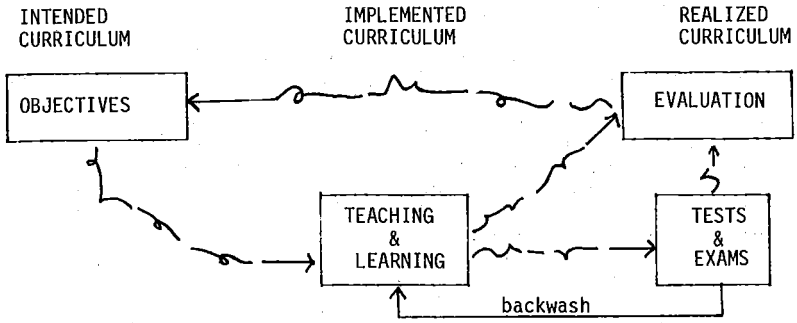
CRITERION-REFERENCED TESTING IN FOREIGN LANGUAGE TEACHING

Sauli Takala

1 Objectives - teaching - evaluation: the problems of match

One of the biggest problems in all educational activity is how to match objectives, teaching and evaluation. In other words, how the "intended curriculum" is transformed into the "implemented curriculum" and finally into the "realized curriculum" is no simple matter. Textbook writers and teachers interpret the intended curriculum and "implement" it to a certain extent: they cover some points thoroughly, some in a more cursory way, do not teach some points at all, and by contrast, teach some contents not mentioned in the intended curriculum. Students, in turn, learn ("realize") the implemented curriculum to a certain extent. The links between these three concepts are illustrated in the following figure.



The broken lines are intended to illustrate that there are no simple ways of getting from one stage in the process to another. There is a lot of interpretation and this means that there is a need to develop explicit and rigorous procedural rules to make such interpretation less ad hoc.

Tests and examinations are distinguished from evaluation. This is meant to convey the idea that evaluation is a broader (superordinate) term than testing. It is also meant to suggest that tests and examinations are not necessary for making evaluations. Evaluation means making judgements about the worth (merit, value) of something. In all walks of life we are constantly making such judgements. In education, evaluation usually means comparing what has been learned to what was intended to be taught.

Such judgements may be based on observation alone, but when properly made and used, tests can naturally give a valuable contribution to making correct evaluations. Tests are only one of the ways of getting data (information) for making evaluations.

2 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 (1980a), 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. Osborn (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 (1973) 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.

3 Stages in test Construction

3.1 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.

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?

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.

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.

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.

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 Methods: 1. Item transformations 2. Mapping sentences 3. Algorithms 4. Item forms 5. Amplified objectives 6. Test specifications	Content limits only partially specified Methods: Instructional and behavioral objectives	Content limits only partically specified Methods: 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 Methods: 1. A priori judgement of item-objective congruence by subject matter experts 2. A posteoriori computation of item statistics	Purpose to detect flawed items Methods: ?	Purpose to select items Methods: A posteoriori 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 Construct Decision	Content Criterion-related Construct Decision	Criterion-related
8. Reliability	1) Consistency of decisions ($\hat{P}_o$, $\hat{k}$) 2) Dependability ($\Phi(\lambda)$) 3) Error of measurement or estimate around domain score using Φ or other indices	Consistency of decisions ($\hat{P}_o$, $\hat{k}$)	Traditional procedures (based on correlation)
9. Score Interpretation	Performance in relation to domain (level of functioning) Performance in relation to required level of mastery	Performance in relation to required level of mastery	Performance in relation to other examinees
10. Item and Test Variance	Not required	Not required	Required

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

A. Activity

1. Receive
2. Send

a/an

B. Channel

1. auditive
2. visual

message which deals with

C. Content/topic

1. self
2. school
3. home town
4. hobbies
5.
6.

and whose

D. Communication Partner

1. addressor
2. addressee

has/is

E. Role relationship between addressor and addressee

1. a higher social status
2. an equal social status
3. a lower social status
4. identical with addressor

and which is

F. Degree of publicity/formality

1. private
2. semi-public
3. public

consisting of

G. Input-output relationship (stimulus-response)

1. repetition of input
2. modification of input
3. internal input

and whose purpose is

H. Function

1. to preserve the message (documentative)
2. to inform (referential)
3. to persuade (emotive)
4. to describe (descriptive)
5.
6.

