ED 193 271 TM 800 586

AUTHOR	Gifford, Janice A.; Hambleton, Ronald K.
TITLE	Construction and Use of Criterion-Referenced Tests in Program Evaluation Studies. Laboratory of Psychometric and Evaluation Research Report No. 102.
INSTITUTION	Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	14 Mar 80
NOTE	41p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; Group Testing; *Item Sampling; Models; Norm Referenced Tests; *Program Evaluation; Sampling; *Test Construction; Test Interpretation; Test Items; *Test Reliability; Test Validity

ABSTRACT

Technical considerations associated with item selection and reliability assessment are considered in relation to criterion-referenced tests constructed to provide group information. The purpose is to emphasize test building and the evaluation of test scores in program evaluation studies. It is stressed that an evaluator employ a performance or achievement instrument sensitive enough to reflect group/individual ability in terms of the specific goals of the program. Criterion-referenced tests are designed for this purpose. Two steps in criterion referenced test development are considered: the approach to item selection, and the assessment of score reliability. Methods for handling them in preparing and using criterion-referenced tests in program evaluation studies are offered. Norm-referenced and criterion-referenced testing within a program evaluation context are compared, and a model for developing/validating criterion-referenced tests is introduced. (Author/GK)

Construction and Use of Criterion-Referenced Tests in Program Evaluation Studies

Janice A. Gifford and Ronald K. Hambleton
University of Massachusetts, Amherst

Abstract

The number of new educational programs has increased dramatically since the mid-sixties. While some programs are minor extensions of older programs, others represent completely new educational approaches. The importance of comprehensive summative and formative evaluations of these new programs is clear. It is equally clear to many administrators, program developers, and evaluators that criterion-referenced tests are an essential type of instrumentation for conducting program evaluation studies. Unfortunately, nearly all of the recently developed criterion-referenced testing technology applies to test development and uses with individual scores (for example, to monitor student progress, to diagnose student learning needs, and to certify students as high school graduates). In program evaluation, group information is of central importance. It is not the case, as some have assumed, that testing technology developed for use of test scores with individuals is optimal for this purpose. We suggest that there is some misdirection in testing projects due to this basic misunderstanding. Four steps in test development are different: (1) approach to item selection, (2) assessment of reliability, (3) standard-setting methods, and (4) methods of test score reporting. The purposes of the paper will be to consider the first two steps and offer methods for handling them in preparing and using criterion-referenced tests in program evaluation studies. In addition, prior to considering the two steps, a brief comparison of norm-referenced testing and criterion-referenced testing within the context of program evaluation is offered and a model for developing and validating criterion-referenced tests is introduced.

3/14/80

Construction and Use of Criterion-Referenced Tests in Program Evaluation Studies[1,2]

Janice A. Gifford and Ronald K. Hambleton[3]
University of Massachusetts, Amherst

The following questions are often addressed in order to determine the effectiveness and hence the impact of an educational program:

- Are the objectives worthwhile?
- Are the stated objectives being achieved?
- How does one program compare with another in accomplishing a common set of objectives?
- What changes should be made to improve program effectiveness?

The formal, systematic search for answers to these and similar questions is termed program evaluation. During the past ten years, several models of evaluation have emerged (Glass & Ellett, 1980). Since there is no single accepted definition of program evaluation, these models differ in varying degrees, in the set of questions addressed, and in the phases of the program implementation that are examined. Rather than attempt to present, compare, and contrast the major evaluation models here, the reader is referred to any of the several excellent discussions of

[1] The project reported herein was performed persuant to a grant from the National Institute of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

[2] Laboratory of Psychometric and Evaluative Research Report No. 102. Amherst, MA: School of Education, University of Massachusetts, 1979.

[3] We would like to thank H. Swaminathan for his helpful comments on an earlier draft of the paper.

evaluation models (Glass & Ellett, 1980; Perloff, Perloff, & Sussna, 1976; Popham, 1975; Worthen & Sanders, 1973). However, in its most general form, program evaluation may be thought of in terms of three phases. Phase one consists of the examination and evaluation of the goals of a program. That is "Are the stated purposes of the program of value?" Phase two consists of the examination and evaluation of the processes of the program. For example, "Are the processes such that they facilitate the attainment of the stated program goals?" Finally, phase three focuses on the measurement of program outcomes. That is, "Have the stated goals and objectives been achieved?"

In order to answer questions raised at any of the three phases, a program evaluator must begin by drawing on many of the measurement techniques commonly used by social, psychological and educational researchers. For example, in order to study the adequacy of the goals of a program, measurement may take the form of needs assessments, attitude scales or preference scales. In phase two, for examination of the process, questionnaires, interview schedules, and observational instruments may be helpful. Attitude scales, performance tasks and paper and pencil achievement measures are examples of techniques available for the measurement of program outcomes.

Since educational programs, in particular, are generally directed toward goals such as the acquisition of particular knowledge or skills, or the advancement to some desired performance level or achievement level by those individuals served by a program, it is crucial that an evaluator employ a performance or achievement instrument sensitive enough to adequately reflect the ability of a group or of the individuals in terms of the specific goals of the program. Unfortunately, norm-referenced paper and pencil instrument development techniques are less than ideal for constructing tests to measure individual and group accomplishments in relation to a set of program goals. Norm-referenced test development methods, which are well-known are aimed toward producing tests to reliably and validly rank or compare examinees. However, evaluators require test development methods that will permit them to design and to use instrumentation to determine what individuals and groups can and cannot do in relation to a set of program goals. Criterion-referenced test development methods provide the answer since criterion-referenced tests are constructed to permit the interpretation of individual or group test scores in relation to a set of well-defined objectives (Popham, 1978a).

