DOCUMENT RESUME

ED 107 722 TM 004 580

AUTHOR Hambleton, Ronald K.; And Others
TITLE Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments.
PUB DATE [Apr 75]
NOTE 102p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$5.70 PLUS POSTAGE
DESCRIPTORS *Course Objectives; *Criterion Referenced Tests; Individualized Instruction; Item Analysis; Literature Reviews; *Measurement Techniques; Psychometrics; Research Needs; Scores; Statistical Analysis; Task Analysis; Test Construction; Testing; *Test Reliability; *Test Validity
IDENTIFIERS Tailored Testing

ABSTRACT

The success of objectives-based programs depends to a considerable extent on how effectively students and teachers assess mastery of objectives and make decisions for future instruction. While educators disagree on the usefulness of criterion-referenced tests the position taken in this monograph is that criterion-referenced tests are useful, and that their usefulness will be enhanced by developing testing methods and decision procedures specifically designed for their use within the context of objectives-based programs. This monograph serves as a review and an integration of existing literature relating to the theory and practice of criterion-referenced testing with an emphasis on psychometric and statistical matters, and provides a foundation on which to design further research studies. Specifically, the material is organized around the following topics: Definitions of criterion-referenced tests and measurements, test development and validation, statistical issues in criterion-referenced measurement, selected psychometric issues, tailored testing research, description of a typical objectives-based program, and suggestions for further research. The two types of criterion-referenced tests focused on are: Estimation of "mastery scores" or "domain scores", and the allocation of individuals to "mastery states" on the objectives in a program. (Author/BJG)

-Symposium Handout-

Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments

Chairman

David L. Passmore

Presenters

Ronald K. Hambleton
Hariharan Swaminathan
James Algina
Douglas Coulson

The chairman and presenters are from the Laboratory of Psychometric and Evaluative Research at the University of Massachusetts, Amherst

Discussants

Ross E. Traub
Ontario Institute for Studies in Education

and

Thomas Donlon
Educational Testing Service

(An invited symposium presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.)

Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments[1]

Ronald K. Hambleton
Hariharan Swaminathan
James Algina
Douglas Coulson
University of Massachusetts

With the need for significant changes in our elementary and secondary schools clearly documented by Project Talent data (Flanagan, Davis, Dailey, Shaycoft, Orr, Goldberg, & Neyman, 1964), we have seen the development and implementation of a diverse collection of alternative educational programs that seek to improve the quality of education by individualizing instruction (Gibbons, 1970; Gronlund, 1974; Heathers, 1972). A common characteristic of many of the new programs is that the curriculum is defined in terms of instructional objectives; a program specified in such a way is referred to as objectives-based. The overall goal of an objectives-based instructional program is to provide an educational program which is maximally adaptive to the requirements of the individual learner. The instructional objectives specify the curriculum and serve as a basis for the development of curriculum materials and achievement tests. Among the best examples of objectives-based programs are Individually Prescribed Instruction (Glaser, 1968, 1970); Program for Learning in

---

[1]This material is an integration of previously published articles by the authors with several of their new contributions. In addition, an attempt was made to place the total material in a broader context of developments to the criterion-referenced testing field.

Accordance with Needs (Flanagan, 1967, 1969) and the Individualized Mathematics Curriculum Project (DeVault, Kriewall, Buchanan, & Quilling, 1969).

Unfortunately, while considerable progress has been made in important areas such as the construction of instructional materials, curriculum design, and computer management, until quite recently (Glaser & Nitko, 1971; Harris, Alkin, & Popham, 1974; Millman, 1974) there have been few reliable guidelines for test construction, test assessment, and test score interpretation, and this in turn has hampered effective implementation of the programs. One of the underlying premises of objectives-based programs is that effective instruction depends, in part, on a knowledge of what skills the student has. It follows that the tests used to monitor student progress should be closely matched to the instruction. Over the years, standard procedures for testing and measurement within the context of traditional educational programs have become well-known to educators; however, the procedures are much less appropriate for use within objectives-based programs (Glaser, 1963; Hambleton & Novick, 1973; Popham & Husek, 1969).

As an alternative, we have seen the introduction of criterion-referenced tests, which are intended to meet the testing and measurement requirements of the new objectives-based programs. In view of the importance of criterion-referenced testing to the success of objectives-based programs, and their newness, it is perhaps not surprising to note the many articles written on the topic and that these articles typically reflect diverse points of view concerning criterion-referenced test definitions, methods of test development, assessment of psychometric properties, and so on. Now with the

important integrating works of Glaser and Nitko (1971), Millman (1974), and Harris, et al. (1974), terminology has been standardized, issues delineated, and many important technical developments identified.

Purposes

Clearly, the success of objectives-based programs depends to a considerable extent upon how effectively students and teachers assess mastery of objectives and make decisions for future instruction. While not all educat.rs agree on the usefulness of criterion-referenced tests (Block, 1971; Ebel, 1971), the position taken in this monograph is that criterion-referenced tests are useful, and that their usefulness will be enhanced by developing testing methods and decision procedures specifically designed for their use within the context of objectives-based programs. Our monograph is intended to serve as a review and an integration of existing literature relating to the theory and practice of criterion-referenced testing with an emphasis on psychometric and statistical matters, and to provide a solid foundation on which to design further research studies. Specifically, the material in the monograph is organized around the following topics: Definitions of criterion-referenced tests and measurements, test development and validation, statistical issues in criterion-referenced measurement, selected psychometric issues, tailored testing research, description of a typical objectives-based program, and suggestions for further research. Whereas there are a multitude of uses for criterion-referenced tests, we have chosen to provide a concentrated study in this monograph of only two: Estimation of "mastery scores" or "domain scores", and the allocation of individuals to "mastery states" on the objectives in a program. Both criterion-referenced test uses directly concern the day-to day management of students through an

objectives-based program.

The monograph is intended to serve as a companion paper to the review by Hambleton (1974) on testing and decision-making procedures within selected objectives-based programs, and to provide an expanded discussion of one of the four major areas of use of criterion-referenced tests described in the excellent monograph by Millman (1974). Millman indicated four major areas of use (needs assessment, individualized instruction, program evaluation, and teacher improvement and personnel evaluation) and there may be others. However, we have limited our discussion to the use of criterion-referenced tests within the context of individualized instructional programs, although the extension to other areas, in some cases, is obvious. Our work also serves as a second response to some of the technical measurement problems posed by Harris, et al. (1974).

Definitions of Criterion-Referenced Tests and Measurements

A criterion-referenced test has been defined in a multitude of ways in the literature. (See, for example, Glaser & Nitko, 1971; Harris & Stewart, 1971; Ivens, 1970; Kriewall, 1969; and Livingston, 1972a.) The intentionally most restrictive definition of a criterion-referenced test was proposed by Harris & Stewart (1971): "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed [p.1]." On the other hand, possibly the least restrictive definition is that by Ivens (1970) who defined a criterion-referenced test as one "comprised of items keyed to a set of behavioral objectives [p.2]." Given the current state of the art, Iven's definition would correspond to what we refer now to as an "objectives-based test" (Donlon, 1974; Millman, 1974) and this kind of test is not going to allow us to make the strongest kind of criterion-referenced interpretation, i.e. treat the score as an indication of the examinee's level of mastery in some well-specified content domain (Traub, 1972). A very useful definition has been proposed by Glaser and Nitko (1971): "A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards." According to Glaser and Nitko, "The performance standards are usually specified by defining some domain of tasks that the student should perform. Representative samples of tasks from this domain are organized into a test. Measurements are taken and are used to make a statement about the performance of each individual relative to that domain [p.653]."

If one accepts the Glaser and Nitko definition of a criterion-referenced test, it is apparent that the test may be constructed of items from more than one domain. An assessment of mastery or an instructional decision for each individual is then made on the basis of the student's performance on items from each domain. Major interest thus rests on the reliability and validity of domain scores. (For more on this, see Baker, 1974; Bormuth, 1970; Hively, Patterson, & Page, 1968; Glaser & Nitko, 1971; Millman, 1974; Popham, 1974; Skager, 1974.)

Following the Glaser and Nitko definition, the construction of a criterion-referenced test requires the sampling of items from well-specified domains of items. The domain "may be extensive or a single, narrow objective, but it must be well defined, which means that content and format limits must be well specified" (Millman, 1974). The specification of the domain is crucial for putting together a criterion-referenced test since only then the criterion-referenced test scores can be interpreted most directly in terms of knowledge of performance tasks. It should be noted that the word "criterion" does not refer to a criterion in the sense of a normative standard but rather to the minimal acceptable level of functioning that an examinee must achieve in order to be assigned to a mastery state on each domain included in the test. Therefore, the term, domain-referenced test, may be less ambiguous than the term, criterion-referenced test. Furthermore, the term "criterion-referenced" may imply that the only use for the test is to make mastery decisions. Estimation of domain scores is another important use.

Distinctions Among Testing Instruments and Measurements

With the availability of a test theory for norm-referenced measurements (e.g., see Lord & Novick, 1968), we have procedures for constructing appropriate measuring instruments, i.e., norm-referenced tests. Do objectives-based programs which require different kinds of measurement (i.e., criterion-referenced measurement) also require new kinds of tests or will the usual norm-referenced tests with alternate procedures for interpreting test scores be appropriate? There is little doubt that different tests are needed, constructed to meet quite different specifications than those typically set for norm-referenced tests (Glaser, 1963). However, it should be noted that a norm-referenced test can be used for criterion-referenced measurement, albeit with some difficulty, since the selection of items is such that many objectives will very likely not be covered on the test or, at best, will be covered with only a few items. It has been noted by at least two writers (Millman, 1974; Traub, 1972) that when items in a norm-referenced test can be matched to objectives, criterion-referenced interpretations of the scores are possible, although they are quite limited in generalizability. A criterion-referenced test constructed by procedures especially designed to facilitate criterion-referenced measurement can and sometimes is used to make norm-referenced measurements. However, a criterion-referenced test is not constructed specifically to maximize the variability of test scores (whereas a norm-referenced test is). Thus, since the distribution of scores on a criterion-referenced test will tend to be more homogeneous, it is obvious that such a test will be less useful for ordering individuals on the measured

ability. In summary, a norm-referenced test can be used to make criterion-referenced measurements, and a criterion-referenced test can be used to make norm-referenced measurements, but neither usage will be particularly satisfactory.

It has been argued that to refer to tests either as norm-referenced or criterion-referenced may be misleading since measurements obtained from either testing instrument can be given a norm-referenced interpretation, criterion-referenced interpretation, or both. The important distinction made was that between norm-referenced measurement and criterion-referenced measurement (Glaser, 1963; Hambleton & Novick, 1973). From a historical perspective, this distinction was important since a methodology for constructing criterion-referenced tests did not exist, at least at the time of Glaser's article. Criterion-referenced tests were constructed in the same manner as norm-referenced tests, and as pointed out above, the usage was not satisfactory. However, in view of the recent developments in the field, it may not be misleading to label tests as either criterion-referenced or norm-referenced. In fact, given the operational definitions, the distinction between criterion-referenced tests and norm-referenced tests may not only be unambiguous but also meaningful.

Further distinctions between norm-referenced and criterion-referenced tests and measurements have been presented by Block (1971), Carver (1974), Ebel (1962, 1971), Glaser and Nitko (1971), Harris (1974a), Hieronymous (1972), Messick (1974), and Popham and Husek (1969).

Estimation of Domain Scores and Allocation of Individuals to Mastery States

Assume that a criterion-referenced test is constructed by randomly sampling items from a well-defined domain of items. There are two basic uses for which the scores obtained from the criterion-referenced test are ideally suited.

Supposing that a student has a true score π, defined, say, as the proportion of items in the domain of items that a student can correctly answer, the problem is to obtain an estimate $\hat{\pi}$ of his score π based on his performance on a random sample of items from the domain. (The true score π need not be defined as the proportion of correct items. Other definitions may be suitable.) Millman (1974) has aptly termed this the "estimation of domain scores." (Other terms for domain score are "level of functioning score" and "true mastery score.") There are several approaches for the estimation of π, and we shall return to a discussion of these estimates in a later section.

The other use of the scores derived from a criterion-referenced test is consistent with the notion that testing is a decision process (Cronbach & Glaser, 1965). It makes sense to assume that each examinee has a true mastery state on each objective covered in the criterion-referenced test. Typically, a cut-off score or threshold score is set to permit the decision-maker to assign examinees, on the basis of their performance on each subset of items measuring an objective covered in the criterion-referenced test, into one of two mutually exclusive categories - masters and non-masters. Here, the examiner's problem is to locate each examinee into the correct mas-

tery category. For the purposes of this discussion, let us assume that there are just two mastery states: Masters and non-masters. (In a later section, we will extend the discussion to include the problem of assigning an examinee into one of k mastery states.) There are two kinds of errors that occur in this classification problem: False-positives and false-negatives. A false-positive error occurs when the examiner estimates an examinee's ability to be above the cutting score when, in fact, it is not. A false-negative error occurs when the examiner estimates an examinee's ability to be below the cutting score when the reverse is true. The seriousness of making a false-positive error depends to some extent on the structure of the instructional objectives. It would seem that this kind of error has the most serious effect on program efficiency when the instructional objectives are hierarchical in nature. On the other hand, the seriousness of making a false-negative error would seem to depend on the length of time a student would be assigned to a remedial program because of his low test performance. The minimization of expected loss would then depend, in the usual way, on the specified losses and the probabilities of incorrect classification. This is then a straightforward exercise in the minimization of what we would call threshold loss. Complete details for assigning examinees to mastery states are described in a later section.

Test Development and Validation

Introduction

In this section of the monograph, we put forth procedures for constructing valid domain-referenced tests. Such tests are used for much different purposes than norm-referenced tests and, consequently, the procedures needed to develop and validate domain-referenced tests will also be different.

In view of the purposes of domain-referenced tests presented in this monograph, content validity becomes the center of validation concerns. While it is appropriate to study the other validites of a domain-referenced test, it is essential that the content validity be carefully established in order that the test yield meaningful scores. Indeed some aspects of the construction process also serve to content validate the test. The symbiotic relationship that exists between domain-referenced test construction procedures and content validity is illustrated by Jackson's (1970) remarks:

> ". . ., the term criterion-referenced [here, domain-referenced] will be used here to apply only to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents. . . .The meaningfulness and reproducibility of test scores derives then from the complete specification of the operations used to measure the quantity involved." (p.3)

Jackson's statement implies that a properly constructed domain-referenced test will result in a meaningful score. Thus, the question of validity, specifically content validity, of a domain-referenced test can only be answered within the context of proper construction procedures. More specifically, the problem that is unique to domain-referenced tests is that of linking the test item to the behavioral

referent and this is a content validation problem. Osburn (1968) stresses the importance of this aspect of domain-referenced testing when he made the following remark,

> "What the test is measuring is operationally defined by the universe of content as embodied in the item generating rules. No recourse to response-inferred concepts such as construct validity, predictive validity, underlying factor structure or latent variables is necessary to answer this vital question".

While we agree in part with Osburn's position, we do not completely reject the usefulness of such response-inferred concepts as predictive (or criterion) validity. These concepts will be discussed later in the monograph.

At this point the reader should be reminded of the important differences between norm-referenced tests and domain-referenced tests. In general, the purpose of a norm-referenced test is to discriminate among individuals on some ability continuum. In order to achieve this purpose there needs to be some variability in the scores. It is clear that without variability among the scores no discriminations can be made.

