item will be that statement which is both entirely accurate and of the most appropriate scope in relation to the other statements given. If a sentence in the text itself qualifies as the best main idea statement which can be formulated about the selection, that sentence may be reiterated as a response option. No more than 20 percent of the items on a given test may have as their correct answer a main idea statement which is a direct restatement of a sentence in the text.

(Source: Popham 1980)

Popham (1980, 1981) feels that test specifications like the one shown in the above constitute a reasonable balance between clarity and conciseness so that busy people like teachers might not be put off by extreme specificity. Test specifications can also contain a supplement, which can give additional guidance in how to select stimuli, how to phrase questions, and so on.

2.2. Construction and Selection of Items

In the construction of items certain general rules have been devised for producing traditional norm-referenced tests. Such advice is presented in a number of books which deal with testing and evaluation. Most of these rules are also applicable to criterion-referenced measurement. The only difference is that more stringent demands are set for the procedure in item generation. It is, for instance, very important to stick to the limits set for the stimulus and response characteristics. Convergent rather than divergent creativity is needed in item generation. Work carried out by Carroll (1968, 1976) is of interest in this respect even if it is not in the mainstream of criterion-referenced measurement. Roid and Haladyna (1980) also provide a useful review of recent advances in the item-writing technology, including computer-based methods (cf. also Millman 1980). They note that the major positive result of the increased attention to the process of item writing is the heightened concern for the logical congruence between instruction and testing.

Once the rules for domain definition and for item generation have been worked out, it is necessary to consider specific items. Unlike in norm-referenced testing, it is necessary in criterion-referenced testing to know what the universe of items is that represents the defined domain content.

This universe can be finite or infinite. As Millman (1973) points out, it is not necessary that the population of items actually exists. What is necessary, though, is that the domain is so well described that a high agreement can be reached about what items are and what are not members of the population.

Further, unlike in norm-referenced and mastery tests, it is necessary to draw a random sample from the universe of all possible items because only this procedure makes it possible to produce an estimate of the examinees' total domain scores. Random sampling of items is needed in order to make it possible to generalize into the whole domain tested. It is generally assumed that 10-20 items are needed to measure a given content domain.

3. Standard-setting as an Issue in Criterion-Referenced Measurement

Standard-setting has been a topic of great controversy within the criterion-referenced movement. The need to set standards for acceptable performance has been especially great in mastery-type instructional programs, in which it is assumed that a certain level of mastery is optimal for both cognitive and affective outcomes (e.g., Block, 1972). Therefore, it would be important to identify masters and non-masters without making too many wrong classifications. In competency-based promotion systems it is also equally important to avoid too many cases of wrong decisions ("false positives", i.e., pseudo-masters and "false negatives", i.e., pseudo-non-masters). The decision-maker has to specify a loss function, in other words, state the relative seriousness of either passing students who lack requisite knowledge and skills or holding back students who in fact should be passed.

Methods available for setting standards have been discussed in several articles (Hambleton, 1980; Hambleton and Eignor, 1978; Hambleton et al, 1978; Jaeger, 1976; Meskauskas, 1976; Millman, 1973) and critically reviewed by Glass (1978). A whole issue of the Journal of Educational Measurement (Vol. 15, No. 4, 1978) was devoted to this problem. Glass provided a critical overview and Scriven, Block, Popham and Hambleton tried to rebut his main thesis that standard setting methods, in spite of their seemingly objective procedures, are basically arbitrary.

It is, in fact, now generally accepted that setting passing scores is arbitrary in the sense that it is based on judgment, but the advocates of standard setting maintain that it is not arbitrary in the sense of "capricious"

of "unjustifiable". They point out that human life is full of situations where informed judgment must be exercised and measurement should not be faulted too much if some of its procedures also must resort to this method. Thus, they claim, what is needed is not the abolishment of standard setting but the improvement of its procedures. Hambleton (1980) classifies them into three groups: judgmental, empirical and combination.

All judgmental methods require that data are collected from subject-matter experts and other qualified judges for setting standards. Individual items are carefully inspected to judge how a minimally competent person would perform on them. Methods proposed by Nedelsky (1954), Angoff (1971), Ebel (1972) and Jaeger (1978) differ on some points, and what is more disturbing, they can lead to quite different passing scores and rates. It has been shown, for instance, by Andrew and Hecht (1976) that when the same judges used both the Ebel and Nedelsky methods, the passing scores varied from 49% of all items to 68% and the passing rates varied from 50% of all examinees to 95%. Such variability is clearly too wide and indicates the need for further work on this problem.