Norm-referenced tests and criterion-referenced tests are designed to achieve different purposes and therefore the approaches to test construction and test score interpretation will also differ. When these two types of tests are used incorrectly, problems arise. For example, Carver (1975) argues convincingly that Coleman et al. (1966) in a well-known and often cited study of the impact of schooling used inappropriate instruments (norm-referenced tests rather than criterion-referenced tests) and therefore the data do not address the important question under study, that of the relationship between school differences and level of achievement.

Following ten years or so of psychometric research, a well-developed technology for building criterion-referenced tests and using the derived test scores exists (e.g., Hambleton & Eignor, 1979a; Popham, 1978a). Unfortunately, this technology is designed to construct tests for use in evaluating the performance of individuals in relation to a set of well-defined goals or competency statements and therefore when group performance is of primary interest, as it is in program evaluation studies, variations from the usual ways for building and using the tests will be necessary. Four steps in test development are different: (1) approach to item selection, (2) assessment of score reliability, (3) standard-setting methods, and (4) methods of test score reporting. The purposes of this paper will be to consider the first two steps and offer methods for handling them in preparing and using criterion-referenced tests in program evaluation studies. In addition, prior to considering the two steps, a brief comparison of norm-referenced testing and criterion-referenced testing within the context of program evaluation is offered and a model for developing and validating criterion-referenced tests is introduced.

Comparison of Criterion-Referenced Tests and Norm-Referenced Tests in the Context of Program Evaluation

The educational program evaluator, in the search for a suitable instrument, will quickly discover that the great majority of instruments are norm-referenced tests. For example, more than 95% of the instruments listed in the Eighth Mental Measurement Yearbook (Buros, 1978) are norm-referenced tests. Although criterion-referenced measures have not been used to the same extent as norm-referenced measures, there is a growing awareness of the importance of criterion-referenced measurement.

Generally, a norm-referenced test cannot be distinguished from a criterion-referenced test by appearance alone. The differences revolve primarily around three areas: specification of test content, the selection of items, and interpretations of the scores. In comparing CRTs to NRTs, it should be kept in mind that the goal of NRTs is to represent "ability" in terms of other individuals, while the goal of CRTsis to represent "ability" in terms of a given domain of content.

The first step in the construction of any test is to specify, in some manner, the content domain to be measured by a test. It is common for developers of both types o. tests to begin with objectives. With criterion-referenced tests, however, it is essential to describe the objectives in considerably more detail. Added clarity can be obtained by offering a sample test item, describing appropriate item content and specifying characteristics and types of answers that can be used as distractors in objective test items. "Expanded objectives" (or "domain specifications") facilitate the preparation of test items to measure objectives and improve the clarity of test score interpretations.

The second phase of test construction involves the development, analysis, and selection of items. With norm-referenced tests, a large set of items is initially constructed to reflect the objectives outlined in step one. Preliminary forms of the test are constructed and administered to examinees similar to those for whom the test is intended. Later, the items are studied in terms of their difficulty and discrimination. Since the major purpose of a norm-referenced test is to compare an individual's performance, knowledge, or skill, to that of some reference group, a suitable norm-referenced test will be constructed with those items that contribute most to maximizing test score variability. Comparisons among examinees are more reliable when test scores are dispersed widely. Hence, the final item selection is dependent not only on the objectives of interest, but also on the statistical characteristics of the available items.

On the other hand, since the universe or domain of items is specifically defined for a criterion-referenced test, item selection, typically, consists of selecting a set of representative items from the domain. If more than one objective is measured by a test, a set of representative items from the domain of items matched to each objective is drawn. Item statistics play a secondary role to item representativeness in criterion-referenced test item selection.

Finally, test scores are reported and used in a way consistent with the test's purpose. A norm referenced test score is reported as a raw score and one or more derived scores (for example, percentile scores, age or grade-equivalent scores, and standard scores). Raw scores alone have very little meaning. Inferences cannot be made as to what the individual knows or does not know. The derived scores give specific information concerning the relation of an individual's knowledge, skill or ability, to that of a particular reference group. The score (or scores) on a criterion-referenced test, however, provides information concerning the relationship of an individual's knowledge, skill or ability to a given specified domain of content.

The intrinsic differences between criterion-referenced and norm-referenced measurement have important implications for their use in the evaluation of educational programs. A major shortcoming of the use of norm-referenced tests in program evaluation results from the discrepancy between the content covered by a test and the content of the program that is being evaluated. The tests that are most commonly used in evaluations are used nationwide and are based on an amalgamation of objectives of programs from all over the country. Each program has different instructional objectives and the instruction of particular objectives may occur at different times. The overlap of instructional objectives and test objectives will not usually be complete and the degree of overlap will change from program to program. This is particularly true in compensatory educational programs, where the objectives may be more basic and specific than the general objectives reflected in norm-referenced tests. Moreover, each curriculum typically depsnds on the people teaching the program and their priorities and emphases. In general, it will be difficult to find a atandardized achievement test where the content closely matches the content goals of a particular program being evaluated. It is not uncommon therefore to hear the charge of "unfairness" when a norm-referenced test is used in program evaluation.