On the other hand, in general, a domain-referenced test may be used to determine an individual's level of functioning or it may be used to make an instructional decision involving the student. Other test uses exist, such as evaluating instruction (Millman, 1974), however, these uses will not be considered in this monograph. The essential aspects of the domain-referenced test in terms of these two uses are that the test items reflect the criterion and that the items were sampled in an appropriate manner from the population of domain items. Variability is not a factor; all the individuals taking the

test could be at a very high level of functioning thus getting most or all the items correct and thereby significantly reducing the variability of scores. However, variability in domain-referenced testing is not a completely useless concept. Indeed, variability will be observed when the sample of examinees is heterogenous in terms of their ability to answer items from a given content domain. By establishing a priori the composition of the examinee sample, the resulting variability will provide additional, helpful information for constructing a good domain-referenced test.

It should also be noted here that the different uses for domain referenced tests do not have differential implications for the construction of the tests. Basically the same construction and content validation procedures are followed regardless of the intended use of the score. However, the intended use of the test will influence the number of items to be selected. This point will be discussed later.

Domain-Referenced Test Construction Steps

Introduction. There are six basic steps in constructing domain-referenced tests: 1. task analysis, 2. definition of the content domain, 3. generation of domain-referenced items, 4. item analysis, 5. item selection, and 6. test reliability and validity. These steps are in close agreement with the steps outlined by Fremer (1974). The remainder of this section will examine in detail each of the domain-referenced test construction steps. These steps will be contrasted, when appropriate, to the analogous norm-referenced test construction step.

Task Analysis. A task analysis separates into manageable components the complex behaviors that are to be tested. Task analysis actu-

ally precedes the test construction process. In domain-referenced testing a task analysis provides a logical basis upon which the content domain definitions may be developed. It puts into perspective the purpose of the test and the characteristics of the examinees.

A simple example of a domain-referenced test task analysis might be a general behavioral objective statement. While behavioral objectives do not provide sufficient detail for writing items, they can serve to delineate the general scope of the content domain. Once the task analysis is completed, the domain-referenced test development steps are a focussing and detailing process.

Definition of the Content Domain. The focussing and detailing process referred to above is essentially defining the content domain. This particular step is the most difficult one as well as the most critical step in constructing a good domain-referenced test. Many approaches to defining a content domain have been suggested in the literature (Osburn, 1968; Hively, et al. 1973; Bormuth, 1970; Guttman and Schlesinger, 1966; Popham, 1974).

Recall that a central factor of a domain-referenced test is that its items are linked to the con[illegible] domain in such a way that responses to the items yield informat[illegible] astery of that domain. However, this essential fact is the so[illegible] of a significant difficulty. Put simply, the difficulty is in establishing a content domain that on the one hand permits explicit items to be written from it and on the other hand is not itself trivial (Ebel, 1971). Establishing a domain is a content specification problem and is closely linked to problems in the discussion that follows.

Our position is to seek a balance between those procedures that specify content via item generation rules (Bormuth, 1970; Hively, et al. 1973) and other procedures that begin with behavioral objectives too general to yield domain-referenced items. The reason for this position is that, first, content delineation that is item specific is too restrictive to be educationally useful, and second, a meaningful domain-referenced interpretation of the scores is not possible with generally stated objectives.

Specifically, we believe that Popham's (1974) notion of an amplified objective provides an excellent balance between the clarity achieved with item generation schemes and the practicality of behavioral objectives. Thus, amplified objectives represent a compromise position in the clarity-practicality dilemma and as such, they are likely to represent the approach adopted by individuals interested in developing domain-referenced tests. The compromise seems essential since it does not appear likely that the notion of specifying content via the use of item generation rules will be applicable to many subject areas. Certainly to date little progress has been made along these lines although as Millman (1974) notes "The task is very difficult, but we have just not had enough experience constructing tests, such as DRT's, to know [the limitations of the approach]".

According to Millman (1974), "An amplified objective is an expanded statement of an educational goal which provides boundary specifications regarding testing situations, response alternatives and criteria of correctness." The amplified objective defines the content to be dealt with, the response format and criteria of correctness. The important aspect of these guidelines is that they are

specific; it is not necessary, however, that they specify a homogeneous content area. Specificity and homogeneity are different concepts. Millman (1974) makes this point, "The domain being referenced by a criterion-referenced test may be extensive or a single, narrow objective, but it must be well defined, which means that content and formal limits must be well specified".

An example of an amplified objective taken from Popham (1974) is:

> "When presented with a series of the following types of statements concerning U.S. - Cuba relationships, the learner will correctly identify those which are true:
>
> a. Economic: dealing with size of mutual imports of tobacco, rice, sugar, wheat for the period 1925-1955.
> b. Political: dealing with status of formal diplomatic relationships from 1925 to the present.
> c. Military: dealing with the post-Castro period emphasizing the Bay of Pigs incident and the USSR missile crises."

Popham says that we may further "amplify" this objective by specifying the kinds of true or false items to be used. Further, it should be noted that even by limiting the set of meaningful test items using amplified objectives there still exists the danger of developing a trivial set of items (Popham, 1974).

Before examining the next step in domain-referenced test construction it would be worthwhile to note that the content domain defined for a norm-referenced test (that is, a test constructed to facilitate norm-referenced interpretations) would seldom be as explicitly defined. However, it would be quite incorrect to state, as some writers have, that the content domain of items for a norm-referenced test is not well-defined. In many cases, it is very well-defined, but not to the same extent as is necessary for the

construction of domain-referenced tests.

Generation of Domain-Referenced Items. Once the domain is defined, the test constructor must generate test items. If the domain were defined in a perfectly precise manner, then the item themselves would not need to be generated. The items would simply be a logical consequence of the domain definition. Unfortunately, however, such precision may never be achieved in practice and we must, therefore, generate items and then develop procedures to check the quality of these items. Examining the quality of the items falls under the next section, item analysis.

Even without a perfectly precise specification of the content domain the test constructor should have an excellent idea of item content and format from the statement of the amplified objective. At this stage of the test construction process the item writer would study the amplified objective and generate a set of items that were believed to reflect the domain specified by the amplified objective. After generating a set of domain-referenced test items in this manner, it is necessary to determine the quality of the items through item analysis procedures described below.

Item Analysis. Generally speaking, the quality of domain-referenced items is determined by the extent to which they reflect, in terms of their content, the domain from which they were derived. Because the domain specification is never completely precise, we must determine the quality of the items in a context independent from the process by which the items were generated. Specifically, what is needed are procedures that will determine the extent to which the items reflect the content domain.

There are two general approaches that may be used to establish the content validity of domain-referenced test items. The first approach involves judging each item by content specialists. The judgements that are made concern the extent of the "match" between the test items and the domains they are designed to measure.

The second item analysis procedure is to apply suggested empirical techniques that have been frequently used in norm-referenced test construction along with some new empirical procedures that have been developed exclusively for use within criterion-referenced test development projects. However, it is important to state that we do not advocate the use of empirical methods to select items that would comprise a particular domain-referenced test. We take this position for two reasons. First, selecting items for a domain-referenced test on the basis of their statistical properties would destroy the requirement that the items are representative of the domain of items. Hence, the proper interpretation of domain-referenced test scores would not be possible. Second, empirical methods provide useful information for detecting "bad" items, but the information by itself, is not sufficient to establish the validity of the domain-referenced test items. Here we highlight some of the important aspects of these two approaches; a more detailed discussion may be found in Coulson and Hambleton (1974) and Rovinelli and Hambleton (1973).

(a) Content Specialist Ratings. Probably the most common approach to item validation, although it is fraught with problems, involves the judgements of two content specialists. One suggested procedure is as follows: We first choose two independent and qualified content specialists to judge the quality of the items. Concurrently the test developer has

drawn up a set of items to measure each of several amplified objectives. The rating data is gathered in the following way. A sheet is prepared with a brief paragraph on the top that describes the objective. Below the description of the instructional objective a single question would appear. For example:

> Below are 10 test items that are believed to measure the instructional objective described above. Please rate each item on a scale from 1 to 4 according to the question below.
>
> "How appropriate or relevant is the item for the instructional objective described above?"
>
> 1. Not at all relevant
> 2. Somewhat relevant
> 3. Quite relevant
> 4. Extremely relevant.

The data collected from the two content specialists is arranged into a contingency table with general element p_{ij} equal to the proportion of items that were classified in category i (1, 2, 3, or 4 above) by the first specialist and category j by the second.

An intuitively appealing measure of agreement between the classification of items made by the content specialists is

$$\sum_{i=1}^{k} p_{ii},$$

where p_{ii} is the proportion of items placed in the ith category by each content specialist and k(=4) is the number of categories. However, this measure of agreement does not take into account the agreement that could be expected by chance alone, and hence does not seem entirely appropriate. The coefficient kappa introduced by Cohen

(1960) takes into account this chance agreement and thus appears to be somewhat more appropriate.

One disadvantage to the approach discussed above is that it cannot be used to provide explicit statistical information on the agreement of judgements for each item. With the availability of more content specialists (i.e., perhaps 10 or more), such information could be obtained. Indeed there exist a multiple of rating forms and statistics to assess the level of agreement among content specialists on the match between items and objectives [for example, see Goodman and Kruskal (1954); Light (1973); Lu (1971); Maxwell and Pilliner (1968).] Applications of these statistics to problems of item validation have been described by Coulson and Hambleton (1974).

(b) Empirical Methods. Empirical methods, such as using discrimination indices (Cox & Vargas, 1966; Crehan, 1974; Wedman, 1973), may provide useful information for detecting "bad" items. Indeed Wedman (1973) gives a compelling argument for using empirical procedures. He argues that even careful domain definition and precise item generation specifications never completely eliminate the subjective judgments that, to great and lesser degrees, influence the test construction process. In order to guard against this subjective element, albeit small, we should complement the domain definition and item generating procedures with empirical evidence on the items.

Essentially, empirical procedures involve the use of various item statistics that measure item difficulty and item discrimination. In all instances, for these statistics to be meaningful, it is necessary to have some item variability across examinees.

There has been some discussion recently on the matter of item and test variance with criterion-referenced tests (Haladyna, 1974;

Millman & Popham, 1974; Woodson, 1974). Our own view, which is in agreement with Millman and Popham (1974) is that item and test variance is unnecessary with a domain-referenced test. The "quality" of the test is determined by the extent of the match between the items in the test and the domain they are intended to measure, and of course whether or not the items represent a random sample of items from the domain of items. From this point of view, item and test variance play no role in the determination of the validity of the test for estimating domain scores. On the other hand, one would expect some variability of scores across a pool of examinees consisting of "masters" and "non-masters" and to the extent that there was no (or limited) variability we might suspect that something was wrong with the test. The test ought to reflect some variability of scores across "masters" and "non-masters" groups although one would not select items to maximize this difference since this would distort the process of estimating domain scores.

(b1) Standard Item Indices. There are a number of standard statistical indices which appear to provide information which can be used to ascertain whether the items are measures of the instructional objectives. When items in a domain are expected to be relatively homogeneous, and there are many times when this is not a reasonable assumption (Macready & Merwin, 1973), it has become a fairly common practice for the test developer to compare estimates of item difficulty parameters, or item discrimination parameters, or both. Since one would expect items measuring an objective equally well to have similar item parameters, estimates of the parameters are compared to detect items that deviate from the norm. Such "deviant" items are given

careful scrutiny. In particular, content specialists' judgments of the item are considered along with the empirical evidence. If the items look acceptable, they are returned to the item domain. A more formal method of comparing item difficulty parameters is considered next.

Brennan and Stolurow (1971) present a set of rules for identifying criterion-referenced test items which are in need of revision. The decision process which they established for deciding which items to revise can be used to determine item validity. However, our particular interest is with their procedure for comparing difficulty levels of items intended to measure the same objective. Brennan and Stolurow (1971) state that the item scores from criterion-referenced tests will most likely not be normally distributed. Therefore, in order to determine if the item difficulties are equal, they propose the use of Cochran's Q test. This statistic can be used to determine whether two or more item difficulties differ significantly among themselves. Cochran's Q is a test of the hypothesis of equal correlated proportions. For a large enough sample of examinees, Q is approximately distributed as a χ^2 variable with n-1 degrees of freedom where n is the number of test items. Rejection of the null hypothesis, however, provides no guidance as to which items are significantly different. This can be achieved by setting up confidence bands for each pair of items.

(b2) Item Change Statistic. The difference between the difficulty level of an item before and after instruction describes another item statistic that seems to have some usefulness in the validation of domain-referenced test items. However, an important point to note is that a large difference between the pretest and posttest item difficulty is not necessary since items may be valid but because of poor instruction, there may be

very little change in difficulty level between the two test administrations. But an analysis of the change in item difficulty is an indication of the validity of the test items. Assuming instruction is effective, one would expect to see a substantial change in item difficulty, if the item is a measure of the intended objective. With several items intended to measure the same objective, one could also compare the item change indices for the purpose of detecting items that seem to be operating differently than the others.

Popham (1971) has proposed a two pronged approach for developing adequate domain-referenced test items: An a priori and a posteriori approach. The a priori approach corresponds to the determination of validity by operationally generating items from an amplified objective. The a posteriori approach consists of empirically determining whether or not items are defective. In his discussion of the a posteriori approach, Popham presented a new means for empirically evaluating criterion-referenced test items. This procedure represents an extension of the item change statistic and consists of constructing the following fourfold table from the results of a pre-posttest administration of a set of items measuring an objective:

		Posttest	
		Incorrect	Correct
Pretest	Incorrect	A	B
	Correct	C	D

A, B, C, and D represent the percentage of examinees obtaining each of the four possible response patterns for an item on the two test administrations.

One then computes the median value across items designed to measure the same objective for each of the four cells. These values are used as expected values and a chi-square statistic is computed for each item by comparing the observed percentages in the four-fold table with the expected values.

This chi-square analysis is used to determine the extent to which the items are homogeneous. Popham states that this procedure was more accurate than visual scanning in locating the atypical items. While Popham (1971) describes other descriptive statistics for use in item analysis, the chi-square analysis for detecting "bad" items seems to be the most promising of his suggestions.

Item Selection. The next step in the test construction process is to select a sample of items from the population of "valid" items defining the domain.

A prior question to the selection of test items is the determination of test length. Since this issue is discussed in some detail in a later section , it suffices to say here that test length is specified to achieve some desired level of "accuracy" of test usage. The particular method of assessin accuracy is of course dependent on the intended use of the test scores- estimating domain scores or allocating examinees to mastery states. (For example, see Fhaner, 1974, for an interesting solution to the latter problem, or Kriewall, 1969, 1972.)

Item selection is essentially a straight forward process and involves the random selection of items from the domain of valid test items that measure the objective. In the case of a complex domain, the test developer may resort to selecting items on the basis of a stratified random sampling plan to achieve a "better" selection of items. It is precisely this

feature of random selection of items from a well-specified domain of items that makes it possible for "strong" criterion-referenced interpretations of the test scores (Millman, 1974; Traub, 1972). Clearly, it is exactly this kind of interpretation that so many educators desire to make. Failure to either completely specify the domain of items measuring an objective or to select items in a random fashion from that domain will vitiate against an appropriate criterion-referenced interpretation of an examinee's test performance.