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 = letter of application

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)

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.

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 (Takala, 1984).

Domain Specification and Item Generation Rules for Vocabulary Size Assessment

Behavior

(1) When given a Finnish word in writing, the student can produce an acceptable English equivalent in writing (recall or active vocabulary). (2) When given an English word in writing, the student can produce an acceptable Finnish equivalent in writing (recognition or passive vocabulary).

Stimulus specification

The vocabulary presented in the core texts and extra (optional) texts in widely used English textbooks is listed. A stratified random sample is selected from the universe of such word lists. The words are presented without providing any context. Some of the words are used to measure both the passive and active knowledge of word meanings.

Response specification

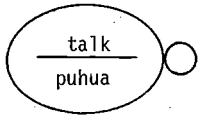
The student has to write the response in the space provided for that purpose.

Scoring

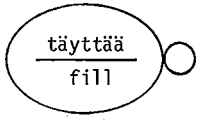
The responses are scored 0 - 1. A semantically acceptable and understandable response, which may contain spelling errors, is scored 1. In scoring active vocabulary, the decision is based on how the written English word would sound if read aloud. Thus, the student will get full marks if he/she has given the English equivalent of the Finnish word "talo" as "haus" instead of "house", since "haus" in Finnish orthography corresponds to the way "house" is pronounced in English.

Sample items

Instructions: "In this test you can show how well you know the English vocabulary included in your course work. Below are presented a number of Finnish words. Your task is to write the English equivalent on the line above the Finnish word. Write the word even if you may not be quite sure about the correct spelling, since spelling mistakes are a minor consideration in scoring."



"Write the Finnish equivalents of the following English words."



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.

3.2 Size of Domain

The proper size of the domain is, as so many issues in testing, ultimately dependent on the purpose of the test (measurement). A fairly large domain is appropriate if we are interested in more general forms of "terminal

behaviors" (i.e., we are doing "summative" evaluation and giving grades). A more limited domain definition is recommended when we are more interested in "en-route" behaviors and need information for deciding whether we need to review some matters with all students or give remedial help to some students ("formative" and "diagnostic" evaluation). Let us illustrate the issue of domain size with some concrete examples. Compare the following domain definitions.

1. Student can speak English
2. Student can ask for information in English
3. Student can ask about (a) time in English
   (b) place
   (c) cost
   (d) another person's feelings
   (e) another person's preferences
   (f) another person's opinions
   (g) another person's advice
   etc., etc.
4. Student can ask for
   (a) conformation of information (i.e., make yes/no questions)
   (b) lacking information (make questions with HOW, WHAT, WHEN, etc.)
5. Student can make questions with WHAT
6. Student knows what WHAT means
7. Student knows how to spell/pronounce WHAT

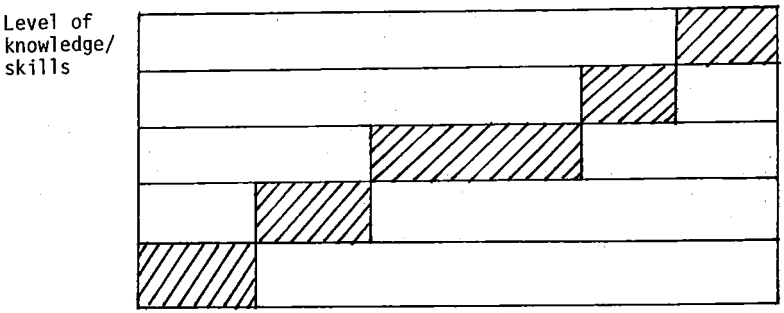
Points 1 and 2 are probably too broad domain definitions to be of much use to language teachers. Levels 3 and 4 are probably of appropriate size for summative evaluation. Levels 5 through 7 might be considered for diagnostic purposes.