A second source of the discrepancy between test content and program objectives arises directly from a major purpose of norm-referenced tests, i.e., to compare an individual's performance, knowledge or skill to that of some reference group. In order to effectively obtain this type of information from a test, the test must be constructed with that purpose in mind. Consequently, norm-referenced tests consist of test items that contribute most to maximizing test score variability. In the process of choosing items that contribute sufficiently to test variability, those contributing less to variability are eliminated. It is clear that items tapping concepts taught successfully by a great number of teachers will contribute little to test score variability (most students will answer the items correctly) and will be eliminated, while the items measuring pure reasoning ability will have greater variability and will be retained. In other words, many instruction-related skills are systematically eliminated, and the variation that remains is primarily due to the effects of non-instruction related variables. When "easy" and "difficult" items are deleted, resulting tests look less like achievement tests and more like aptitude tests (Popham, 1978b). If an instrument is to be sensitive to the learning process, its content must be carefully matched to that of the program. Since, at present, many programs to be evaluated are innovative, not only are the instructional methods different, but often the goals and objectives of these programs are different from those of the traditional program. As a result, a norm-referenced test score may be inapproriate since it does not indicate knowledge in terms of the instruction. It would often be a mistake to judge a new program according to the standards of a traditional program.

Criterion-referenced tests, however, are constructed or can be selected specifically to match the goals and objectives of a program, and since item quality depends exclusively on the ability of the item to reflect the domain, this match is not lost in the item selection process as it may be in a norm-referenced test. Consequently, criterion-referenced test scores, assuming the test from which the scores are derived is constructed and administered properly, are valid indicators of performance or achievement in relation to the instructional objectives of the program.

Perhaps the greatest advantage of using criterion-referenced measurement in the evaluation of educational programs result from the range and quality of information obtainable from the test scores. Because of the match between the test content and instructional objectives, criterion-referenced scores permit a description of an individual in terms of clearly specified domains of content. For example, it may be said that a student has mastered 60% of a set of program objectives. However, it is not always the case that information is required on each individual or all objectives. Particularly in program evaluation, an evaluator often will want to know how some group of students in general has been affected by an educational program rather than any given individual. Since this is the case, it is possible with criterion-referenced testing to make very efficient use of items. A procedure referred to as "item-examinee sampling" provides for optimal efficiency in information gathering when there are practical limits on the number of items that can be reasonably administered to an individual. This topic will be considered in detail in a later section.

In most evaluations of educational programs it is not only important to know something about the achievement of those served by a program in terms of the prescribed objectives, it is also valuable to be able to compare the performance of individual in a program to the performance of various other groups. Even though criterion-referenced tests are not constructed specifically to maximize variability of test scores and the frequency distribution of the test scores may be homogeneous and hence less useful for ranking individuals, norm-referenced interpretations of criterion-referenced test scores can be made and can be of considerable value. As long as objectives are held in common, comparisons of criterion-referenced test scores among examinees or groups of examinees can be made.

Articles by Ebel (1978), Popham (1978b), and Mehrens and Ebel (1979) provide additional insights into the topics considered in this section.

Steps in Criterion-Referenced Teat Development

In this section the essential atepa in criterion-referenced test development are introduced. A 12 atep model is presented in Figure 1 (Hambleton & Eignor, 1979b). The importance of each step in the model depends upon the size and scope of the test development and validation project. An agency with the responsibility of producing tests for state-wide use will proceed through the steps in a rather different way than will a small consulting firm or a group of researchers.

In brief, the twelve ateps are as followa:

Step 1--Objectives must be prepared or selected before the test development procesa can begin.

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

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

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

Step 5--A systematic assessment of items prepared in steps 3 and 4 is conducted to determine item validities. Essentially, the taak ia to determine the content validity of the test items.

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

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

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

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

1. Preparation and/or Selection of Objectives

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

3. Writing Test Items "Matched" to Objectives

4. Editing Test Items

5. Determining Content Validity of the Test Items

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

6. Additional Editing of Test Items

7. Test Assembly

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

8. Setting Standards for Interpreting Examinee Information

9. Test Administration

10. Collection of Reliability, Validity and Norms Information

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

12. Periodic Collection of Additional Technical Information

Figure 1. Steps for Developing and Validating Criterion-Referenced Test Scores (From Hambleton & Eignor, 1979b).

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

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

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

Hambleton and Eignor (1979a, 1979b) and Popham (1978a) describe in detail how to carry out the 12 steps in constructing tests to describe the performance of individuals. Methods for constructing tests for use in program evaluation studies are not nearly so well-developed. In the next two sections methods will be proposed for handling two of the four steps, item selection and reliability assessment, which are handled differently when building tests to describe the performance of groups.

Approach to Item Selection

Introduction

When decisiona are to be made concerning an entire educational or social program, group information rather than individual information is of primary concern. There are two very important types of group information available when criterion-referenced tests are employed. The firat of these is the average domain score for the entire group on each of the relevant objectives (and across the set of objectives of interest). An examinee's domain acore ia his/her proportion-correct score in the domain of items measuring the objective. An estimate of the average domain score for a group on a particular objective not only gives an excellent description of a group in terms of the specific objective, but can be used to make comparisons over time, comparisons to other groups, comparisons among objectives or comparison of the group'a performance to some desired standard of performance (possibly set by the instructors of the program of study). For example, a target may be set for a group of examinees to achieve an average domain score of .70 on an objective, that is, a 70% average performance level on items measuring the objective. It would be helpful after program implementation to compare the average domain score of the group to the chosen standard or target.