Test Reliability and Validity. The problem of establishing domain-referenced test reliability will be considered in a later section of the monograph.

If procedures described earlier are followed closely, content validity should be guaranteed. Nevertheless, it would be desirable to check the content validity and this can be done using a technique described by Cronbach (1971).

The Cronbach method involves two independent test constructors (or teams of test constructors) developing a domain-referenced test from the same domain specifications. The two resulting tests are then administered to the same group of examinees and a correlation coefficient is computed between the two sets of domain-referenced test scores. The correlation coefficient provides a statistical indication of the content validity of the test.

The main disadvantage of this procedure is that it requires that two domain-referenced tests be constructed. If the two tests were constructed along the guidelines suggested here, the correlation study would be rather expensive to conduct.

When the criterion-referenced tests are being used to make instructional decisions, studies should also be designed to investigate their predictive validities. (For more on this, see Brennan, 1974; Millman, 1974.)

Statistical Issues in Criterion-Referenced Measurement

Estimation of Examinee Domain Scores

There are several methods available for the estimation of a domain score for an individual. The basic problem is, given an examinee's observed score on a criterion-referenced test, to determine his score had he been administered all the items in the domain of items.

(a) Proportion-Correct Estimate

The simplest and the most obvious estimate of the ith examinee's true mastery score, π_i, defined as the proportion of items in the domain of items measuring the objective that the examinee can answer correctly, is his observed proportion score, $\hat{\pi}_i$. This estimate is obtained by dividing the examinee's test score, x_i (the number of items answered correctly), by the total number, n, of the items measuring the objective included in the test. Appealing as it may seem in view of the fact that the proportion-correct score is an unbiased estimate of the true mastery or domain score, this estimate is extremely unreliable when the number of items on which the estimate is based is small. For this reason, procedures that take into account other available information in order to produce improved estimates, especially in the case when there are only few items in the test, would be more desirable.

(b) Classical Model II Estimate

One of the first attempts to produce an estimate of the true score of an examinee using the information obtained from the group to which an individual belongs was made by Kelley in 1927. This is

the well-known regression estimate of true score (Lord and Novick, 1968, pp. 65), which is the weighted sum of two components - one based on the examinee's observed score and the other based on the mean of the group to which he belongs. Jackson (1972) modified this procedure for use with binary data, by transforming the test score x_i into g_i via the arcsine transformation, known as the Freeman-Tukey transformation, given by

$$g_i = \frac{1}{2}\left(\sin^{-1}\sqrt{\frac{x_i}{n+1}} + \sin^{-1}\sqrt{\frac{x_{i+1}}{n+1}}\right) . \qquad (1)$$

As a result of this transformation, the true mastery score is transformed onto γ_i, where,

$$\gamma_i = \sin^{-1}\sqrt{\pi_i} . \qquad (2)$$

If $.15 \leq \pi_i \leq .85$, and if n, the number of test items, is at least eight, then the distribution of g_i is approximately normal with a mean approximately equal to the transformed true mastery score, γ_i, and known variance

$$v = (4n + 2)^{-1} .$$

The model II estimate, or the Jackson estimate becomes, in terms of γ,

$$\hat{\gamma}_i = [g_i \hat{\zeta} + (4n + 2)^{-1} g_.] \; / \; [\hat{\zeta} + (4n + 2)^{-1}] , \qquad (3)$$

where $g_.$, the sample mean based on a sample of N examinees is given by

$$g_{\cdot} = N^{-1} \sum_{i=1}^{N} g_i \ , \tag{4}$$

and $\hat{\phi}$, the sample variance of the γ's, is given by

$$\hat{\phi} = (N - 1)^{-1} \sum_{i=1}^{N} (g_i - g_{\cdot})^2 - (4n + 2)^{-1} \ . \tag{5}$$

Once $\hat{\gamma}_i$ is obtained, $\hat{\pi}_i$ is determined from the expression

$$\hat{\pi}_i = (1 + .5/n) \sin^2 \hat{\gamma}_i - .25/n. \tag{6}$$

For a detailed discussion of this estimate, the reader is referred to Novick and Jackson (1974, pp. 352) and Novick, Lewis, & Jackson (1973).

(c) Bayesian Model II Estimate

The Jackson estimate given above is not ideal since it does not take into account any prior information that may be available. In addition, it may happen that $\hat{\phi}$ estimated using (5) is negative, in which case the solution will not be meaningful. Novick et al. (1973) utilizing the transformations (1) and (2), obtained a Bayesian solution for the estimation of the mastery score that not only takes into account the direct and collateral information, but also any prior information that may be available. In addition, this procedure avoids the problem of negative estimates for ϕ.

Since the distribution of g_i has known variance but unknown mean γ_i, the distribution of g_i is customarily expressed as a conditional distribution i.e.,

$$g_i \mid \gamma_i \sim N(\gamma_i, v) \tag{7}$$

where $N(\gamma_i, v)$ represents the normal distribution with mean γ_i and variance v. The Bayesian estimates are based on the revised belief about the parameters after the data are obtained. The revised belief about the parameters after the data are obtained is summarized in the form of the posterior distribution of the parameters.

As a consequence of Bayes Theorem, the posterior joint distribution $h(\gamma_1, \gamma_2, \ldots, \gamma_N \mid \text{Data})$, is readily expressed in terms of the prior distribution $f(\gamma_1, \gamma_2, \ldots, \gamma_N)$ as

$$h(\gamma_1, \gamma_2, \ldots, \gamma_N \mid \text{Data}) \propto g(\text{Data} \mid \gamma_1, \gamma_2, \ldots, \gamma_N)\ f\ (\gamma_1, \gamma_2, \ldots, \gamma_N). \tag{8}$$

The expression $g(\text{Data} \mid \gamma_1, \gamma_2, \ldots, \gamma_N)$ is known as the likelihood function and is a statement of the joint probability of observing the data conditional upon the unknown parameters $\gamma_1, \gamma_2, \ldots, \gamma_N$. The product of the N distributions given by equation (7), where N is the number of examinees in the sample, yields the likelihood function.

In order to obtain the posterior distribution of γ_i, it is necessary to specify the prior knowledge about the distribution of γ_i, or $f(\gamma_1, \gamma_2, \ldots, \gamma_N)$. In order to do this, it is assumed that the transformed "true" scores $\gamma_1, \gamma_2, \ldots, \gamma_N$ of the N individuals are exchangeable. This amounts to saying that the prior belief about one γ_i is no different from the belief about any other γ_j and implies the assumption that γ_i is a random sample from some distribution. In particular, it is assumed that the prior distribution of γ_i is normal with unknown mean α and unknown variance ϕ. Thus, the specification of the prior distribution of γ_i is dependent upon the knowledge of the mean α and the variance ϕ. However, Novick et al. (1973) have suggested that the prior belief

about α may not be important as the specifications of the prior belief about ϕ and may be represented by a uniform distribution. The above authors have further assumed that it is reasonable to represent the belief about ϕ by an inverse chi-square distribution with ν degrees of freedom and scale parameter λ (see Novick and Jackson, 1974, for an extensive discussion of this distribution). Specification of the prior belief about ϕ thus requires the specification of only the two parameters, ν and λ.

Novick et al. (1973) have considered in detail the problem of setting values of the parameters, ν and λ. Based on various considerations, these authors recommend setting $\nu = 8$. The mean $\bar{\phi}$, of the inverse chi-square distribution is given by $\lambda / (\nu-2)$, and once ν is known, λ can be set equal to $(\nu-2)\ \bar{\phi}$. To estimate $\bar{\phi}$ it is necessary to indicate the amount of information that is available about π. This is accomplished by specifying a value M, where M is considered to be the π value of the typical examinee in the sample. The next step is to specify the number of test items, t, that would have to be administered to the examinee in order to obtain as much information about π as is deemed to be available. Now, transformed estimates of π, from a t-item test are distributed normally on the γ-metric with variance $(4t + 2)^{-1}$. Hence, $(4t + 2)^{-1}$ can be taken as an estimate of $\bar{\phi}$ and subsequently λ can be specified.

Specification of ν and λ in essence determines the prior distribution $f(\gamma)$ of $\gamma_1, \gamma_2, \ldots, \gamma_N$. Substituting this in equation (8), Novick et al. (1973) obtained the joint posterior distribution of the parameters, and hence the joint modal estimate of γ_i.

The joint modal estimate γ_i is obtained by solving the equation

$$\gamma_i = \frac{g_i\left[\frac{\lambda + \Sigma(\gamma_i - \gamma_\cdot)^2}{N + \nu - 1}\right] + \gamma_\cdot\left[\frac{1}{(4n + 2)}\right]}{\left[\frac{\lambda + \Sigma(\gamma_i - \gamma_\cdot)^2}{N + \nu - 1}\right] + \left[\frac{1}{(4n + 2)}\right]} \qquad (9)$$

where

$$\gamma_\cdot = N^{-1} \sum_{i=1}^{N} \gamma_i \qquad (10)$$

This equation for γ_i has to be solved iteratively, and has been found (Novick, et al. 1973) to yield a satisfactory solution after only a few iterations.

(d) Marginal Mean Estimate

The Bayesian model II estimate discussed above is useful for making joint decisions about a set of N examinees. However, in criterion-referenced testing situations, separate decisions about each individual have to be made and hence separate or marginal estimates of true mastery or domain scores, are required.

Lewis, Wang, and Novick (1973) have obtained a marginal mean estimate of the true mastery score, given by

$$\hat{\gamma}_i = g_\cdot + \rho^*(g_i - g_\cdot) \quad . \qquad (11)$$

The quantity ρ^* is dependent on the parameters ν and λ and on the data; once the parameters are set, ρ^* can be read directly from tables prepared by Wang (1973). Again, once $\hat{\gamma}_i$ is obtained $\hat{\pi}_i$ is determined using equation (6).

(e) "Quasi" Bayesian Estimates

In obtaining the joint modal estimates and the marginal mean

estimates, Novick, et al. (1973) and Lewis, et al. (1973) assumed that the prior beliefs about α and ϕ could be expressed in the form of distributions. There are several variations to this theme. If instead of specifying the prior beliefs in the form of distributions, values for α and ϕ can be specified on the basis of previous experience, then the expressions corresponding to the Bayesian marginal mean estimates are readily obtained, and these estimates are relatively easy to compute.

These estimates are based on the prior specification of α and ϕ. Specification of α introduces relatively few complications, but the exact specification of ϕ poses a problem. This is not a quantity most practitioners are familiar with. However, the interrogation procedure described by Novick and Jackson (1974) can be effectively used to yield this information. These quasi-Bayesian estimates are derived on the assumptions that, 1. the prior belief about α can be expressed as a uniform distribution, and ϕ can be specified exactly, and, 2. both α and ϕ can be specified exactly. In the first case, it can be shown that the marginal mean estimate $\hat{\gamma}_i$ is given by

$$\hat{\gamma}_i = \frac{g_i \phi + (4n+2)^{-1} g.}{\phi + (4n+2)^{-1}} \quad . \tag{12a}$$

In the second case, the marginal mean estimate, $\hat{\gamma}_i$, becomes

$$\hat{\gamma}_i = \frac{g_i \phi + (4n+2)^{-1} \alpha}{\phi + (4n+2)^{-1}} \quad . \tag{12b}$$

The similarity between the marginal mean estimates (12a) and (12b) and the Jackson estimate (3) is obvious. In fact, it is interesting

to note that the Jackson estimate is in reality an empirical Bayes estimate and a version of it has been given by Rao (1965).

Allocation of Examinees to Mastery States

Let us consider now the situation where one is interested in assigning an examinee to one of several mastery states or categories. In view of the discussion in the last section, it may appear tempting to first estimate the examinee's domain score or mastery score, compare it with the cut-off scores, and then, in the case of two categories, classify the examinees as either a master or a non-master. Unfortunately, this approach is not very satisfactory. The estimates for the domain scores may be based on a loss function completely inappropriate for that associated with making decisions. For instance, the joint modal estimate and the marginal mean estimates are based on a zero-one loss function and a squared-error loss function, respectively. In making decisions, how far the examinee is from, say, the cut-off score is of no concern. Instead, the main concern is whether the examinee is above or below the cutting-score. Hence, an appropriate loss function in the decision-theoretic process is the threshold loss function. This together with losses or costs associated with misclassifications make obvious the fact, that in order to classify students into categories, a decision-theoretic procedure has to be used.

We shall first consider the problem of classifying an examinee into one of two categories. As in the previous section, the observed scores x_i are transformed into g_i by the arc sine transformation. Let $\gamma(=\sin^{-1}\sqrt{\pi})$ denote the transformed domain score π, and π_o to be cut-off score. If γ_o $(=\sin^{-1}\sqrt{\pi_o})$ is the transformed cut-off score, examinees with true scores γ less than γ_o are classified as true non-

masters, and true masters otherwise. Conforming with the notation employed by Hambleton and Novick (1973) we define the two-valued parameter ω to denote the mastery state of the examinee. The parameter ω assumes one of two values, ω_1 or ω_2. If the examinee is a non-master, i.e., if $\gamma < \gamma_o$, we set

$$\omega = \omega_1,$$

and if he is a master, i.e., $\gamma \geq \gamma_o$, we set

$$\omega = \omega_2.$$

Both γ and ω are, of course, unobservable quantities. Our approach is to produce, using Bayesian statistical methods the posterior distribution representing our belief about the location of the parameter γ. Using this distribution and with a cutting score defined, we can produce probabilities representing the chances of an examinee being located in each mastery state.

In classifying an examinee the decision-maker may take one of two actions - retain the examinee for instruction or advance the examinee to the next segment of instruction. The action "retain" will be denoted by a_1 and the action "advance" by a_2. The decision-maker can commit one of two kinds of errors. If the individual is in reality a non-master (in state ω_1), the decision-maker can classify the individual as a master (in state ω_2) or if in reality the individual is a master (in state ω_2), the decision-maker can classify the individual as a non-master (in state ω_1). In order to arrive at a rule for selecting actions a_1 or a_2, it is necessary to specify the losses associated with these two kinds of misclassifications.

Conforming with the usage and notation of decision theory, we

shall employ the notation $L(\omega_i, a_j)$ to denote the non-negative loss function which describes the loss incurred when action a_j is taken for the individual who is in state ω_i. Thus,

$$L(\omega_i, a_2) = \ell_{12},$$

and

$$L(\omega_2, a_1) = \ell_{21}.$$

with

$$L(\omega_1, a_1)\ L(\omega_2, a_2) = 0.$$

A good classification procedure is obviously one which minimizes in some sense or other the total loss incurred. That is, we shall choose that action for which the expected loss

$$E_\omega L(\omega, a)$$

is a minimum.