Since empirical methods are seldom used alone, they will not be discussed in this paper. The combination method uses both judgmental and empirical data. In the "Borderline Group Method" the judges are first asked to think of a minimally acceptable performance on the measured content area. They are then asked to give a list of those students whose performance is so close to the borderline that it is difficult to classify them with confidence. The test is then administered and the median score for the borderline group is taken as the passing score.

In the "Contrasting Groups Method" the judges are first asked to determine in their minds the minimally acceptable performance level and then identify those students who can be classified clearly either as masters or non-masters. Empirical test data are then obtained for both groups and the point of intersection of the two score distributions is taken as the passing standard. The present author used this method in 1979 in the first national assessment of English as a foreign language in Finland in an attempt to study how teachers' judgments could be used in establishing a common core syllabus for English. The results have not yet been analysed.

4. Validity as an Issue in Criterion-Referenced Measurement

Criterion-referenced tests are more and more often used in monitoring individual progress through objectives-based instructional programs (formative testing), to diagnose learning problems (diagnostic testing), to evaluate educational and social programs (program evaluation), and to assess level of performance on certification and licensing examinations. The usefulness of such applications depends heavily on the validity of the procedures undertaken in such testing.

According to Hambleton (1980) validity considerations in criterion-referenced testing arise at three steps: 1) the selection of objectives (content domain), 2) the measurement of objectives (content domains) included in the criterion-referenced test, and 3) the uses of test scores.

Validity is a difficult topic in all measurement and criterion-referenced measurement is no exception. Terminology varies quite a lot so that different terms are used to designate the same characteristic and the same term is used to designate somewhat different things. There are also some fundamental confusions that have persisted for a long time.

As Cronbach (1971), Messick (1975) and Linn (1979) have pointed out, a major conceptual confusion arises from the fact that content validity is focused on test forms rather than test scores, on instruments rather than measurements. In Linn's words "questions of validity are questions for the soundness of the interpretations of a measure... Thus, it is the interpretation rather than the measure that is validated. Measurement results may have many interpretations which differ in their degree of validity and in the type of evidence required for the validation process" (Linn, 1979, p. 109). For this reason, Messick states that content coverage is an important consideration in test construction and interpretation but it does not itself provide validity. He would prefer the term "content relevance" or "content representativeness", since they do not really provide evidence for the validity of the interpretation of scores.

Popham (1978) uses the term "domain-selection validity" to refer to the question of how well the results obtained can be generalized to as many other domains as possible. It thus resembles "construct validity" to some extent, although the latter is a more theoretical concept. Since testing for many reasons ought to be limited to a minimum, it is important to measure such domains and use such techniques which permit maximum generalization across

domains of content. Domain-selection validity can be assessed by asking experts to give judgements on the relevance of selected domains.

Popham (1978) proposes the term "descriptive validity" to indicate the representativeness of measured content. In traditional norm-referenced testing no quantitative indices are usually given to describe content representativeness (cf. Table 1). In criterion-referenced testing, judges can be used to assess to what extent items are congruent with the test specification. Hambleton (1980) provides some useful methods for doing this. In some areas, where it is possible to specify completely a pool of valid test items, the representativeness of items can be ensured by drawing a random sample from the item pool. This was the procedure adopted when the present author studied students' active and passive vocabulary of English in the Finnish comprehensive school in 1979.

Hambleton (1980) uses the term "decision validity" to refer to the decisions made on the basis of scores. Popham (1978) uses the term "functional validity" in much the same sense. Decision validity in criterion-referenced testing is often related to standard setting (minimum passing scores). Since that question is discussed elsewhere in this paper (section 3, p. 3), it will not be dealt with further in this context. A good review of decision-consistency is in Subkoviak (1980). Hambleton and Eignor (1978) and Walker (1978) review and assess standards and guidelines for evaluating criterion-referenced tests and test manuals.