Following Baker (1974) and Popham (1975), it should be emphasized that we should not test trivial matters because that may lead to excessive testing. We should only test important matters and reserve as much time as possible for teaching and learning. Our tests should include only such items that contain those features and elements whose variation makes a difference in student response (Millman 1974). To take a simple example, it probably makes no difference whether we use "he" or "she" to test whether students can use the s-form in the simple present tense. If this is so, it is superfluous to have two items, one with "he" and one with "she".

Trying to concretize further the issue of domain size, it is probably not useful to treat as a domain something that can be taught and learned in one lesson. On the other hand, if something takes a whole term to teach, that probably is best divided into more than one domain.

3.3 Levels of Measurement

Knowledge and skill are not dichotomous phenomena, i.e., it is seldom the case that we either know and can do something or do not know or cannot do it. There are various degrees of knowledge and various levels of skill. Knowledge can be partial or complete, and skill can range from that of a novice to that of master (expert). One of the most important points to keep in mind in all measurement and testing is that it should sensitively portray such a range of knowledge and skills. We do a disservice to teachers and students, and undermine the role of the school, if we measure only at higher levels and thus underestimate the effectiveness of teaching and learning. This principle can be illustrated by the following figure.



The author has discussed this question in greater detail in his dissertation which dealt with vocabulary learning. Due to space limitations, an interested reader is referred to that publication (Takala, 1984, 55-57; 65-67; 84-85). See also Appendix 1.

4 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 Hala-

dyna (1980; 1982) 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.

5 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 domain) 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 thing. 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 interpretation 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 somewhat beyond the scope of this paper 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.

6 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). $\hat{\Phi}(\lambda)$ is related to the reliability of criterion-referenced test scores and $\hat{\Phi}$ 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 $\hat{\Phi}(\lambda)$ characterizes the dependability of decisions, or estimates, based on the testing procedure. Its magnitude depends, in part, on chance agreement. The index $\hat{\Phi}$ 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.

7 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 on the presented items. Several requirements must be 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.

If we continue to do serious work on testing, we can move from what some might describe as the modern "test cult" more and more towards "test culture". Test culture is characterized by several desirable features. First, there is an awareness of the importance of knowing why one is testing in the first place. Second, there is an awareness of the problems of how valid interpretations (conclusions) can be made on the basis of obtained, more or less reliable, scores. Third, there is an awareness of problems of the generalizability of the results to the whole content universe and to the whole student population. Fourth, and finally, there is an absence of dogmatism and taboos concerning test types. Test culture is mature when we are aware of every aspect of testing and evaluation being riddled with problems but we are only pleased to have been able to reach such a level of awareness.

References

Anderson, C. 1972. "How to construct achievement tests to assess comprehension". Review of Educational Research 42: 145-170.

Berk, R. A. 1980a. Introduction. In R. A. Berk (ed.) 1980. 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.) 1980. Criterion-referenced measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press 49-79.

Bloom, B. S. 1968. Learning for mastery. Evaluation Comment, 1 (1).

Bloom, B. S., J. T. Hastings and G. F. Madaus (ed.). 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.). 1980. Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press 186-202.

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.) 1968. Language testing symposium: A psycho-linguistic approach. London: Oxford University Press 46-68.

- 1976. Psychometric tests as cognitive tasks: A new "structure of intellect". In L. B. Resnick (ed.). 1976. 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.). 1971. Educational Measurement. Washington, D.C.: American Council of Education 443-507.

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. 1962. "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.). 1971. Educational measurement. Washington, D. C.: American Council on Education. 652-670.

Guttman, L. 1969. Integration of test design and analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Servide.

Hambleton, R. K. 1980. Test score validity and standard-setting methods. In R. A. Berk (ed.). 1980. 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: 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.

Linn, R. L. 1979. Issues of validity in measurement for competency-based programs. In M. A. Bunda and J. R. Sanders (eds.). 1979. 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.