We see that if action a_1 is taken, then the expected loss, $E_\omega L(\omega, a_1)$, is given by

$$E_\omega L(\omega, a_1) = 0 \cdot \text{Prob}[\omega = \omega_1] + \ell_{21} \text{ Prob } [\omega = \omega_2]$$

$$= \ell_{21} \text{ Prob}[\gamma \geq \gamma_o]. \tag{13}$$

Similarly, if action a_2 is taken, then the expected loss, $E_\omega L(\omega, a_2)$ is given by

$$E_\omega L(\omega, a_2) = \ell_{12} \text{ Prob}[\omega = \omega_1] + 0 \cdot \text{Prob}[\omega = \omega_2]$$

$$= \ell_{12} \text{ Prob}[\gamma < \gamma_o]. \tag{14}$$

We take action a_1 if

$$E_\omega L(\omega, a_1) < E_\omega L(\omega, a_2) ,$$

or equivalently, if

$$\ell_{21} \text{ Prob}[\gamma \geq \gamma_o] < \ell_{12} \text{ Prob}[\gamma < \gamma_o]. \tag{15}$$

Similarly, we take action a_2 if

$$\ell_{12} \text{ Prob}[\gamma < \gamma_o] < \ell_{21} \text{ Prob}[\gamma \geq \gamma_o]. \tag{16}$$

If it so happened that

$$\ell_{12} \text{ Prob}[\gamma < \gamma_o] = \ell_{21} \text{ Prob}[\gamma \geq \gamma_o],$$

we would be indifferent as to which action to take.

Swaminathan, Hambleton, and Algina (1975) generalized this two category problem to one where examinees are classified into one of several categories. Suppose that there are k categories into which the examinees are to be classified and consequently k actions to be taken. For example, when k=3, the decision-maker may be interested in classifying examinees as masters, partial masters, or non-masters. The appropriate actions may be to advance the masters, retain the partial masters for a brief review and retain the non-masters for remedial work.

In order to separate examinees into k categories or k states, $\omega_1, \omega_2, \ldots, \omega_k$, we need k-1 cut-off scores. Denote these by $\pi_{o1}, \pi_{o2}, \ldots, \pi_{ok-1}$. Hence, an examinee is in state ω_1, if his true proportion score π is less than π_{o1}, in state ω_2 if his score π is between π_{o1} and π_{o2}, and so on. In general an examinee is in state

ω_i if $\gamma_{oi-1} \leq \gamma < \gamma_{oi}$. In addition, we denote the set of k actions to be $a_1, a_2, \ldots, a_j, \ldots, a_k$. Action a_j is to be taken if the examinee is classified into state ω_j.

Associated with misclassifications is the loss function $L(\omega_i, a_j)$. If an action a_j is taken for an individual who in reality is in state ω_i, the loss is ℓ_{ij} so that

$$L(\omega_i, a_j) = \ell_{ij}.$$

These losses are conveniently displayed in Table 1. As before, we choose the action which has the smallest expected loss. Here again we utilize the transformation presented in equation (1).

For action a_j, the expected loss is given by

$$E_\omega L(\omega, a_j) = \sum_{p=1}^{k} \ell_{pj} \text{ Prob } [\gamma_{op-1} \leq \gamma < \gamma_{op}] \tag{17}$$

where $\gamma_{oo} = -\infty$, and $\gamma_{ok} = +\infty$. Thus action a_j is chosen if

$$\sum_{p=1}^{k} \ell_{pj} \text{ Prob}[\gamma_{op-1} \leq \gamma < \gamma_{op}] < \sum_{p=1}^{k} \ell_{pm} \text{ Prob}[\gamma_{op-1} \leq \gamma < \gamma_{op}], (m=1,2,\ldots,k, \ m \neq j). \tag{18}$$

The probabilities given in Equations (13) through (18) are really posterior probabilities and should be so stated. Thus,

$$\text{Prob } [\gamma_{op-1} \leq \gamma < \gamma_{op}]$$

in Equation (18) should be written as

$$\text{Prob } [\gamma_{op-1} \leq \gamma < \gamma_{op} \mid \text{Data}] \ .$$

Once the posterior distribution of γ is determined, the above probability is determined as the area under the probability density curve between γ_{op-1} and γ_{op}.

Table 1

Loss Table for a
Multi-Action Problem

State	Action					
	a_1	a_2	...	a_j	...	a_k
ω_1 $(Y < Y_{o1})$	0	ℓ_{12}	...	ℓ_{1j}	...	ℓ_{1k}
ω_2 $(Y_{o1} \leq Y < Y_{o2})$	ℓ_{21}	0	...	ℓ_{2j}	...	ℓ_{2k}
...	...	...	...	...	...	...
ω_i $(Y_{oi-1} \leq Y < Y_{oi})$	ℓ_{i1}	ℓ_{i2}	...	ℓ_{ij}	...	ℓ_{ik}
...	...	...	...	...	...	...
ω_k $(Y_{ok-1} \leq Y)$	ℓ_{k1}	ℓ_{k2}	...	ℓ_{kj}	...	0

The next stage in the decision theoretic process is to obtain this posterior distribution of parameter, γ, for each individual, or, the posterior marginal distribution. The posterior joint distribution of the parameters, given the prior and the likelihood function, is obtained by using Equation (8) given previously. Once the joint distribution is obtained, the marginal distribution is obtained by integrating out all the irrelevant parameters.

Several procedures are available for the determination of posterior marginal distributions and, hence, posterior marginal probabilities. The first method is that given by Lewis et al. (1973). Utilizing the distributions and assumptions given in connection with the Bayesian model II estimates in a previous section, Lewis et al. (1973) derived an approximation to the posterior marginal distribution. They showed that the posterior marginal distribution of γ_i, is approximately normal, i.e.,

$$\gamma_i \mid \text{Data} \sim N(\mu_i, \sigma_i^2) \tag{19}$$

where

$$\mu_i = g_. + \rho^*(g_i - g_.), \tag{20}$$

and

$$\sigma_i^2 = \frac{1 + (N - 1)\,\rho^*}{(4n + 2)\,N} + (g_i - g_.)^2\,\sigma^{*2}. \tag{21}$$

(This approximation is reasonably good when the number of test items

exceeds seven.) The quantity g_i is defined by Equation (4). The quantities ρ^* and σ^{*2} in expressions (20) and (21) are dependent on the parameters ν and λ of the inverse chi-square distribution of λ, and have to be computed by numerical integration. As mentioned earlier, the tables prepared by Wang (1973) can be used so that on specifying ν and λ, ρ^* and σ^{*2} may be obtained.

Returning to the problem of classification of students into k mastery categories, we first transform the (k-1) specified cut-off score π_{op} into γ_{op}, given by

$$\gamma_{op} = \sin^{-1}\sqrt{\pi_{op}}, \qquad p = 1,\ldots,k-1. \tag{22}$$

The next step is to calculate the probabilities of the type given by Equation (16), (17), and (18). It is clear that for any examinee,

$$\text{Prob}[\pi_{op-1} \leq \pi < \pi_{op} \mid \text{Data}] = \text{Prob}[\gamma_{op-1} \leq \gamma < \gamma_{op} \mid \text{Data}]. \tag{23}$$

For the ith examinee, we define the quantity z_{oji} as

$$z_{oji} = \frac{\gamma_{oj} - \mu_i}{\sigma_i}, \tag{24}$$

with μ_i and σ_i^2 defined by Equations (20) and (21). The quantity z_{oji} is merely the normal deviate corresponding to the cut-off score j for examinee i. Since the posterior distribution is approximately normal with mean μ_i and variance σ_i^2,

$$\text{Prob}[\gamma_{op-1} \leq \gamma_i < \gamma_{op} \mid \text{Data}] = \text{Prob}[z_{op-1i} \leq z_i < z_{opi} \mid \text{Data}]. \tag{25}$$

That is, the probability that γ_i is between γ_{op-1} and γ_{op} is approximately equal to the probability that a standardized normal variate is between the z scores z_{op-1} and z_{op}. Hence, for each examinee i, the quantity

$$E_\omega L(\omega,a_j) = \sum_{p=1}^{k} \ell_{pj} \text{ Prob}[z_{op-1i} < \gamma_i < z_{opi} \mid \text{Data}] \qquad (26)$$

is calculated for each action j (j=1, 2,...,k). These k expected losses are than compared with one another, and the action for which the expected loss is the least is chosen as the appropriate action.

In order to illustrate the procedure consider the following hypothetical example. The data and results for this example are summarized in Tables 2 and 3.

Suppose that a set of 10 items representative of the domain of items measuring an objective is administered to a group of 25 examinees, and that the examinees are to be classified into one of three categories, masters, partial masters, and non-masters. The losses associated with wrongly classifying the examinee are given in Table 4. Also, assume that the cut-off scores π_{o1} and π_{o2} are .60 and .80, respectively. First, the observed scores, x_i are transformed into ξ_i, and the cut-off scores π_{o1}, and π_{o2} into γ_{o1}, and γ_{o2}. Next, the prior belief about ϕ is specified. As indicated earlier, this is done by choosing ν and λ, the parameters of the distribution that is used to represent the belief about ϕ. In order to determine ν and λ, the length of the test that would be required to yield as

Table 2

Analysis of a Hypothetical Set of Data: n=10, m=25

Number of Items Correct x_i	Frequency	Transformed Observed Score g_i	Marginal Mean μ_i	Marginal Standard Deviation σ_i
4	2	.695	.836	.121
5	4	.785	.881	.118
6	5	.875	.933	.118
7	4	.980	.989	.115
8	4	1.083	1.043	.115
9	3	1.202	1.107	.118
10	3	1.392	1.211	.125

$\overline{g} = m^{-1}\Sigma g_i = .998$

$\gamma_o = \sin^{-1}\sqrt{\pi_o} = 1.107$

Table 3

Decision Making in a Three-Way Classification Problem

Number of Items Correct	Prob[π_1 < .6 | Data]	Prob[.6 ≤ π_1 < .8 | Data]	Prob[π_1 ≥ .8 | Data]	Expected Losses			Action
				Action a_1	Action a_2	Action a_3	
7	.019	.510	.471	1.452	.509	1.077	Retain Briefly
8	.006	.399	.595	1.589	.607	.816	Retain Briefly
9	.000	.156	.844	1.844	.844	.312	Advance
10	.000	.015	.985	1.985	.985	.030	Advance

Table 4

Losses for the Three-Action Problem

State	Action		
	a_1 (Remedial Work)	a_2 (Brief Review)	a_3 (Advance)
Non-Master	0	2	3
Partial Master	1	0	2
Master	2	1	0

much information as one feels one has about any examinee's true mastery score π_i is decided. Suppose that, it is decided that a five-item test would be required. Hence, t=5 and, $(4t+2)^{-1} = .0454$, is the value for $\bar{\phi}$. Since, in general, a good value for ν is eight, the value for λ is .2727 [$\lambda = (\nu-2)\ \bar{\phi}$]. The tables prepared by Wang (1973) give $\rho^* = .5335$ and $\sigma^{*2} = .0159$. The next step is to compute μ_i and σ_i using equations (20) and (21). Finally, the standardized normal deviate given by equation (24) is obtained and using the tables of the standardized normal distribution the approximate probabilities, Prob[$\pi_i < .6 \mid$ Data], and Prob[$.6 \leq \pi_i < .8 \mid$ Data], Prob[$\pi_i > .8 \mid$ Data], are calculated.

The hypothetical probabilities reported in Table 3 are the probabilities associated with an examinee being in any one of these three categories. These probabilities, when combined with the loss structure presented in Table 4, would result in examinees with seven or eight correct items being retained for a brief review and examinees with a score of nine or ten items correct being moved ahead.

The Bayesian method outlined above is one of several methods that could be used to provide the posterior probabilities necessary for the decision-theoretic approach. Other methods that could be used to produce the posterior probabilities can be developed along the lines indicated in the previous section. One obvious procedure is to obtain the posterior probabilities under the assumption that instead of specifying the prior beliefs about α and ϕ in the form of a distribution, the parameters that characterize the distribution of π_i, values for α and ϕ can be specified exactly. In this case,

the posterior marginal distribution of γ_i is normal with mean

$$\frac{\alpha v + g_i \phi}{\phi + v},$$

and variance

$$\frac{v + \alpha}{v\alpha},$$

i.e.,

$$\gamma_i \mid \alpha, \phi, \text{Data} \sim N\left(\frac{\alpha v + g_i \phi}{\phi + v}, \frac{v + \phi}{v\phi}\right) \tag{27}$$

Once the posterior marginal mean and variances are obtained, the cut-off scores are transformed and the posterior probabilities obtained for each examinee. The expected loss for each action is obtained as given by Equation (26) and the appropriate decisions made.

Another method for obtaining the posterior probabilities is to assume that the variance ϕ of the distribution of γ_i is specified exactly but that the distribution of α is uniform. This test amounts to saying that although we have prior beliefs about ϕ, and we are ignorant about α. In this case, the posterior marginal distribution of γ_i is also normal, and is given by

$$\gamma_i \mid \phi, \text{Data} \sim N\left(\frac{vg_. + g_i \phi}{\phi + v}, \frac{v(\phi + N^{-1}v)}{\phi + v}\right). \tag{28}$$

Again, the posterior probabilities are obtained in the manner described above, and the appropriate decisions made.

The posterior marginal distribution can be obtained more directly if, instead of transforming the observed score x_i into g_i by the arc-

sine transformation, we worked directly with the proportions. In this case, the Beta-binomial analysis outlined by Novick and Jackson (1974) and Novick and Lewis (1974) can be utilized effectively to produce the posterior probabilities. For details of this procedure, we refer the reader to the above references.

It should be pointed out that more recently Lewis, Wang, and Novick (1974) have developed an extension of the procedure for deriving the posterior marginal distribution by incorporating the prior information on the parameter α. They assumed, in addition to all the assumptions made for obtaining the joint modal and marginal mean estimates, that

$$\alpha \sim N(\mu, \phi/\eta) \ . \tag{29}$$

The quantity η together with μ and the parameters λ and ν for specifying the distribution of ϕ have to be supplied by the user. This procedure shows great promise and needs to be studied carefully.

Application of a Bayesian Decision-Theoretic Procedure

The procedures described in the previous section should be feasible with objectives-based programs that have a small computer of the type typically used to manage instruction (see, for example, Baker, 1971). We shall attempt to demonstrate the feasibility of the procedure by briefly outlining the steps a hypothetical instructional designer would take. Let us suppose that an instructional designer is interested in making decisions on students' status with respect to a particular set of program objectives. Test items measuring each objective are organized into a criterion-referenced test and administered to the students. We assume that the test items are binary scored and represent

a random sample of items from the domain of items that measure each objective. For each objective, the designer must specify the number and the location of the mastery states on the mastery score interval [0, 1]. This is done by defining the cutting scores. In addition, the instructional designer specifies the losses attached to classifying an individual incorrectly. A loss matrix of the kind shown in Table 1 is developed and provided to the computer. Some rough guidelines for developing the loss matrix have been described by Hambleton and Novick (1973). Finally, it is necessary for the designer to specify his prior beliefs about the distribution of ability on each objective covered in the test. This is one step where the instructional designer needs to be extremely careful. The effects of poor choice of priors on the decision process is not known at this point, and it remains to be determined under what conditions a poor choice of priors will result in worse decisions than not using Bayesian methods at all. Clearly, further research is necessary to develop efficient methods for accurately assessing prior beliefs.

Using any one of a variety of input devices (i.e., optical scanning sheets, mark sense cards or computer cards) the examinee test item responses are read by the computer and the Bayesian decision theoretic procedure implemented. The computer program can be designed to provide the output necessary to monitor student progress through the instructional program. A statement of domain scores and mastery allocations on objectives for each student can be produced and this information can be used to guide a student through the next segment of his instruction.