5. Reliability as an Issue in Criterion-Referenced Measurement

Traditional methods of estimating reliability in norm-referenced measurement are usually based on correlational analyses where variance is a key concept. Since there may be relatively little variation in the scores of criterion-referenced tests, correlation-based estimates may not be ideally suitable for the estimation of reliability.

As Berk (1980) has noted there are at least three major conceptualizations of criterion-referenced test reliability: 1) consistency of mastery-non-mastery decisions across repeated measures with one test form or parallel test forms, 2) consistency of squared deviations of individual scores from the cut-off scores across parallel or randomly parallel test forms, 3) consistency of individual scores across parallel or randomly parallel test forms.

Subkoviak (1980) gives a good survey of five methods of determining decision-consistency reliability. Usually only two statistics are used in this context: P_0 , which indicates the proportion of individuals consistently classified as masters and non-masters across parallel test forms, and κ , which estimates the proportion of individuals consistently classified beyond that expected by chance. Thus, P_0 estimates the overall consistency whereas κ estimates consistency due to testing alone. The choice of the index has to be based on whether one wants an estimate of overall consistency of decisions for whatever reason or of the contribution of the test alone. In most cases, it is probably advisable to report both estimates.

Brennan (1980) reviews the generalizability theory approach to reliability, which builds on the work by Cronbach and his associates (1972). Generalizability theory is based on the analysis of variance model and focuses on the estimation of various variance components in different types of test x items designs. Generalizability theory allows for the existence of many types and sources of error and it does not require strictly parallel tests for reliability estimation. Only randomly parallel tests are required.

As in the case of the decision-consistency approach, there are two indices of reliability (or dependability): $\phi(\lambda)$ provides an estimate of the dependability of mastery-non-mastery decisions based on the testing procedure (λ represents the cut-off score), and ϕ the "general purpose" index that is independent of the cut-off score and which can be used to estimate individual domain scores (a major interest in the present writer's study of the size of students' active and passive vocabulary). $\phi(\lambda)$ is related to the reliability of criterion-referenced test scores and ϕ is associated with the reliability of domain score estimates. The former indicates how closely the scores for any examinee can be expected to agree, the latter the degree of agreement with chance agreement removed. Thus $\phi(\lambda)$ characterizes the dependability of decisions, or estimates, based on the testing procedure. Its magnitude depends, in part, on chance agreement. The index ϕ characterizes the contribution of the testing procedure to the dependability of decisions, over and above what can be expected on the basis of chance agreement (Brennan, 1980).

As in the case of the decision-consistency approach, it might be useful to give both estimates. Brennan (1980) also strongly recommends that variance components too should always be reported.

6. Uses of Criterion-Referenced Measurement

The several possible applications of criterion-referenced measurement are mainly due to the increased rigor and precision in the description of important subject-matter domains and of behavior related to them. Some of the most common uses of CRM are described below drawing mainly on Millman (1974) and Popham (1978).

CRM can be used in needs assessments, which help in setting educational priorities. Need can be defined as the difference between an expected and the present observed situations. The latter can best be ascertained by means of CRTs, which possess a high degree of content representativeness. It also follows that CRM can be used in individualized teaching programs to assess students' current status with respect to objectives.

One of the most promising uses of CRM is in the area of large-scale program evaluation. Since CRM puts such rigorous demands on the item-program congruence, it is ideally suited to reveal the effectiveness of instruction or the lack of it. CRM with random samples of items from well-defined content domains provides reliable estimates of students' domain scores and makes reliable and valid generalizations to whole teaching programs possible. It furnishes reliable qualitative and quantitative information on learning and thus CRM will be of great help in efforts to develop education.

Methods developed within the so-called modern test theory movement, which has worked out new methods for avoiding some of the problems and limitations of classical test theory, (for instance, latent trait theory, generalizability theory, Bayesian methods), make it possible to shorten testing time. This is possible by either using the instructors' earlier knowledge of students in the estimation of their level of functioning (Bayesian methods) or by applying multilevel testing procedures or both combined. In multilevel testing items of varying difficulty, carefully prepared from some well-defined domain of content, are divided into a few groups. Each student takes the group of intermediate difficulty and then goes to either an easier or more difficult set depending on how difficult the intermediate set was for him or her. Lord (1976) notes that testing time can be reduced to one half and the number of items needed can be dropped from 100 to 20. Multilevel testing also has the positive affective effect of not shocking students with too difficult items or boring them with too easy ones.