Mager, R. F. 1962. Preparing instructional objectives. Palo Alto: Fearon Publishers.

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.) 1974. Evaluation in education: Current applications. Berlekey: McCutchan. 310-397.

- 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.). 1980. Criterion-referenced measurement: The state of the art. Baltimore and London: The Johns Hopkins University Press. 32-43.

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.). 1980. 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.). 1981. 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.

Roid, G. H., and T. M. Haladyna. 1980. "The emergence of an item-writing technology". Review of Educational Research, 50: 293-314.

Roid, G. H. & Haladyna, T. M. 1982. A technology for test-item writing. New York: Academic Press.

Subkoviak, M. J. 1980. Decision-consistency approaches. In R. A. Berk (ed.). 1980. 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: 209-230.

Takala, S. 1984. Evaluation of students' knowledge of English vocabulary in the Finnish comprehensive school (Tech. Rep. No. 350). University of Jyväskylä. Institute for Educational Research.

Walker, C. B. 1978. Standards for Evaluating Criterion-Referenced Tests. Center for the Study of Evaluation, CSE Report No. 103.

CONTENTS	LANGUAGE KNOWLEDGE			LANGUAGE SKILLS				COMMUNICATION SKILLS (LANGUAGE USE SKILLS)	
	Levels of language/ discrete tests			Sub-skills/ integrative tests				Pragmatic tests	
LEVEL OF ACTIVITY	PHONOLOGY	MORPHOLOGY & SYNTAX	SEMANTICS	Listening comprehension	Speaking	Reading comprehension	Writing	Non-verbal	Verbal
MECHANICAL SKILLS (discriminate, repeat, recognize, recall)	Which is different? Which are identical? Imitate! Repeat!	Recall conjugation, paradigms Do in accordance with the model! Recite a rule	Choose/ give -L1 equivalent -L2 equivalent	Write down phone numbers Spell names	Imitate! Repeat! Respond to...	Which is different? Which are the same?	Copy!	Conventional gestures and facial expressions in conversational interactivity Interpret! Use gestures ..!	Conventional verbal expressions customary in verbal interaction Interpret ! Act (speak/write) in a manner expected in the context!
	Minimal pairs	Multiple choice Matching Reorganization	Multiple choice Matching					Eg. smiling (What were going to say?) or a gesture with the thumb (In that direction)	Dictation Proof-reading Linguistic rituals (fillers etc)
KNOWLEDGE UNDERSTANDING (recognize, recall, classify, state..)	Is x voiced/ voiceless? Is x short or long? Where is main stress? What intonation?	What tense? Sing/plural? What case? What does x refer to? Which structure fits the context?	What word class? What is the meaning of x(using word formation)?	What was said? What facts were stated? What opinions were expressed?	State/present/ inform/express..	What was written? What facts/ opinions were expressed?	Write, state, present, express	Respond/act in a way customay in context x	Respond and act(speak/ write) in a way customary in the context
	Multiple choice Matching			True-false Multiple ch. Matching Classification Question-answer Error spotting	Reading aloud Guided speaking	True-false Multiple choice Matching Classification Quest-answer Error spotting	Fill-in Transform. Guided writing	Eg. movement of hand with the palm upwards (Ladies first! Would you like to say something as this point?)	Taking notes Interpreting Chatting Reporting Cloze
APPLICATION (perform, execute, demonstrate, produce, invent, create..)	Read aloud	What structure would you use in this context?	What is the corresponding adj., adverb, verb etc of this word?	What does the speaker really wish to communicate?	Produce, create, express	What does the writer really want to say?	Produce, create..	Respond and act creatively	Respond and act creatively
		Fill-in Transformation Multiple choice	Quest-answer Fill-in Multiple ch.	Selection of gist Making inferences Correcting errors Criticizing Commenting	Role playing Oral presentation	Underlining main points Making inferences Correcting errors Criticizing text	Reproduction	Eg. smiling in this context. X says to Y: "How about some coffee?" Y smiles and transforms implied request into an offer	Interview Conversation Summary Narrating Free writing Fluent translation