The decision-theoretic procedure outlined in the last section provides a framework within which Bayesian statistical methods can be employed with criterion-referenced tests to improve the quality of decision-making in objectives-based instructional programs. The incorporation of losses introduces the decision-maker's values into the decision process. The Bayesian methods incorporate the prior knowledge of the decision maker and utilize the data from all examinees, thereby effectively increasing the amount of information the decision maker has without requiring the administration of additional test items. However, it should be pointed out that research is needed to establish the robustness of the Bayesian statistical model with respect to deviations of the data from the underlying assumptions. We also note that the Bayesian statistical model described in this monograph is only one of several models that could be used (for example, see, Novick and Lewis, 1974, for another) within our decision-theoretic framework. Further study of these additional models would seem to be highly appropriate.

Selected Psychometric Issues

Of fairly obvious concern for both the theory and practice of criterion-referenced measurement are the following issues: (1) concepts of error of measurement, (2) reliability, (3) determination of appropriate test length, and (4) determination of cut-off scores. This section is intended to provide both a review and discussion of the literature concerning each of these issues.

Concepts of Error of Measurement for Criterion-Referenced Tests

A framework for discussing errors of measurement of criterion-referenced tests would need to include at least three dimensions. The first has to do with the use of the test: Estimation of domain score or allocation to mastery states; errors have to be defined differently for these two uses of the test. The second dimension is concerned with the particular view of probability that one adopts. If the view of subjective probability is adopted, the concept of error of measurement is related to the properties of the posterior distribution for the true score that is being estimated. If the frequency view of probability is adopted, then the concept of error of measurement is related to the observed score distribution for the examinee. The final dimension concerns whether information about the error is desired for the individual, the group or both. However, the discussion of measurement error will be principally in terms of the first dimension, although the latter two dimensions will be briefly referred to.

Earlier in the monograph we identified two uses of criterion-referenced tests. In this section we shall first discuss the concept of error associated

with estimating the examinee's domain score. Many theorists in criterion-referenced measurement have insisted that the items on a criterion-referenced test should be interpretable as a random sample from some domain of items that may be described with a high degree of specificity. They argue that when this situation obtains, the observed proportion correct score may be considered to be an unbiased estimate of the domain score. The situation, in which tests are constructed by random sampling from a domain of items, is clearly one example of the class of situations for which generalizability theory was intended (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) The brief treatment of generalizability theory given in chapter eight of Lord and Novick (1968), which is concerned with nominally (or randomly) parallel tests, is sufficient for our limited aims in this monograph.

Lord and Novick (1968) discuss the notion of generic true score which we shall use to define the domain score, π_a, i.e.,

$$\pi_a \equiv E_j Y_{ja} , \tag{30}$$

where Y_{ja} is a random variable for examinee a defined over tests constructed by random sampling of items and E is the expectation operator. The generic error of measurement is

$$\epsilon_{ja} \equiv Y_{ja} - \pi_a \tag{31}$$

which is the deviation of the observed score for examinee a on test j from his generic true score. The generic error of measurement is the

quantity of interest when our purpose is to estimate the examinee's domain score since it contains information about the accuracy of the domain score estimates. Lord and Novick (1968) give the following linear model for the observed score

$$Y_{ja(k)} = \mu + (\pi_a - \mu) + (\tau_j - \mu) + \alpha_{ja} + e_{ja(k)} \qquad (32)$$

where τ_j is the mean of the jth test, α_{ja} is the interaction between person a and test j and $e_{ja(k)}$ is the specific error of measurement on the kth replication of the test. This model implies the identity

$$\varepsilon_{ja} = e_{ja(k)} + (\tau_j - \mu) + \alpha_{ja} \quad . \qquad (33)$$

From the definition of generic error and this identity, Lord and Novick (1968) derive a number of interesting properties for ε_{ja}. One property is

$$E_j \varepsilon_{ja} = 0 \quad , \qquad (34)$$

that is, over randomly sampled tests the expected value of the generic error of measurement is zero and hence the observed score is an unbiased estimate of the domain score. However, the expected value for any given sample of items over replications is given by

$$E_k \varepsilon_{ja} = E_k (e_{ja(k)} + (\tau_j - \mu) + \alpha_{ja}) \qquad (35)$$

$$= \tau_j - \mu + \alpha_{ja} \quad . \qquad (36)$$

Thus, on any administration of test j for person a there is a bias due to

the test difficulty term, $(\tau_j - \mu)$, and the interaction term. It is clear that estimating this bias should be one concern of the users of criterion-referenced tests.

Other important properties of the generic error of measurement may be enumerated. However, rather than listing these properties we refer the reader to Lord and Novick (1968) and point out that the properties of interest depend critically on whether the investigator is interested in group or individual error distributions, and whether the error is defined with respect to replications or randomly parallel tests.

Having defined and discussed to some extent the error of measurement, the important consideration of a loss function arises next. A loss function may be described as a function that weights the error incurred in estimating a parameter, and in this case the loss function weights the error of measurement incurred in estimating a domain score. If we decide that the squared-error loss function provides a reasonable quantification of the loss incurred by the error of measurement, the procedures given in chapter eight of Lord and Novick (1968) will be useful to estimate parameters concerned with the error of measurement.

The above discussion implicitly assumes that the frequency view of probability is adopted. However, it is equally reasonable to consider the "error of measurement" from a subjective view of probability. Within the framework of subjective probability, philosophical considerations imply that the concern should be with the quality of information we have about the individual's true score rather than the "error of measurement." One method of quantifying the quality of information is in terms of the limits of c percent highest density region of the posterior distribution of the

domain score. If we are satisfied with our knowledge that there is a c percent probability that π_a lies within these limits, then the test is providing the information we desire. If the region is too wide, a longer test is required, while if the region is narrower than we require, a shorter test may be used.

In the previous section we introduced a linear model to point out the possible bias in the estimation of an examinee's domain score. To discuss the issue within the framwork of subjective probability, we need to investigate the Bayesian procedures for the analyses of such linear models. The Bayesian models discussed earlier in the monograph may not be appropriate for this purpose since a linear model such as that given by Equation (32) may not be implied by the Bayesian models. Therefore, we will not discuss the possibility of a bias in Bayesian estimators due to an unrepresentative sample of items.

The second purpose of criterion-referenced testing is that of classifying examinees into mutually exclusive categories or mastery states. As outlined earlier, typically k-1 cut-off scores are specified to separate the examinees into k categories. In the case of a single cut-off score, the examinees with domain scores greater than the cut-off score have mastered the instructional material to a desired level of proficiency, while those with domain scores below the cut-off score have not achieved the required level of proficiency. The problem is to use the results of a criterion-referenced test to decide on which side of the cut-off score each examinee's domain score lies.

There are at least two possible concepts for error of measurement when the purpose is to classify individuals into mastery states. The

first concept is based on the accuracy of decisions while the second concept is based on the consistency of decisions made on repeated administrations of a criterion-referenced test. The concept of decision-making accuracy implies that an error occurs whenever an individual is incorrectly classified. A plausible loss function for this error of measurement is the threshold loss function. However, Novick and Lewis (1974) suggest three additional loss functions that might be used:

(1) A threshold loss function with an indifference region in which there is zero loss for false positive or false negative errors,

(2) A negative squared-exponential loss used with the root arcsine transformation parameter,

$$\gamma = \sin^{-1}\sqrt{\pi} \ ,$$

(3) A cumulative Beta distribution loss function.

From the concept of decision-making consistency it follows that errors should be defined in terms of inconsistencies in allocation of examinees to mastery states across repeated administrations of a criterion-referenced test. An error occurs if an examinee is classified in different mastery categories on different administrations of a criterion-referenced test. Here again a threshold loss function is a reasonable loss function. However, again additional loss functions should be considered. In particular, the threshold loss function with an indifference region may be useful.

It should be realized that the concept of error based on decision-

making consistency is very different from that based on decision making accuracy. Inconsistent classifications imply that a misclassification has occurred on one of the classifications, but consistent classifications do not necessarily imply that accurate decisions have been made, for it is entirely possible to be consistently inaccurate. Inaccurate but consistent decisions may occur whenever a Bayesian decision-theoretic procedure is used for classification. The choice of loss ratio, violations of the Bayesian model assumptions, improper specifications of priors, and regression effects acting either alone or in conjunction, can create consistently inaccurate decisions. The possibility of consistently inaccurate decisions also occurs when the sample proportion correct score is used to make classificatory decisions. If we adopt the definition of error of measurement given by Equation (31), then the covariance of the generic errors of measurement over examinees on two tests will in general be non-zero, even though the expected value of such covariances over all pairs of tests in an infinite population of tests will be zero (Lord & Novick, 1968). Since we have correlated errors, the possibility exists that consistently inaccurate decisions may be made on the basis of the observed proportion correct score.

Reliability of Criterion-Referenced Tests

Lord and Novick (1968) point out that the standard error of measurement provides meaningful information about the degree of inaccuracy of a norm-referenced test only when we have knowledge of the observed score variance for the group we are interested in. If we do not, the reliability

coefficient provides more meaningful information. This state of affairs is a reflection of the relative interpretation of norm-referenced test scores. However, properly constructed criterion-referenced tests yield absolute interpretations and when we are estimating domain scores, a quantity such as the standard error of measurement will always provide meaningful information about the degree of inaccuracy of the test (Harris, 1972). Both the probability of misclassification and the probability of inconsistent classification provide needed information about the "reliability" of the test. There have been several reliability indices proposed in the educational measurement literature that are related to decision-making accuracy and decision-making consistency, and some of these are discussed below.

Suppose that we administer a criterion-referenced test to a population of examinees on two occasions and classify the examinees into one of k mutually exclusive mastery states at each administration and denote the proportion of examinees placed in the ith mastery state on the first administration and in the jth mastery state on the second administration, by p_{ij}. An intuitively appealing measure of agreement between the decisions made on the two administrations is

$$\sum_{i=1}^{k} p_{ii} ,$$

where p_{ii} is the proportion of examinees placed in the ith mastery state on both test administrations. However, as noted by Swaminathan, Hambleton, and Algina (1974), this measure of agreement does not take into account the agreement that could be expected by chance alone, and hence does not seem entirely appropriate. The coefficient

κ introduced by Cohen (1960) takes into account this chance agreement and thus appears to be somewhat more appropriate (Swaminathan, et al. 1974). The coefficient κ, an expression for reliability of criterion-referenced tests, is defined as

$$\kappa = (p_o - p_c) / (1 - p_c), \tag{37}$$

where p_o, the observed proportion of agreement is given by

$$p_o = \sum_{i=1}^{k} p_{ii}, \tag{38}$$

and p_c, the expected proportion of agreement is given by

$$p_c = \sum_{i=1}^{k} p_{i.} \, p_{.i} \, . \tag{39}$$

It should be noted that $p_{i.}$ and $p_{.i}$ represent the proportions of examinees assigned to the mastery state i on the first and second test administration, respectively.

Since p_o is the observed proportion of agreement and p_c is the expected proportion of agreement, κ defined in equation (37) can be thought of as the proportion of agreement that exists, over and above that which can be expected by chance alone. It should be stressed that κ is based on the observed and expected proportions along the main diagonal of the joint proportion matrix. It is unaffected by discrepancies that exist in off-diagonal entries (for a further discussion, see Light, 1973).

The properties of κ have been discussed in detail by Cohen (1960, 1968) and Fleiss, Cohen, and Everitt (1969). It suffices to note here that the upper limit of κ is + 1 and may only occur when the marginal

proportions for different administrations are equal. However, if any examinee is classified differently on repeated administrations, the value of κ will be less than +1.

In the derivation of the κ statistic, all inconsistent classifications are weighted equally. The quantity κ_w or weighted Kappa, which was introduced by Cohen (1968) represents an extension which permits differential weighting of different kinds of misclassification.

The work of Swaminathan et al. (1974) clearly is based on the concept of reliability as decision-making consistency. Criterion-referenced test users who adopt these authors' concept and coefficient of reliability should keep firmly in mind that consistent decisions are not necessarily accurate decisions. Also, these authors point out that κ is dependent on factors such as the method for assigning examinees to mastery states, selection of the cutting score, test length and the heterogeneity of the group. Hence, they recommend that when reporting κ, other information such as cutting scores and student ability as measured by the test be reported along with the reliability index.

Harris (1974b) introduced an index of efficiency for a mastery test. Harris argues that a necessary characteristic of a mastery test is that it should sort students into two categories and that if it is a valid test, it should sort students into the correct two categories, as determined by some criterion data. As a consequence, he proposes that, lacking criterion data, it may be informative to examine how well a test sorts students into mastery categories, where the cutting score for classification is some number of items correct. The index of efficiency is defined as

$$\mu_c^2 = \frac{SS_b}{SS_b+SS_w}$$

which is equivalent to a squared point biserial coefficient between total score and a dichotomous variable indicating criterion group. Harris (1974b) points out that the largest μ_c^2 over all possible classifications of the examinees is an upper bound to the validity of the mastery test when validity is measured by an analogous index.

Harris' discussion of the index of efficiency implies that it may serve as a coefficient of decision-making accuracy since, in general, a large μ_c^2 indicates a high decision-making accuracy. However, μ_c^2, interpreted as a coefficient of decision-making accuracy may be misleading in some situations. For instance, if all the examinees are say, masters, μ_c^2 may turn out to be relatively small even though the decisions may be substantially accurate. Thus we would underestimate the utility of the test for making mastery decisions. A situation that plausibly occurs in criterion-referenced testing is to have the test scores have a bimodal distribution. Let us assume that two non-overlapping distributions that accurately indicate mastery occur. If there is any within distribution variability, μ_c^2 will be less than one, but we will be making accurate decisions on the basis of the test. While it is clear that μ_c^2 will be relatively large in this situation, it still underestimates the decision-making accuracy of the test. Finally it may be possible that in using μ_c^2 to compare the decision-making accuracy of two tests, in at least some cases, μ_c^2 may be nigher for the test with which we would make less accurate decisions. These difficulties arise because μ_c^2 is based on a squared error loss function, whereas the threshold loss function appears to be more appropriate when criterion-referenced tests are

used to make mastery decisions. Thus, although the applicability of μ_c^2 to a single test and its ease of computation make it attractive, care in interpretation must be taken if an investigator adopts μ_c^2 as a measure of decision-making accuracy.

Another interesting suggestion for reliability estimation comes from the work of Livingston (1972a, 1972b, 1972c). He proposed a reliability coefficient which is based on squared deviations of scores from the cut-off score rather than the mean as is done in the derivation of reliability for norm-referenced tests in classical test theory. The result is a reliability coefficient which has several of the important properties of a classical estimate of reliability. In fact, it can be easily shown that the classical reliability is simply a special case of the new reliability coefficient. However, several psychometricians (e.g., Harris, 1972; Shavelson, Block, & Ravitch, 1972) have expressed doubts concerning the usefulness of Livingston's reliability estimate. For example, while Livingston's reliability estimate may be higher than a classical reliability estimate for a criterion-referenced test, the standard error of the test is the same, regardless of the approach to reliability estimation. Hambleton and Novick (1973) note that they feel Livingston misses the point for much of criterion-referenced testing. They suggest that it is not "to know how far (a student's) score deviates from a fixed standard." Certainly, Livingston's definition of the purpose of criterion-referenced testing is different from the two primary uses reviewed in this monograph. In fact, we are aware of no objectives-based programs that use criterion-referenced tests in a way suggested by Livingston.