7. Current Problems and Issues in Criterion-Referenced Measurement

As Popham (1978) points out one of the greatest problems in the development of criterion-referenced measurement is its difficulty and laboriousness. Where to get the resources for activity which presupposes highly trained full-time measurement experts and takes a lot of time? Popham (1980) says that he is distressed that he is unable to teach people how to go about the conceptualization of tested domains. In his own words, "at no point in the test development process for criterion-referenced measures is it more apparent that we are employing art, rather than science, than when the general nature of the behavioral domain to be tested is initially conceptualized" (p. 26).

It seems to the present writer that Popham is overemphasizing the "artistic" aspect of domain conceptualization. It seems likely that the reason for the felt difficulty is mainly due to the lack of a theoretical grasp of the structure and nature of the tested subject matter. If there were a better theoretical conception of the content structure and of the cognitive structure of some school subject, surely domain specifications would not need to be so much "artistic endeavors of no small shakes" (Popham, 1980, p. 27). It is, however, hard to find persons who master both the theoretical structure of subject-matter and the structure of the cognitive processes involved in its learning and use. Usually one is an expert in only one of these two aspects. After several years of work in curriculum construction, curriculum evaluation and textbook writing the present writer is convinced that the state of art in subject-specific domain specification in several school subjects is very low and serious work in this area has hardly been started. There is an urgent need for developing the "psychologies" of specific school subjects if there is to be any real progress in curriculum construction, teaching and evaluation.

Another related problem is the codification of guidelines for the construction of criterion-referenced tests and for their use. This would also be of great help in the training of test constructors and test users.

In addition to such content-specific problems, there are a number of technical problems that need to be studied. These include methods of estimating the validity and reliability of different uses of criterion-referenced tests; the use of computers in generating test items; and the employment of new ideas of modern test theory in criterion-referenced measurement.

Criterion-referenced measurement is so new if it is compared with norm-referenced measurement, and similarly modern test theory is new in relation to classical test theory, that both have a number of "unsolved problems and

"problematic solutions", as Popham so aptly puts it. There is intensive work being done all over the world to produce less problematic solutions to such problems and there is no need to doubt that such solutions will be forthcoming.

8. Discussion

Criterion-referenced measurement and norm-referenced measurement share a number of features. As in several other fields, for instance, in curriculum construction, new approaches usually mean only new emphases. At first there is a tendency to exaggerate differences. It is possible that this is inevitable when a new idea is introduced. Karl Popper has suggested that certain dogmatism may have an important part to play in the development of science, because giving up an idea too soon may mean that its merits and weaknesses are not given a sufficient chance of showing themselves. A scientist should not be too ready to adopt a new idea or to abandon an old one without persisting in some seemingly dogmatic stance for some time for the sake of argument. We should know how to play the believing and doubting games in a balanced way.

Criterion-referenced measurement shows some characteristics of this initial dogmatism. At first it was categorically stated that CRM does not need such concepts as item and score variance; that empirical item analyses are not needed; that norm data should not be gathered; and that content validity is the most important aspect of CRM. It was soon admitted, however, that these claims were overstated. Item variance usually occurs and serves a useful purpose in CRM testing as well as in norm-referenced testing. Similarly, it was conceded that norm data are not embarrassing for CRM. On the contrary, they add useful information and can help to interpret how "good" is "good enough". A posteriori empirical item analyses complement a priori judgemental (rational-logical) item analysis and help to detect flawed items. And, finally, content validity is not the all-important consideration in CRM. While content representativeness is a necessary characteristic of CRM it does not guarantee the validity of interpretations based on CRT scores.

Criterion-referenced measurement has the special advantage that it provides an exact description of a person's performance level in an entire domain and not only in the presented items. Several requirements must be

75

fulfilled before such an interpretation is possible. First, there has to be a detailed description of the measured domain. Second, there must be a detailed description of the instrument, which includes the specification of the stimulus and response parts and of the scoring system. Third, items must be generated that have a high item-objective congruence and which are also a representative random or stratified random sample from the item pool. If CRM is used for program evaluation there must also be a representative sample of students from the entire population. In the latter case it is advisable to use matrix sampling with several parallel test versions rotated in the class.