The aecond type of information available through the use of criterion-referenced testing is the percentage of people in a program who are classified as masters on any given objective. An individual is classified as a master or non-master by comparing the individual's domain score estimate to a cut-off score positioned on the domain score scale. It is thus helpful to know what percentage of those taking part in a given program can be classified as masters. Again, comparisons over time, comparisons between groups, comparisons among objectives, or comparisons of the group to some standard are extremely useful in the evaluation of effectiveness of program implementation.

Besides average domain scores and percent of masters on each or perhaps only the most important program objectives, program evaluators usually have an interest also in the variability and distribution of domain scores, and in the percent of examinees in a group mastering a specified number of objectives at a specified level of performance.

It should be noted, however, that in order to gather the types of information described above, each student or sample of students should be tested by several items for each objective. Testing time can quickly become prohibitive. It would not be unusual, for example, to have 100 objectives, each tested with 10 items, resulting in a total of 1000 items, far too many to reasonably administer to any group of people. It is possible, however, to utilize sampling plans in order to gather information more efficiently.

The simplest sampling technique is to choose a random, or stratified random sample of examinees from the examinee population, and administer the entire test to the sample. This is known as examinee-sampling. Although this procedure reduces the total amount of testing, each individual that is selected may still be tested to an unreasonable extent. An improvement on this is another sampling procedure referred to as item-sampling. Here, items are randomly selected (or perhaps stratified on difficulty level) from the domain of items measuring each objective and administered to all examinees. This is actually the situation that occurs in criterion-referenced measurement.

A representative set of items is selected from the domain of items measuring an objective in order to make inferences about the entire domain. Unfortunately, since the number of objectives to be measured by a test is often large, the number of test items measuring any single objective is likely to be quite small and therefore adequate domain coverage of an objective is difficult to ensure.

Fortunately, it is possible for the evaluator to provide an accurate description of the program with respect to the given objectives, while administering only a fraction of the total number of items to any given individual. This procedure consists of the simultaneous application of the two previously mentioned sampling procedures and is called item-examinee or matrix sampling. A randomly selected group of items is administered to a randomly selected group of examinees. A further refinement, which results in better estimates of population parameters, is referred to as multiple matrix sampling. In this case, the item-examinee sampling procedure is repeated a number of times. A first set of randomly selected items is assigned to a first group of randomly selected examinees, followed by the assignment of a second set of items to a second group of examinees and so on. Estimates of parameters of interest are calculated for each matrix and then pooled, resulting in estimates that can be used to make inferences about all examinees on all items. Considerable research has demonstrated the feasibility, desirability and efficiency of matrix and multiple matrix sampling procedures (Shoemaker, 1973a; Sirotnik, 1974).

There are several item-examinee sampling designs that program evaluators may find particularly useful for applications involving criterion-referenced testing. Next, some practical considerations in choosing a design will be discussed, followed by a presentation of specific designs.

Some Preliminary Considerations

There are many practical aspects to be considered in choosing an efficient sampling plan.[1] Total testing time, the number of objectives, the number of items per objective, and the number of examinees must all be considered in light of the desired degrees of precision for the statistics of interest. The amount of time allotted for testing is very often restricted, if not by conditions intrinsic to the program itself, then by the length of time one can expect an examinee to respond to test items. The number of objectives to be tested must also be considered and decisions made as to whether or not it is critical that each and every objective be tested.

In some situations, it is more important to have more reliable information on a subset of objectives rather than less reliable information on all objectives. This may be particularly true when it is of interest to report the percent of examinees who are classified as masters on a given objective. In order to classify an examinee as a master or non-master reliably, several items must be used for a given objective. Since this may result in an unreasonably long test, it may be necessary to establish priorities for the objectives, and measure most completely, only those objectives basic to the purposes of the program. This may be accomplished particularly if the objectives are structured hierarchically, that is, mastery of one objective is a prerequisite to mastery of others. Priorities can be established to reflect this.

[1] A sampling plan describes the number of different tests that will be constructed, the number of items in the tests, and the number of examinees who will be administered each test.

In contrast, it may be more important in some situations to report information on all objectives of a program. Since the number of objectives is often large, obtaining information on each objective and at the same time maintaining a test of reasonable length may not be feasible. This problem can be overcome, through the use of multiple matrix sampling.

In designing a sampling plan, since the number of examinees, items, objectivesand items per objective have a direct bearing on the precision of estimates, the evaluator often must arrive at a compromise, sometimes sacrificing precision in order to arrive at a feasible test plan. Other aspects that need to be taken into account relate to the nature of the objectives and test items of interest. Objectives tested through use of items that require special directions, practice questions, or verbal presentation can have an effect on the development of a sampling plan. In these cases it is an inefficient use of time to sample a very small number of items from the objective. It is perhaps more reasonable to test each examinee group with fewer objectives and more items per objective. The complexity of the domain also has an effect on the number of items selected to measure an objective and the method of item selection. More items must be used with complex domains to insure item representativeness. Stratification of the items may be necessary to insure complete, yet efficient, coverage of the item domain.

Selection of Designs

In this section of the paper a few designs are presented that are particularly suited for use with criterion-referenced testing in program evaluation. The notation and definitions used here will be in keeping with that suggested by Shoemaker and Knapp (1974). The number of items in the domain of items measuring an objective is denoted by K, the number of items measuring an objective in a test by k, the total number of examinees in the population is denoted by N, and the number of examinees taking each test by n. A particular sampling plan for the collection of test data in relation to an objective, then, can be represented as t/k/n where t is the number of tests.