Determination of Test Length

As in classical test theory, test length for a criterion-referenced test is set to achieve some desired level of "accuracy" with the test scores. In the case where estimation of domain scores is of concern, the relationships among domain scores, errors of measurement, and test length as summarized in the item-sampling model are well known (Lord and Novick, 1968) and provide a basis for determining test length.

When using criterion-referenced tests to assign examinees to mastery states, the problem of determining test length is related to the size of misclassification errors one is willing to tolerate. One way to assure low probabilities of misclassification is to make the tests very long. However, since there are a relatively large number of tests administered in objectives-based programs, very long tests are not feasible.

Of course an additional constraint imposed on the determination of test length is the relatively large number of tests that are needed within an objectives-based program and so it would seem useful to study the problem of setting test lengths within a total testing program framework (see for example, Hambleton, 1974).

There have been three approaches to the problem of determining test length reported in the literature. One issue that distinguishes the approaches is the concept of probability that underlies each approach. The Bayesian approach of Novick and Lewis (1974) employs the subjective meaning of probability, while the approaches of Millman (1972, 1973) and of Fahner (1974) employ the frequency view of probability.

Millman (1972, 1973) considered the error properties of mastery

decisions made by comparing an observed proportion correct score with a mastery cut-off score. By introducing the binomial test model, one can determine the probability of misclassification, conditional upon an examinee's true score, an advancement score and the number of items in the test. (Advancement score is distinguished from cut-off score in the following way: The advancement score is the minimum number of items that an examinee needs to answer correctly to be assigned to a mastery state. The cut-off score is the point on the true mastery or domain score scale used to sort examinees into mastery and non-mastery states.) By varying test length and the advancement score, an investigator can determine the test length and advancement score that produces a desired probability of misclassification for a <u>given</u> domain score. The primary problem in applying the tables prepared by Millman (1972) is that one would need to have a good prior estimate of the domain score. Other problems have been suggested by Novick and Lewis (1974): They report that for certain combinations of cut-off scores and test length, changing one or both to decrease the probability of misclassification for those above the cut-off score will actually increase the probability of misclassification for those below the cut-off score. In order to choose the appropriate combination of test length and advancement score, one must have some idea of whether the preponderance of students are above or below the cut-off score and of the relative costs of misclassification. However, the first requirement can only be satisfied with prior information on the ability level of the group of examinees. Novick and Lewis (1974) suggest that is would be useful to have some systematic way of incorporating prior knowledge into the test length determination problem.

Novick and Lewis (1974) provide such a method based on the Bayesian Beta-binomial model. Their approach may be described as follows: For a fixed prior, fixed cut-off score, and fixed loss ratio, identify those combinations of test length and advancement score that "just favor" the decision to classify the examinee as a master. By "just favor" we mean that the difference in expected losses for a mastery classification and a non-mastery classification lies in the interval $[0, -r]$, where r is set by the instructional designer. Then using the two criteria below choose the optimal combination of test length and advancement score:

(1) Disregard test lengths that are absurd in the context that the testing takes place (in all cases test lengths less than 25 items are recommended),

(2) Choose a combination of test length and advancement score that will be reasonable for a class of appropriate prior distributions.

Clearly the results of such a procedure are dependent upon the chosen prior distribution. In fact, because of criterion (2) above the results for any one prior distribution is dependent on the class of appropriate priors. Novick and Lewis (1974) provide these guidelines for choosing priors:

(1) choose a prior such that $E(\cdot) = \pi_0$,

(2) choose priors such that $p(\cdot \geq \pi_0)$ is just greater than .50,

(3) choose a class of priors with properties 1 and 2 but which differ in their variance.

The results also depend on the loss ratio, and the general result is that

longer tests and higher advancement scores are required with greater loss ratios. Also, the results depend on the cut-off score but a general trend does not really emerge.

Novick and Lewis (1974) mention the important trade off between instructional time and testing time. If instructional time is increased, the expected value of the prior distribution should increase. A prior with a greater expected value permits shorter tests, or if the tests remain the same length this prior will, in general, reduce the risk of misclassification. However, the saving from either of the latter, or some combination thereof has to be balanced against the cost of additional instruction.

Novick and Lewis make three summary remarks:

(1) In most situations, a level of functioning of something less than .85 is satisfactory. A value as low as .75 would be highly desirable. This could be accomplished by redefining the task domain slightly so as to eliminate very easy items.

(2) [Instruction] should be carefully monitored so that expected group performance will be just slightly higher than the specified criterion level. This will keep [instruction] time and testing time relatively short.

(3) The program should be structured so that very high loss ratios are not appropriate. That is, individual modules should not be overly dependent on preceding ones.

As Novick and Lewis suggest, it remains to be determined whether these three concerns can be adequately handled within the context of objectives-based programs. To the extent that they can, the Novick-Lewis results should be quite useful. Although it may be obvious, it is perhaps worthwhile to mention also that strictly speaking, the test length recommendations in Novick and Lewis (1974) are applicable only if the Beta-binomial model is to be used in decision making. We just don't know how optimal the recommendations derived from the model

are for the other Bayesian models reported in the literature (Novick, et al. 1973; Lewis, et al. 1973, 1974).

Fahner (1974) has proposed a procedure that is similar to that proposed by Millman but which avoids the formal difficulty of estimating the value of an examinee's domain score prior to obtaining any data. Fahner's approach is a modification of the procedure employed in significance-testing. The basic procedure is to determine a critical score c and the test-length n_o such that

$$\text{Prob}[Y_{ga} > c \mid \pi] \leq \alpha \text{ for all } \pi \leq \pi_o$$

and

$$\text{Prob}[Y_{ga} \leq c \mid \pi] \leq \beta \text{ for all } \pi > \pi_o ,$$

where α and β are the largest acceptable risk levels and Y_{ga} is the observed domain score of examinee a on test g. Since it is not possible to keep both α and β at acceptable levels when the number of items in the test is less than that in the domain, Fahner suggests specifying two values, π_1 and π_2, such that the errors in deciding $\pi > \pi_o$ when in fact $\pi_1 < \pi \leq \pi_o$, and $\pi \leq \pi_o$ when in fact $\pi_o < \pi < \pi_2$, are not very serious. The interval $[\pi_1, \pi_2]$ is thus an indifference region. Once π_1 and π_2 are specified, the normal approximation to the binomial distribution can be used to determine c and n_o, the length of the test.

A difficulty which is shared by the Millman, Novick-Lewis, and the Fahner approaches is the choice to work with the binomial model. We use performance on a random sample of items to generalize to performance on a domain of items. In studying the adequacy of the generalization we may concern ourselves with the results that might

have occurred using different random samples of items. In this context the binomial error model is justified. However, if we concern ourselves with the results that might have occurred on a different administration of the same test, the compound binomial model is more appropriate. Which kind of alternative results should we consider? We feel there is merit in studying the results that might have occurred on different administrations of the same test, since this is the only test on which decisions are actually made. There are two important implications of the choice of a model for measurement error. First, the errors of measurement derived from the compound binomial model are somewhat smaller than with the binomial model so that the recommendations based on the Beta-binomial may be quite conservative. (This is especially true when one recalls that Novick and Lewis (1974), in the interest of making uniform test length recommendations over a class of priors, have already provided conservative recommendations.) Second, the possible bias of the observed score as an estimate of the domain score and the effect of that bias on the likelihood function for the observed score has been ignored.

An important problem related to test length, but which has not been examined in the literature on criterion-referenced testing is the problem of allocating the total time available for testing to the various tests that are to be administered in the instructional program.

Determination of Cut-off Scores

The problem of determining cut-off scores is an extremely important problem for criterion-referenced testing although it has received only limited attention from researchers. Perhaps the most important ramification of the choice of cut-off scores is the psychological effect it has on students. In addition, changes in the cut-off score affects the "reliability"

and the "validity" of the test scores.

Millman (1973) considers five factors in the setting of cut-off scores: Performance of others, item content, educational consequences, psychological and financial costs, errors due to guessing and item sampling.

With respect to "performance of others," Millman (1973) discusses two possible procedures. The first is to set the cut-off score so that a predetermined percentage of the students "pass." However, this procedure is inconsistent with the philosophy of objectives-based programs and therefore it would not seem to be applicable. A second procedure is to identify a group of students who have already "mastered" the material. This group is administered the test and the cut-off score is chosen as the raw score corresponding to a chosen percentile score. Again, the applicability of this procedure to most objectives-based programs seems dubious, but there may be some situations in which the procedure is reasonable.

The second factor is "item content." This approach requires the instructional designer to inspect the items and to determine the subjective probability that some sub-population of the students would get some sub-population of the items correct. (This includes the possibility of deciding that all students should get a particular item correct.) Passing scores are then determined by either a conjunctive or compensatory model. In the conjunctive model, multiple cut-off scores are determined as expected scores within each item group, while for the compensatory model a single cut off score is determined as the expected value over all items.

This approach does have some relevancy in objectives-based programs.

The schemes involved under the heading "educational consequences" involve determining the cut-off score that maximizes independent learning criteria. Millman suggests, amongst other things, the guideline that higher cut-off scores are required for fundamental or prerequisite skills. He also argues that skills that are not prerequisite should not have cut-off scores.

Consideration of psychological and financial costs leads to the suggestion that a low cut-off score be set when remediation costs are high. In situations with lower remediation costs or higher costs for false advancements, higher cut-off scores can be considered. The Bayesian approach considers a fixed threshold score and varies the advancement score to contend with loss ratios, while Millman's approach leads to changing the threshold score itself.

The last factor considered by Millman concerns error due to guessing and item sampling. He tentatively suggests a correction for guessing to contend with the guessing source of error. The error introduced by item sampling is a bias due to systematically disregarding some of the types of questions and content in the domain. Reasons for leaving such items out of the test may be difficulty of construction, inconvenience of administration, or simply ignorance of the extent of the domain. Millman reasonably suggests adjusting the cut-off score for the bias, although he does not treat the question of determining the bias. He also does not explicitly consider the possibility of getting a poor sample of items by random sampling.

An empirical approach to the problem of studying the effects of cut-off scores was completed by Block (1972). He completed an interesting

study which was motivated in part by Bormuth's (1971) contention that rational techniques of determining cut-off scores, that can be defended logically and empirically, must be developed and in part by Cahen's (1970) suggestion that one way the assessment of learning outcomes for an instructional segment can be accomplished is by examining how well the segment has prepared students for future learning.

The learning materials in the experiment were three units of programmed text material on matrix algebra topics appropriate for eighth grade students. Five experimental groups differed with regard to the mastery cut-off score set for the groups. The cut-off scores were .65, .75, .85, and .95. In a particular experimental group all students were required to surpass the cut-off score. This was accomplished by self-directed review sessions. An additional control group did not have a cut-off score established and was not permitted to review.

Block (1972) studied the degree to which varying cut-off scores during segments of instruction influence end of learning criteria. Six criterion variables were selected for study: Achievement, time needed to learn, transfer, retention, interest, and attitude. The results are rather interesting but somewhat limited in generalizability. The results revealed that groups subjected to higher cut-off scores during instruction performed better on the achievement, retention, and transfer tests. On the interest and attitude measures, there was a trend for interests and attitudes to increase until the .85 group and then to level off (it should be noted that the .75 group fared very poorly on the transfer, interest and attitude measures, suggesting some extra-experimental influence). Therefore, the results suggest that different cut-off scores

may be necessary to achieve different outcome measures.

Tailored Testing Research

The considerable amount of testing required to successfully implement objectives-based programs has been criticized, but to some extent this amount of testing can be justified on the grounds that testing is an integral part of the instructional process. Nevertheless, research is needed on procedures that offer the potential for reducing time but which do not result in any appreciable loss in the quality of decision-making from test results. Earlier in the monograph we discussed the use of Bayesian statistical methods as a basis for improving estimation and decision-making. When it is possible to arrange the objectives of an objectives-based instructional program into learning hierarchies (White, 1973, 1974) another promising procedure is that of tailored testing (Ferguson, 1969; Lord, 1970; Nitko, 1974).

Tailored testing has been defined as a strategy for testing in which the sequence and number of test items a student receives are dependent on his performance on earlier items. In testing objectives organized into a learning hierarchy, one can make inferences about student mastery of objectives in the hierarchy which have not been tested. If, for example, a student is tested and found to have proficiency in a specified objective, all objectives prerequisite to it can also be considered mastered. If the examinee lacks proficiency in an objective it can be inferred that all objectives to which it is a prerequisite are also unmastered.

Work on tailored testing has only recently attracted the attention of educational researchers. While there were several studies in the 1950's and early 1960's, Frederic Lord's recent work in improving the precision of measuring an examinee's ability while decreasing the amount of testing time (Lord, 1970, 1971 a, b, c) has done much to bring attention to tailored testing. Recently, Wood (1973) provided a comprehensive review of this line of research.

Ferguson's work in 1969 typifies a second line of research on tailored testing. It is an adaptation of tailored testing to situations in which the testing problem is one of classifying individuals into mastery states rather than precisely estimating their ability. It is this second line of research that has direct application to testing problems in objectives-based programs. Ferguson (1969, 1971) was concerned with classifying students with respect to mastery or non-mastery at each level of proficiency on the learning hierarchy. To accomplish this, computer-based tailored testing was applied to a hierarchy of skills in an objectives-based curriculum. The routing strategy that Ferguson used was complex and required a computer to perform the actual routing. What he found was a 60% savings in time in the computerized administration using a variety of branched test models. A study of the consistency of classifying students with respect to mastery or non-mastery of specific objectives revealed that consistency of mastery decisions was higher when the decisions were made using tailored testing strategies than with a conventional testing procedure. The validity of the tailored testing approach was also found to be high.

In a recent study, Spineti and Hambleton (in press) investigated the interactive effects of several factors on the quality of decision-making and on the amount of testing time in a tailored testing situation. To enable the study of a large number of tailored testing strategies in different testing situations, computer simulation techniques were employed. Factors selected for study because they were considered to be important in the overall effectiveness of a tailored testing strategy included test length, cutting score, and starting point. (Test length is defined as the number of items administered to a student to assess mastery of an objective; cutting score is defined as the point on the mastery score scale used to separate students into mastery and non-mastery states; and starting point is the place in the learning hierarchy where testing is initiated.) Various values of each factor were combined to generate a multitude of tailored testing strategies for study with two learning hierarchies and three different distributions of true mastery scores across the hierarchies. (Of the many learning hierarchies that are available in the educational literature, the learning structures for hydrolysis of salts (Gagné, 1965) and addition-subtraction (Ferguson, 1969) were selected. The two learning hierarchies are shown in Figures 1 and 2.) The criteria chosen to evaluate the effectiveness of each tailored testing strategy were the accuracy of classification decisions relating to mastery, and the amount of testing time.