One of the greatest attractions of CRM for the present writer is its emphasis on the conceptualization of measured domains. This lends support to his personal claim, which goes back several years, that one of the greatest obstacles for the development of teaching is the lack of theoretically sound conceptualizations of the units and processes in learning a particular subject matter. He would, therefore, fully agree with the view recently put forward by Popham:

When created by instructionally astute developers, a criterion-referenced test can lay out so lucidly a set of teachable skills that the test itself becomes a potent force for instructional improvement. Instead of being an afterthought for use at the close of instruction, a properly conceptualized criterion-referenced test can stimulate measurement-driven instructional enhancement. Test developers can literally create test items so that they agree with one or more instructionally powerful explanatory constructs which teachers can then employ during their lessons... . This sort of focused instructional enterprise is not teaching-to-the-test in the negative sense that one teaches toward a particular set of test items. Rather, this approach constitutes teaching-to-the-skill, a highly effective and thoroughly defensible instructional strategy" (Popham, 1981, pp. 106-107).

Thus it might be that "the testing tail wagging the teaching dog" may not be such a problem or the embarrassment it is often taken to be if the tail is fully compatible with the dog. The present writer's personal experience with curriculum construction and evaluation, and with the in-service education of teachers in Finland suggests that the most effective and fastest way to promote desirable changes in teaching is to make sure that testing and tests display the characteristics of desirable student performance. Tests are the most concrete ways of signaling to teachers and students what the desirable content and forms of learning are.

Focusing on testing may be more effective than focusing on curricula and teaching materials since testing has a more limited scope and it is,

therefore, possible to produce very carefully constructed tests that are, in a sense, modules of teaching. Such tests can serve as examples for preparing units of teaching and for individual lessons. By concentrating on important aspects of the subject matter it is possible to produce such modules which can also serve as a stimulus for textbook writers. While individual units and modules do not constitute an entire syllabus, they are useful wholes as such and can serve as useful models. Practical experience shows that it is much more difficult to seek to conceptualize an entire curriculum with similar rigor and it is also a huge task to produce a textbook package with a similarly consistent approach. Thus testing may, indeed, be a sensible starting point and lead to improved curricula and textbooks. At the very least, the potential contribution of work done within testing and measurement to curriculum design and instruction should not be ignored.

REFERENCES

- Anderson, C. 1972. How to Construct Achievement Tests to Assess Comprehension. Review of Educational Research, 42, 145-170.
- Andrew, B.J., and J.T.A. Hecht. 1976. A Preliminary Investigation of Two Procedures for Setting Examination Standards. Educational and Psychological Measurement, 36, 45-50.
- Angoff, W.H. 1971. Scales, Norms and Equivalent Scores. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council of Education.
- Berk, R.A. 1980a. Introduction. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 3-9.
- . 1980b. Item Analysis. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 49-79.
- Block, J.H. 1972. Student Evaluation: Toward the Setting of Mastery Performance Standards. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Bloom, B.S. 1968. Learning for Mastery. Evaluation Comment, Vol. 1, No. 1.
- Bloom, B.S., J.T. Hastings, G.F. Madaus (Eds.) 1971. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill.
- Bormuth, R. 1970. On the Theory of Achievement Test Items. Chicago: University of Chicago Press.
- Brennan, R.L. 1980. Applications of Generalizability Theory. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 186-232.
- Carroll, J.B. 1963. A Model of School Learning. Teachers College Record, 64, 723-733.
- . 1968. The Psychology of Second Language Testing. In A. Davies (Ed.) Language Testing Symposium: A Psycholinguistic Approach. London: Oxford University Press, 46-68.
- . 1976. Psychometric Tests as Cognitive Tasks: A New "Structure of Intellect". In L.B. Resnick (Ed.) The Nature of Intelligence. New York: Lawrence Erlbaum, 27-56.
- Carver, R.P. 1974. Two Dimensions of Tests: Psychometric and Edumetric. American Psychologist, 29, 512-518.
- Cronbach, L.J. 1957. The Two Disciplines of Scientific Psychology. American Psychologist, 12, 671-684.