Multiple matrix sampling plans can be with or without replacement on both the item and examinee dimensions. In evaluation settings, it is important, when sampling examinees, to choose a given examinee only once. This ensures maximum coverage of examinees and reduces testing time on the part cf an examinee while avoiding confounding effects due to an examinee taking more than one test. Similarly, sampling of items without replacement is important to ensure domain coverage and avoid overlapping tests. Thus, it is apparent that in evaluation settings, sampling of items and examinees without replacement is the most meaningful and feasible sampling plan to consider. For the purposes of this paper, we shall therefore assume that all sampling is without replacement.

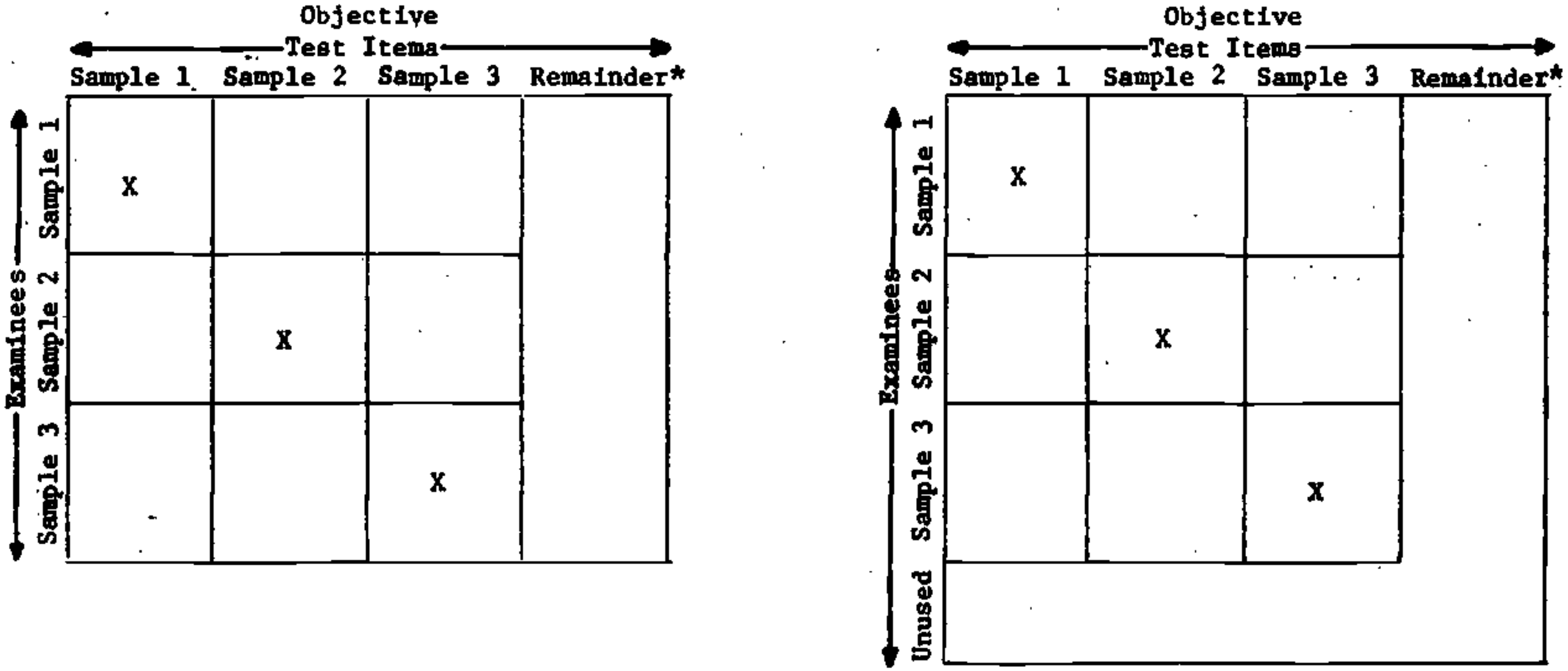
If each of the K items of an item domain is assigned to at least one test the sampling is said to be exhaustive in the item dimension. Likewise, examinee sampling may be referred to as either exhaustive or non-exhaustive depending on whether or not the entire group of N examinees is tested. The choice of sampling is largely dependent on the type of inferences the evaluator wishes to draw from the resulting data. In the particular application of criterion-referenced measurement tothe evaluation of programs, the inference to be made from the item dimension is different from that in a typical item sampling plan. Ordinarily, item sampling is used to estimate a groups' performance on a fixed length teat (K items) by looking at performance on tests of length k. The important point here is that the inference is made to some particular set of K items. In criterion-referenced meaaurement, however, the inference of interest is not to some fixed set of items but to a well-defined but very large domain of test items. Consequently, items are in effect randomly chosen from the well-defined domain and used to estimate examinees' succesa on the domain of interest (Sirotnik, 1974). It is clear, then, that since the inference is to be made to the entire domain from a sample of items, item sampling is non-exhaustive when criterion-referenced interpretationa are to be made of the scores. It should also be noted that, in the evalustion of a program, information about many domains is often required. The multiple matrix sampling must occur within each domain, aince generalizations are to be made to each domain of interest.