The simulation results indicated that it is possible to obtain a reduction of more than 50% in testing time without any loss in decision-making accuracy, when compared to a conventional testing procedure, by

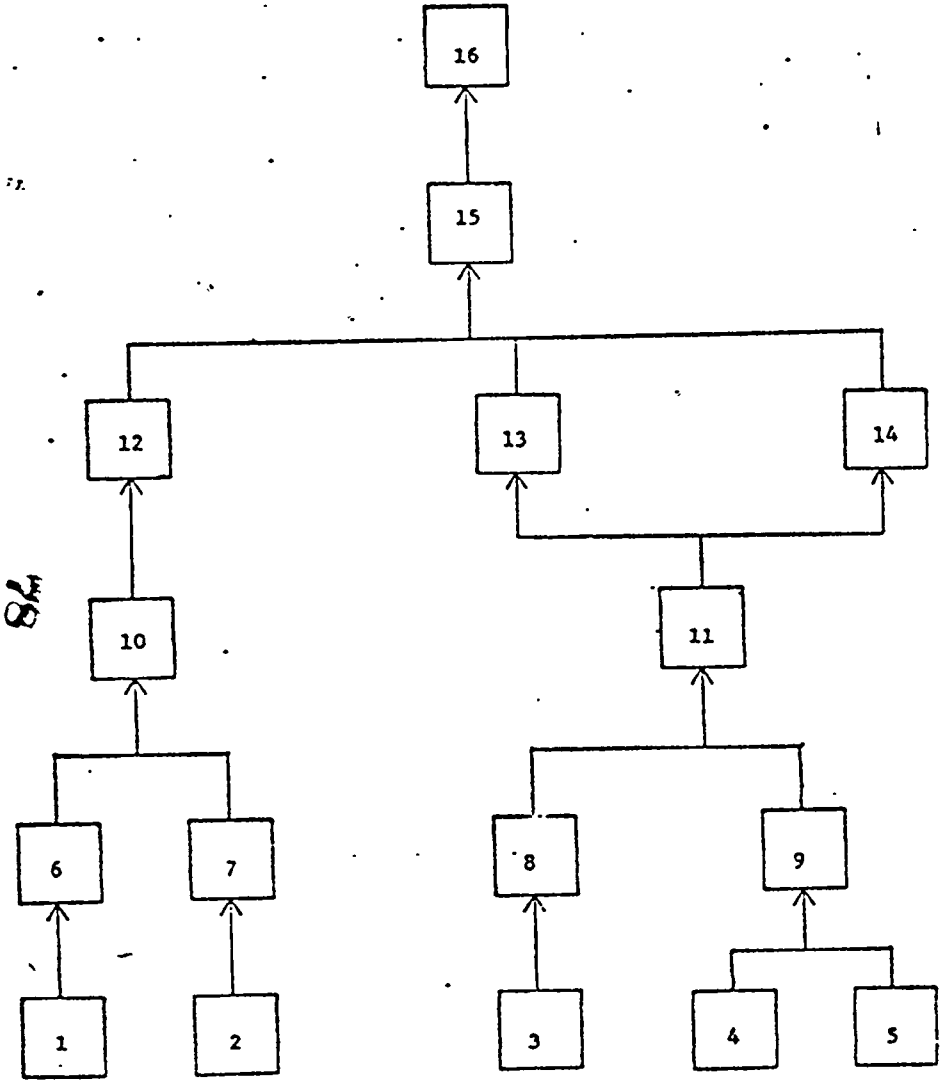

Figure 1. Gagné's Hydrolysis of Salts Hierarchy

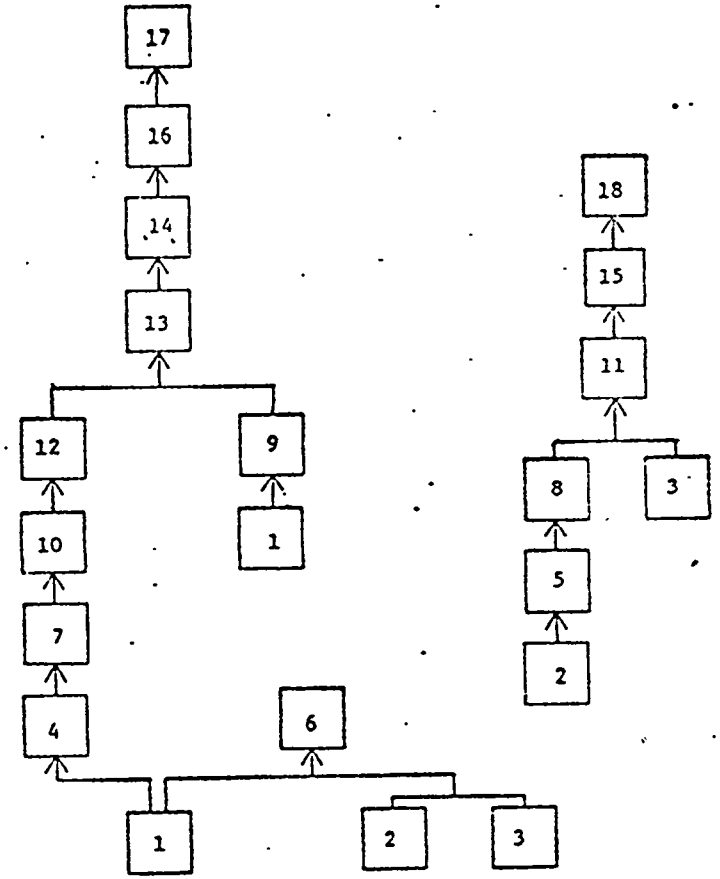

Figure 2. Ferguson's Addition-Subtraction Hierarchy

implementing a tailored testing strategy. In addition, the study of starting points revealed that it was generally best to begin testing in the middle of a learning hierarchy regardless of the ability distribution of examinees across the learning hierarchy. In summary, it was dramatically clear from the numerous simulations, that there was considerable saving in testing time gained through implementing a tailored testing strategy. And, whereas the Ferguson tailored testing strategies could only be implemented with the aid of computer testing terminals, the Spineti-Hambleton tailored testing strategies were simple enough that they could be implemented in the regular classroom with the aid of a "programmed instruction type" booklet.

Among the problems that remain to be resolved in the area of tailored testing research, two seem particularly important. The first involves an extension of the Ferguson and Spineti-Hambleton work. Of most importance we see a need for further study of routing methods and stopping rules. The Spineti-Hambleton study made use of only the simplest routing methods and stopping rules, therefore there is substantial area (and need) for extensions. In addition, it would likely be useful to consider test models in the simulation of test data that incorporate a guessing factor since it is well-known that guessing plays a part in individual test performance.

A second line of research would involve some empirical research on tailored testing in the schools. The design of such study would involve developing a programmed instruction booklet which would include test items designed to measure specific objectives in a learning hierarchy, a self-scoring device, and routing directions. Among the factors that could be investigated in an empirical study are test length, mastery

cut-off score, and routing method. In addition, it would be interesting to study the merits, in terms of overall testing efficiency, of having individuals generate their own starting points for testing in the learning hierarchy.

Description of a Typical Objectives-Based Program

Introduction

As mentioned earlier in the monograph, the trend toward individualization of instruction in elementary and secondary education has resulted in the development of a diverse collection of attractive alternative models (Gibbons, 1970; Gronlund, 1974; Heathers, 1972), many which are objectives-based. According to their supporters, these models offer new approaches to student learning than can provide almost all students with rewarding school experiences. All of these models, as well as many others, represent significant steps forward in improving learning by individualizing instruction. They strive to involve the student actively in the learning process; they allow students in the same class to be at different points in the curriculum; and they permit the teacher to give more individual attention.

To give the reader a flavor for the scope of criterion-referenced testing within an objectives-based program we have included a detailed review of the testing and decision-making procedures within the Individually Prescribed Instruction Program (Glaser, 1968).

The Learning Research and Development Center (LRDC) at the University of Pittsburgh initiated the Individually Prescribed Instruction Project during the early 1960's at the Oakleaf School, in cooperation with the Baldwin-Whitehall Public School District never Pittsburgh. Major contributors to the project over the years have included Robert Glaser, John Bolvin, C. M. Lindvall, and Richard Cox. As of 1974, the IPI program has been adopted by over 250 schools around the country.

Instructional Paradigm

It is instructive, first of all, to describe the structure of the mathematics curriculum. Cooley and Glaser (1969) report that the mathematics curriculum consists of 430 specified instructional objectives. These objectives are grouped into 83 units. (In the 1972 version of the program, there were 359 objectives organized into 71 units.) Each unit is an instructional entity, which the student works through at any one time. There are 5 objectives per unit, on the average, the range being 1 to 14. A collection of units covering different subject areas in mathematics comprises a level; the levels may be thought of as roughly comparable to school grades. For illustrative purposes, we have presented in Table 5 the number of objectives for each unit in the IPI mathematics curriculum.

The teacher is faced with the problem of locating for each student that point in the curriculum where he can most profitably begin instruction. Also, the teacher is responsible for the continuous diagnosis of student mastery as the student proceeds through his program of study.

At the beginning of each school year, the teacher places the student within the curriculum; that is, the teacher identifies the units in each content area for which instruction is required. After completing the gross placement, a single unit is selected as the starting point for instruction, and a diagnostic instrument is administered to assess the student's competencies on objectives within the unit. The outcome of the unit test is information appropriate for prescribing instruction on

TABLE 5

Number of Objectives for Each Unit in the IPI Mathematics Curriculum[1]

Content Area	*Levels*							
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Numeration	12	10	8	8	8	3	8	4
Place Value		3	5	10	7	5	2	1
Addition	3	10	5	8	6	2	3	2
Subtraction			4	6	3	1	3	1
Multiplication				8	11	10	6	3
Division				7	7	9	5	5
Combination of Processes			6	5	7	4	5	6
Fractions	3	2	4	6	6	14	5	2
Money		4	4	6	4	1		
Time		3	2	7	9	5	3	1
Systems of Measurement		4	3	5	7	3	2	
Geometry		2	2	3	9	10	7	9
Special Topics			1	3	3	5	4	5

[1] Reproduced by permission from Lindvall, Cox, and Bolvin (1970)

each objective in the unit. In addition, it is also necessary to select the particular set of resources for the student. In theory, resources that match the individual's "learning style" are selected. Within each unit, there are short tests to monitor the student's progress. Finally, upon completion of initial instruction in each unit, assessment and diagnostic testing takes place. In the next section, the tests and the mechanisms for making these decisions are reviewed.

Testing Model Description

Various research reports over the last couple of years have dealt with the testing model and its development (Cox & Boston, 1967; Glaser & Nitko, 1971; Lindvall et al., 1970). A flow chart of the testing model is presented in Figure 3. To monitor a student through the program the following criterion-referenced tests are used: Placement tests, unit pretests, unit posttests, and curriculum-embedded tests. All of the tests are criterion-referenced, with student performance on the tests compared to performance standards for the purpose of decision-making.

Let us now consider in detail the four kinds of tests and the method for student diagnosis.

Placements Tests When a new student enters the program, it is necessary to place the student at the appropriate level of instruction in each of the content areas. (Glaser and Nitko (1971) called this stage-one placement testing.) Typically, this is done by administering a placement test that covers all of the subject areas at a particular level (see Table 5). Factors affecting the selection of a level for

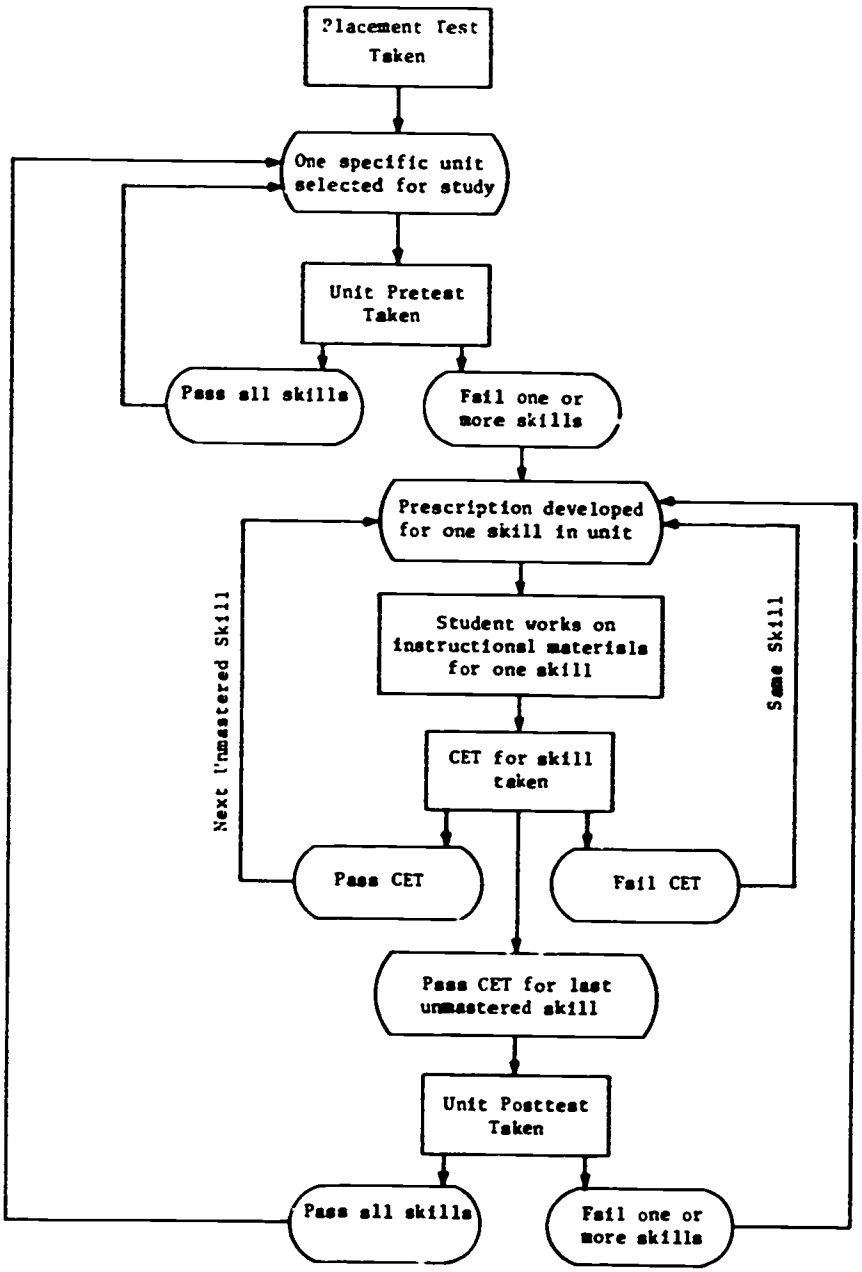

Figure 3. Flowchart of steps in monitoring student progress in the IPI program. (Reproduced, by permission, from Lindvall and Cox, 1969.)

placement testing of a student include student age, past performance, and teacher judgment. Generally, the placement test covers the most difficult or most characteristic objectives within each area. Placement tests are administered until a unit profile identifying a student's competencies within each area is complete. At present, the somewhat arbitrary 80-85% proficiency level is used for most tests in the IPI system.

Student test scores on items measuring objectives in each unit and area in the placement test are used to develop a program of study. The standard procedure is to assign a student to instruction on units in which placement test performance on items measuring a few representative objectives in the units is between 20% and 80%. If the score is less than 20% for a given unit, the unit test in the area at the next lowest level is administered and the same criterion is applied. In the case where a student has a score of 80% or over, testing the unit in the area at the next highest level is initiated. (Further information is provided by Lindvall and Cox, 1970; Weisgerber, 1971; and Cox and Boston, 1967.)

In summary, we note that the placement test has the following characteristics: It provides a gross level of achievement for any student in the curriculum, and it provides information for proper placement of students in the curriculum.