- , 1971. Test Validation. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council of Education.
- Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam. 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.
- Ebel, R.L. 1971. Content Standard Test Scores. Educational and Psychological Measurement, 22, 15-25.
- , 1972. Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall.
- Gagne, R.M., J.R. Mayor, H.L. Garstens, and N.E. Paradise. 1962. Factors in Acquiring Knowledge of a Mathematical Task. Psychological Monographs, Vol. 76, No. 526.
- Glaser, R. 1963. Instructional Technology and the Measurement of Learning Outcomes: Some Questions. American Psychologist, 18, 519-521.
- Glaser, R., and A. Nitko. 1971. Measurement in Learning and Instruction. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council of Education, 652-670.
- Glass, G.V. 1978. Standards and Criteria. Journal of Educational Measurement, 15, 4, 237-261.
- Guttman, L. 1969. Intergration of Test Design and Analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service.
- Hambleton, R.K. 1980. Test Score Validity and Standard-Setting Methods. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 80-123.
- Hambleton, R.K., and D.R. Eignor. 1978. Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals. Journal of Educational Measurement, 15, 4, 321-327.
- Hambleton, R.K., H. Swaminathan, J. Algina, and D.B. Coulson. 1978. Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments. Review of Educational Research, 48, 1, 1-47.
- Hively, E., G. Maxwell, G. Rabehl, D. Sension, and S. Lundin. 1973. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnemast Project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California.
- Jaeger, R.M. 1976. Measurement Consequences of Selected Standard-Setting Models. Florida Journal of Educational Research, 18, 22-27.
- , 1978. A Proposal for Setting a Standard on the North Carolina High School Competency Test. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill.

- Linn, R.L. 1979. Issues of Validity in Measurement for Competency-Based Programs. In M.A. Bunda and J.R. Sanders (Eds.) Practices and Problems in Competency-Based Measurement. Washington, D.C.: National Council on Measurement in Education, 108-123.
- Lord, F.M. 1976. Test Theory and the Public Interest. Proceedings of the 1976 ETS Invitational Conference. Princeton, N.J.: Educational Testing Service, 17-30.
- Mager, R.F. 1962. Preparing Instructional Objectives. Palo Alto: Fearon Publishers.
- Meskauskas, J.A. 1976. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard-Setting. Review of Educational Research, 46, 133-158.
- Messick, S.A. 1975. The Standard Problem: Meaning and Values in Measurement and Evaluation. American Psychologist, 30, 955-966.
- Millman, J. 1973. Passing Scores and Test Lengths for Domain-Referenced Tests. Review of Educational Research, 43, 205-216.
- , 1974. Criterion-Referenced Measurement. In W.J. Popham (Ed.) Evaluation in Education: Current Applications. Berkeley: McCutchan.
- , 1978. Determinants of Item Difficulty: A Preliminary Investigation. Center for the Study of Evaluation, CSE Report No. 114.
- , 1980. Computer-Based Item Generation. In R.A. Berk (ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 32-43.
- Nedelsky, L. 1954. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 14, 3-19.
- Osborn, H.G. 1968. Item Sampling for Achievement Testing. Educational and Psychological Measurement, 28, 95-104.
- Popham, W.J. 1978. Criterion-Referenced Measurement. Englewood Cliffs, N.J.: Prentice Hall.
- , 1980. Domain Specification Strategies. In R.A. Berk (ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 15-31.
- , 1981. Measurement Essentials for the Essentials of Education. In L.Y. Mercier (Ed.) The Essentials Approach: Rethinking the Curriculum for the 80's. U.S. Department of Education, 97-115.
- Popham, W.J., and T.R. Husek. 1969. Implications of Criterion-Referenced Measurement. Journal of Educational Measurement, 6, 1-9.
- Resnick, L.B. 1967. Design of an Early Learning Curriculum. Working Paper 16. University of Pittsburgh: Learning Research and Development Center.

- Roid, G., and T. Haladyna. 1980. The Emergence of an Item-Writing Technology. Review of Educational Research, 50, 2, 293-314.
- Subkoviak, M.J. 1980. Decision-Consistency Approaches. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 129-185.
- Takala, S. 1982. On the Origins, Communicative Parameters and Processes of Writing. Evaluation in Education, 5, 3, 209-230.
- Walker, C.B. 1978. Standards for Evaluating Criterion-Referenced Tests. Center for the Study of Evaluation, CSE Report No. 103.