The sampling of examinees, can be one of three types; exhaustive, non-exhaustive from a finite population, and non-exhaustive from an infinite population. An example of an exhaustive sampling plan is when every person in a program is tested on some subset of items keyed to an objective. For example, a population of 1000 examinees is divided into four subgroups of 250 examinees each, and each group is administered 5 items randomly choaen from a domain. Each examinee receives 5 items and the information from this is used to make an inference about those 1000 examinees on the entire domain. The second type of examinee sampling comes about when sampling is done from a fixed population of examinees and not all those in the population are tested. For example, suppose there are 100 objectives to be tested on a group of 1000 examinees. The population of 1000 examinees can be divided into two random samples, each sample of examinees responding to items representing one half of the total objectives. Although 500 examinees in each case are tested on an objective or domain, the inference is to be made to the original population of 1000 examinees. Within a given objective, examinee sampling is non-exhaustive. This design is particularly applicable when information on many objectives is collected simultaneously, since each objective is tested on only a sample of the population. An obvious extension of the above is non-exhaustive sampling from an infinite population. This design is appropriate whenever the size of the examinee sample is small in relation to the size of the population. A major advantage of this plan is that it simplifies statistical computation. Schematic representations of several types of sampling plans considered so far are presented in Figure 2.

As mentioned earlier, it is important, when choosing a design to implement, to consider carefully, the nature of the information needed. Several types of information will be addressed next. These are: (1) the mean and variance of domain scores on an objective, (2) the entire domain score distribution on an objective, (3) percent of masters on an objective, and (4) percent of examinees mastering a given percent of objectives.

Figure 2. Representation of several types of sampling plans.

At the Objective Level



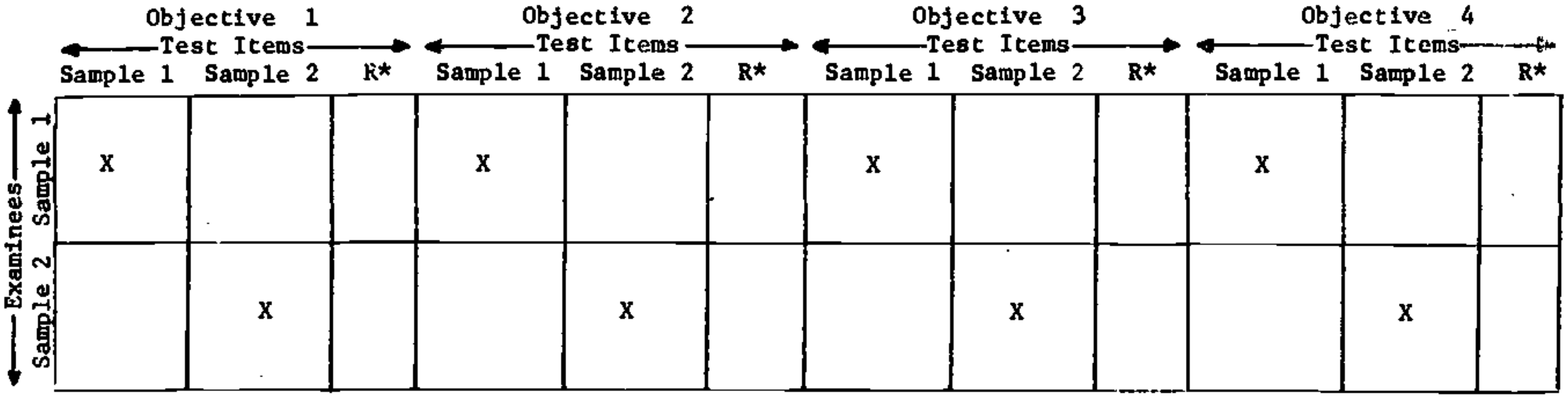
Multiple Matrix Samples (3): Exhaustive Examinee Sampling; Non-Exhaustive Item Sampling

Multiple Matrix Samples (3): Non-exhaustive Examinee Sampling; Non-Exhaustive Item Sampling

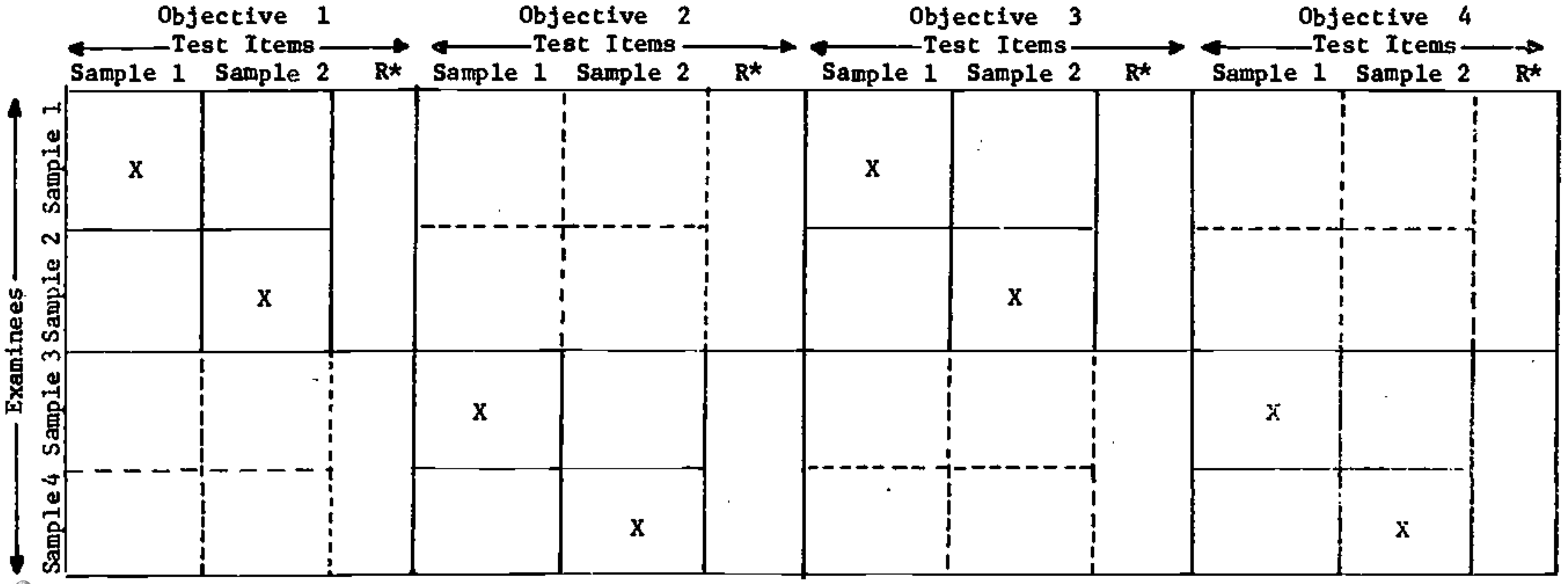
Since the test items that measure an objective are only a sample of the items from the domain of items of interest, all sampling plans will be non-exhaustive of the item domain.