Unit Pretests and Posttests. Having received an initial prescription of units, a student proceeds next to take a pretest for a unit at the lowest level of mastery in his profile. (Glaser and Nitko (1971) call this stage-two placement texting.)

A student is prescribed instruction in each objective in the unit for which he fails to achieve an 85% mastery level on the pretest. A mastery score on each objective for a student is calculated as the percentage of items on the test measuring the objective that the student answers correctly. In the case where the student demonstrates mastery of each objective, he is moved on to the next unit in his profile, where he again takes a pretest.

The unit posttests are simply alternate forms of the unit pretests and are administered to students as they complete instruction on the unit. A student receives a mastery score for each objective in the unit. He is required to repeat instruction on any objective where he fails to achieve an 85% mastery score. The student is directed to the next unit in his profile if he demonstrates mastery on each objective covered in the unit posttest. The next unit prescribed is almost always one at the lowest level of mastery (or grade level). Those who repeat instruction on one or more of the objectives must take the unit posttest again before moving on in their program.

Let us briefly consider the losses involved in making different decisions on the basis of unit testing data. It should be recalled that the unit tests are used to measure student performance on each objective or skill included in the unit with several test items. A student who is mistakenly assigned to a mastery state on an objective covered in the pretest will not likely have the same error in assignment based on the posttest, and so, on the basis of his posttest performance, the student will be assigned instruction on the objective. However, to the extent that the objective is a prerequisite to other objectives in the student's program of study on the unit, he is going

to have some instructional problems. Perhaps this is one place where Bayesian statistical procedures might be useful. They could be used to produce an "improved" profile of test scores across the objectives measured by the unit pretest. Essentially, test performance on an objective that was not consistent with the performance on other objectives in the unit could be modified somewhat. On the average, better mastery-type decisions would result. Likewise, this strategy could be used on the unit posttests.

As far as assigning a student to instruction on objectives he has already mastered, it should be noted that this is likely to be frustrating to the student; however, the majority of false-negative errors occur because students are close to the cutting score.

False-positive errors on the posttest are important if the objectives on which errors are made are prerequisites to other objectives in future units. It should be added that false-positive errors seem to be less serious if they are made on objectives that are terminal objectives (i.e., an objective is terminal if it is not a prerequisite to any other objective in the program). As compared to false-positive errors, false-negative errors are correspondingly less serious because the student can quickly move through the remedial materials and retake the posttest.

In summary, pretests and posttests are available for each unit of instruction. The proper pretest is administered on the basis of a student's curriculum profile, and learning tasks for each objective (or skill, as it is called in the IPI program) within the unit are assigned (or not assigned) on the basis of a student's performance on items measuring the objective.

Curriculum-Embedded Tests. As the student proceeds through a unit of instruction, his progress is monitored. This is done by the use of curriculum-embedded tests (CET). As used in the mathematics IPI program, a CET is primarily a measure of performance on one specific objective. There are usually several test items to measure the objective. A review of the CETs in Level E of the program revealed that there are, on the average, about three items measuring the primary objective covered in the CET. The range is from two to five items. If a student receives a score of 85%, he is permitted to move on to the next presecribed objective. Otherwise, the student is sent back for additional work before taking an alternate form of the CET.

A second purpose of the CET is to assess, albeit in a fairly crude way, whether or not the student has mastered the next objective in the specified sequence for studying the objectives covered in the unit. If the second objective included in the CET is not one the student has been assigned to study, he is moved on to be pretested on the second half of a CET that covers the next objective in the student's program of study. Regardless of which CET a student takes, if he scores above 85% on the items tested, instruction on the objective is not required. Essentially, this means that a student must score 100% since there are normally only about two items included in the test to cover the second objective. This additional pretesting of an objective in the CET gives students a chance to demonstrate mastery of new skills not specifically covered in the instruction up to that point and to eliminate that instruction from his program.

Summary and Suggestions for Further Research

The successful implementation of objectives-based programs depends, in part, upon the availability of appropriate procedures for developing and utilizing criterion-referenced tests for monitoring student progress. The organization and discussion of the available literature on topics such as the uses of criterion-referenced tests, test development, statistical issues in criterion-referenced measurement, validity, reliability, and tailored testing, provided in the monograph, should facilitate the continued development and improvement of criterion-referenced testing in the field. Remaining to be resolved, however, are many technical and practical issues. Let us consider the technical issues first.

First, we are quite enthusiastic about the contributions of Bayesian methods for improving estimation of domain scores and allocation of examinees to mastery states problems, and there is a growing number of impressive results to support our enthusiasm (for example, Novick and Jackson, 1974; Novick and Lewis, 1974). However, we still have some concerns about the overall gains that might accrue in view of the complexity of the procedures, the robustness of the Bayesian models in testing situations where the underlying assumptions of the model are not met (for example, when one has very short tests), and the sensitivity of the Bayesian models to the specification of priors. We note that several of these concerns have been addressed, in part, by Lewis, Wang, and Novick (1974) and we are aware of other studies in progress that also address our concerns.

A second problem, which has not been studied at all in the context of criterion-referenced testing, is an instance of the bandwidth-fidelity dilemma (Cronbach & Gleser, 1965). With a variety of decisions of varying importance to be made in an individualized instructional program and with a limited amount of testing time available, how does one go about determining the "best" distribution of testing time? Does one try to collect considerable test data to make the few most important decisions, or does one try to distribute the available testing time in such a way as to collect a little information relative to each decision? A solution to this important problem is required for an efficient testing program. Determination of test lengths for each domain without regard for the size and scope of the total testing program could produce a serious imbalance between testing and instructional time. Hambleton and Swaminathan (in progress) are studying the problem of distributing testing time across a wide variety of tests (where the tests vary in reliability, validity, and importance to the testing program). The main problem that arises is that it is difficult to obtain a suitable criterion to reflect the "effectiveness" of the testing program.

Third, within objectives-based instructional programs where the objectives can be arranged into learning hierarchies, the strategy of branched testing would seem to offer considerable potential for decreasing the amount of testing while improving its quality. Some of the practical problems have been resolved in the Pittsburgh IPI Program so that the technique can now be used on a limited basis.

Nevertheless, many problems remain before adoption should or can proceed within other programs. For example, it would be necessary to develop a nonautomated modified version of branched testing for schools without computers. Also, we need to know much more than we know now about setting starting places, step sizes, stopping rules, etc., before we can effectively use branched testing in an instructional setting.

Finally, there are many uses for criterion-referenced tests besides the two studied in our monograph. And so it remains to provide a similar review and integration of technical contributions for these uses. For example, the use of criterion-referenced tests in program evaluation will most likely involve methods of item selection and test design different from those mentioned in this monograph. It appears that the methods of matrix sampling could be employed very effectively for item selection in the context of program evaluation.

It seems clear at this point in time that we have sufficient theory and practical guidelines to implement a highly efficient criterion-referenced testing program within the context of objectives-based programs. However, to date, no one has come close to implementing such a testing program. Among the questions that stand in the way of the successful implementation of such a testing program are the following: What skills do classroom teachers need to have in order to implement a criterion-referenced testing program with all of the special refinements (e.g., Bayesian methods, tailored testing, etc.) and how should we train them? Will it be possible to develop domain spe-

cifications in content areas besides mathematics? Even in the area of mathematics where most of the important work has been done (see for example, Hively, et al. 1973) there have been questions raised about the extent to which the notion of domain specifications and subsequent test development can be extended to the more complex mathematics objectives. Another question has to do with whether or not the details of the Bayesian decision-theoretic procedure for allocating examinees to mastery states can be put in a form that teachers will understand and be able to implement. For example, can we train teachers to specify their prior beliefs about abilities of examinees and losses associated with misclassification errors? Prior information for a Bayesian solution might include the student's past performance in the program, scores on other objectives included in the test, the overall performance of the group of students, etc. It is critical that such details be completely checked out for their appropriateness and presented in a clear form to the teachers.

References

Airasian, P. W., & Madaus, G. F. Criterion-referenced testing in the classroom. Measurement in Education, 1972, 3, 1-8.

Alkin, M. C. "Criterion-referenced measurement" and other such terms. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-16.

Baker, F. B. Computer-based instructional management systems: A first look. Review of Educational Research, 1971, 41, 51-70.

Block J. H. Criterion-referenced measurements: Potential. chool Review, 1971, 69, 289-298.

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

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bormuth, J. R. Development of standards of readability: Toward a rational criterion of passage performance. Final Report, USDHEW, Project No. 9-0237. Chicago: The University of Chicago, 1971.

Brennan, R. L. The evaluation of mastery test items. U. S. Office of Education, Project No. 2B118, 1974.

Brennan, R. L., & Stolurow, L. M. An empirical decision process for formative evaluation. Research Memorandum No. 4. Harvard CAI Laboratory, Cambridge, Mass., 1971.

Cahen, L. Comments on Professor Messick's paper. In M. C. Wittrock, & D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart and Winston, 1970.

Carver, R. P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. Psychological Bulletin, 1968, 70, 213-220.

Cooley, W. W., & Glaser, R. The computer and individualized instruction. Science, 1969, 166, 574-582.

Coulson, D. B., & Hambleton, R. K. On the validation of criterion-referenced tests designed to measure individual mastery. Paper presented at the annual meeting of the American Psychological Association, New Orleans, 1974.

Cox, R. C., & Boston, M. E. Diagnosis of pupil achievement in the Individually Prescribed Instruction Project. Working Paper 15. Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1967.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Crehan, K. D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11, 225-262.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions. (2nd Ed.) Urbana, Ill.: University of Illinois Press, 1965.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, 1972.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. The British Journal of Statistical Psychology, 1973, 16, 137-163.

DeVault, M. V., Kriewall, T. E., Buchanan, A. E., & Quilling, M. R. Teacher's manual: Computer management for individualized instruction in mathematics and reading. Madison, Wisconsin: Research and Development Center for Cognitive Learning, University of Wisconsin, 1969.

Donlon, T. F. Some needs for clearer terminology in criterion-referenced testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 3, 11-17.

Ebel, R. L. Criterion-referenced measurements: Limitations School Review, 1971, 69, 282-288.

Ebel, R. L. Evaluation and educational objectives. Journal of Educational Measurement, 1973, 10, 273-279.

Ferguson, R. L. The development, implementation and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Fhaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Flanagan, J. C. Functional education for the seventies. Phi Delta Kappan, 1967, 49, 27-32.

Flanagan, J. C. Program for learning in accordance with needs. Psychology in the schools, 1969, 6, 133-136.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., & Neyman, C. A., Jr. The American high school student. Cooperative Research Project No. 635, U. S. Office of Education. Pittsburgh: American Institutes for Research and University of Pittsburgh, 1964.

Fleiss, J. L., Cohen J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 72, 323-327.

Fremer, J. Handbook for conducting task analyses and developing criterion-referenced tests of language skills. PR 74-12. Princeton, New Jersey: Educational Testing Service, 1974.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1965.

Gibbons, M. What is individualized instruction? Interchange, 1970, 1, 28-52.

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

Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Glaser, R. Evaluation of instruction and changing educational models. In M. C. Wittrock, & D. E. Wiley (Eds.), The evaluation of instruction. New York: Holt, Rinehart and Winston, 1970.

Glaser R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Goodman, . A., & Kruskal, W. H. Measures of association for cross classification. American Statistical Association Journal, 1954, 49, 732-764.

Gronlund, N. E. Individualizing classroom instruction. New York: Macmillan Publishing Co., 1974.

Guttman, L., & Schlesinger, I. M. Development of diagnostic, analytical and mechanical ability tests through facet design and analysis. U.S. Office of Health, Education and Welfare, Project No. OE-15-1-64, 1966.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

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

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

Harris, C. W. Problems of objectives-based measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the study of evaluation, University of California, 1974. (a)

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974. (b)

Harris, C. W., Alkin, M. C., & Popham, W. J., Problems in criterion-referenced measurement. CSE monograph series in evaluation. No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Harris, M. L., & Stewart, D. M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, 1971.

Heathers, G. Overview of innovations in organization for learning. Interchange, 1972, 3, 47-68.

Henrysson, S., & Wedman, I. Some problems in construction and evaluation of criterion-referenced tests. Scandinavian Journal of Educational Research, 1974, 18, 1-12.

Hieronymous, A. N. Today's testing: What do we know how to do? In Proceedings of the 1971 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1972.

Hively, E., Maxwell, G., Rabehl, G., Senison, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. CSE mongraph series in evaluation. No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

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

Hsu, T. C., & Carlson, M. Oakleaf School Project: Computer-assisted achievement testing (A Research Proposal.) Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1972.

Ivens, S. H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Jackson, P. H. Simple approximations in the estimation of many parameters. British Journal of Mathematical and Statistical Psychology, 1972, 25, 213-229.

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

Kriewall, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

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

Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. (Submitted for publication, 1974)

Light, R. J. Issues in the analysis of qualitative data. In R. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Lindvall, C. M., & Cox, R. The role of evaluation in programs for individualized instruction. In R. W. Tyler (Ed.), Educational evaluation: New roles, new means. Sixty-eighth Yearbook, Part II. Chicago: National Society for the Study of Education, 1969.

Lindvall, C. M., Cox, R. C., & Bolvin, J. O. Evaluation as a tool in curriculum development: The IPI evaluation program. AERA monograph series on curriculum evaluation, No. 5. Chicago: Rand McNally, 1970.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests". Journal of Educational Measurement, 1972, 9, 31. (b)

Livingston, S. A. Reply to Shavelson, Block and Ravitch's "Criterion-referenced testing: Comments on reliability." Journal of Educational Measurement, 1972, 1, 139-140. (c)

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row, 1970.

Lord, F. M. Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-21. (a)

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (b)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813. (c)

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lu, K. H. A measure of agreement among subjective judgments. Educational and Psychological Measurement, 1971, 31, 75-84.

Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain-referenced testing. Educational and Psychological Measurement, 1973, 33, 351-360.

Maxwell, A. E., & Pilliner, A. E. G. Deriving coefficients and agreement for ratings. British Journal of Mathematical and Statistical Psychology, 1968, 21, 105-116.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. Research Bulletin 74-77. Princeton, N. J.: Educational Testing Service, 1974.

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Millman, J. Determining test length: Passing scores and test lengths for objectives-based tests. Instructional objectives exchange, Los Angeles, California, 1972.

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

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

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

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

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Popham, W. J. (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Popham, W. J. Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

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

Rao, C. R. Linear statistical inference and its applications. New York: Wiley, 1965.

Rovinelli, R., & Hambleton, R. K. Some procedures for the validation of criterion-referenced test items. Final Report. Albany, N.Y.: Bureau of School and Cultural Research, New York State Education Department, 1973.

Shavelson, R. J., Block, J. H., & Ravitch, M. M. Criterion referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly and interpretation. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Spineti, J. P., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objectives-based instructional programs. Educational and Psychological Measurement, in press.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, in press.

Traub, R. E. Criterion-referenced measurement: Something old and something new. A paper prepared for an invited public address at the University of Victoria, 1972.

Wang, M. M. Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program, 1973.

Wedman, I. Reliability, validity and discrimination measures for criterion-referenced tests. Educational Reports, Umea, No. 4, 1973.

White, R. T. Research into learning hierarchies. Review of Educational Research, 1973, 43, 361-375.

White, R. T. The validation of a learning hierarchy. American Educational Research Journal, 1974, 11, 121-136.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.