*Remaining items (unused or unwritten) in the domain of items measuring an objective.

Figure 2 (continued)



Multiple Matrix Samples (2); Across Objectives (4); Exhaustive-Examinee Sampling; Non-Exhaustive Item Sampling.



Multiple Matrix Samples (2); Across Objectives (4); Non-Exhaustive Examinee Sampling; Non-Exhaustive Item Sampling.

(1) Estimation of the Mean and Variance of Domain Scores on an Objective

A parameter of considerable importance is the average domain score for the population of examinees on an objective.

The unbiased estimate of the average domain score μ is given by

$$\hat{\mu} = \frac{1}{t} \sum_{\ell=1}^{t} \overline{X}_{\ell} \qquad [1]$$

where the average domain score for test ℓ, $\overline{X}_{\ell}$, is given by

$$\overline{X}_{\ell} = \frac{1}{nk} \sum_{j=1}^{k} \sum_{i=1}^{n} X_{ij\ell}$$

The quantity $X_{ij\ell}$ is the score of the ith individual on the jth item measuring the objective under study in the ℓth test.

A convenient and intuitively appealing way to approach the estimation of the examinee domain score variance was presented by Sirotnik (1970). He rederived the formulae, presented earlier by Lord and Novick (1968), using an examinee-by-item analysis of variance design. Examinees and items are seen as random effects and the item and examinee population can be viewed as either finite or infinite depending on the design at hand. Through the usual analysis of variance procedure the mean square due to examinees (MS_E), mean square due to items (MS_I) and the residual or mean square due to interaction (MS_{EI}) can be calculated. From these, variance components of interest can be obtained as follows (for finite populations of examinees and items)

$$\hat{\sigma}^2_E = \frac{N-1}{N}\left[\frac{MS_E - (1 - \frac{k}{K})\ MS_{EI}}{k}\right] \quad ,$$

$$\hat{\sigma}^2_I = \frac{K-1}{K}\left[\frac{MS_I - (1 - \frac{n}{N})\ MS_{EI}}{n}\right] \quad ,$$

and

$$\hat{\sigma}^2_{EI} = \frac{(N-1)\ (K-1)}{NK}\ MS_{EI} \quad .$$

The statistic $\hat{\sigma}^2_E$ is then, the estimate of the population variance. If the size of the population of items (K) approaches infinity,

$$\hat{\sigma}^2_E = \frac{N-1}{N}\left[\frac{MS_E - MS_{EI}}{k}\right] \quad ,$$

and if both the population of items and the population of examinees is infinite, the variance is given as follows:

$$\hat{\sigma}^2_E = (MS_E - MS_{EI})/k$$

An estimate of domain score variance is obtained from each of the t tests and the t values are averaged resulting in a more stable estimate of the population parameter.

It is often of interest to compare the average domain score of a group to some established standard. For example, an average domain score of at least .80 may be required for "success" on a given objective. A statistical comparison of the estimate of average domain score to the standard would be helpful. Rather than comparing the estimate of average domain score to a standard, it may be of interest to compare two groups, for example, experimental and control groups. Although it is possible to test several other hypotheses using estimates obtained through multiple matrix procedures, the hypotheses discussed here are (a) $H_o: \mu = c$

versus $H_a: \mu \neq c$

and

(b) $H_o: \mu_1 = \mu_2$

versus $H_a: \mu_1 \neq \mu_2$

To test hypotheses concerning the parameter, μ, the estimate of the standard error must be calculated. The analysis of variance formulation can again be used to estimate the three variance components, σ_I^2, variance due to items, σ_E^2, variance due to examinees, and σ_{EI}^2, variance due to item-examinee interaction, for each test. As was mentioned earlier, these variance estimates are pooled across tests to yield a pooled variance estimate. The standard error of estimate of the mean domain score can then be expressed by:

$$\hat{\sigma}_{\hat{\mu}} = \{\frac{1}{tnk(N-1)(K-1)} \left[k(K-1)(N-nt)\hat{\sigma}_E^2 + n(N-1)(K-kt)\hat{\sigma}_I^2 + [(N-n)(K-k) + nk(t-1)]\hat{\sigma}_{EI}^2\right] \}^{\frac{1}{2}} \qquad [2]$$

Examinee sampling that is non-exhaustive, yet from a finite population and item-sampling that is non-exhaustive from an infinite population result in

$$\hat{\sigma}_{\hat{\mu}_1-\hat{\mu}_2}=\{\frac{2}{tnk(N-1)}\left[k(N-nt)\hat{\sigma}^2_E + n\ (N-1)\hat{\sigma}^2_I + (N-n)\hat{\sigma}^2_{EI}\right]\}^{\frac{1}{2}} \quad . \qquad [8]$$

Finally, if both the number of examinees and items are allowed to approach infinity, the expression simplifies to

$$\hat{\sigma}_{\hat{\mu}_1-\hat{\mu}_2}=\{\frac{2}{tnk}\left[k\hat{\sigma}^2_E + n\hat{\sigma}^2_I + \hat{\sigma}^2_{EI}\right]\}^{\frac{1}{2}} \quad . \qquad [9]$$

After choosing the correct standard error of the estimates, the test statistic can be calculated as follows

$$z = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_{\hat{\mu}_1-\hat{\mu}_2}} \quad .$$

This quantity is approximately distributed normally. Hence, the computed z value can be compared with the tabulated values and the appropriate decision concerning the hypothesis can be made.

In practice, the sampling of examinees will be usually <u>exhaustive</u> or <u>non-exhaustive</u> from a <u>finite</u> population and the sampling of items will be non-exhaustive from an infinite pool of itemsand therefore Equations [3] and [4], will prove to be more useful than either equation [2] or [5]. Likewise, when comparing two groups, equations [7] and [8] will be applicable more often than equation [6] or [9].

(2) Estimation of Domain Score Distribution on an Objective

Multiple matrix sampling procedures were introduced initially to enable test constructors to obtain better test score norms (Lord. 1962). By requiring schools to administer fewer test items it was felt that more representative test norms could be obtained because fewer schools would decline to participate in a norming study. This would result in more representative samples of examinees to estimate the distribution of test scores in the examinee population of interest.

Although matrix sampling was developed primarily for purposes of norm-referenced measurement, the evaluator who is using criterion-referenced measurement, may find the estimation of the entire distribution to be valuable. There are times when describing group performance on an objective by a mean and variance alone is insufficient. Information about particular percentiles may be needed. For example, it may be of interest to know the proportion of students who have domain scores above a value of .80 on a particular objective.

Several approaches to the estimation of an entire distribution have been investigated (Brandenburg & Forsyth, 1974a). Lord (1962) presented a successful application of item sampling using the negative hypergeometric distribution to estimate a test score distribution. This procedure is relatively straightforward since the distribution is fitted to the three parameters, mean, variance and number of items. Further work with the negative hypergeometric distribution was conducted by Shoemaker (1970). He systematically varied the number of tests, number of items per test, and the number of examinees receiving each test and studied the fit of the estimated distribution to the actual distribution. Shoemaker concluded that for small numbers of observations the fit is variable, but as the number of observations increases beyond a certain point (1.23% of the norm data base in this study) all procedures produce equivalent results.

Brandenburg and Forsyth (1974b) studied the use of multiple matrix sampling to estimate the parameters of the negative hypergeometric distribution. They compared the distribution to that obtained through estimation of the parameters of the Pearson Type I distribution. In order to specify a particular Pearson Type I curve, the first four moments must be estimated. These parameters were estimated through use of Lord's (1960) formulae. Brandenburg and Forsyth concluded that, in general, the Pearson Type I model tended to yield the better fit of the two models. Since the Pearson Type I procedure requires estimation of the first four moments, more items are required per test in order to get a stable estimate of the distribution. When the number of items per test is relatively small, the negative hypergeometric may be more appropriate since only two moments of distribution need to be estimated. More study is needed with regard to the effect of the choice of the sampling design on the fit of the models to the actual distribution. In particular, the study of the fit of the two models to various skewed distributions is critical, since much criterion-referenced test data seems to be either positively or negatively skewed.

(3) Estimation of Percent of Masters on an Objective

One of the major purposes of criterion-referenced measurement is to provide a mastery/non-mastery decision for a given individual. In program evaluation, however, information on a given individual is not critical. Reliable group information is what is needed to make program decisions. For example, if 85% of the population served by a particular program, achieved mastery status, the program must be accomplishing something. Whether or not 85% is an adequate level of mastery must be concluded by comparing the value to some previously established standard.

If every person in the entire population is tested with enough items to make a reliable mastery decision, the percent of masters can be obtained by simply calculating the percent of students classified as masters. Then, the obtained percent can be directly compared to some standard set by the program designers. If, however, it is impossible to test all examinees on all objectives with enough items on each to make reliable mastery decisions, it is necessary to do some careful sampling.

One solution is to carry out examinee-sampling on each objective, make reliable mastery decisions on the chosen sample of examinees by using a sufficient number of test items, and use the proportion of masters in the examinee sample to estimate the proportion of masters in the entire population.

The multiple matrix sampling plans presented earlier can apply in this situation, as long as enough items on given objectives are administered to individuals. But there are some special considerations when the variable of interest is the percent of examinees in the population who have achieved some minimum level of performance on an objective. When the number of items administered to a student does not allow for setting a performance standard equal to the one which applies to the domain of items measuring an objective, the resulting percent estimate will be biased. For example, suppose the performance standard is .80 and two items are administered per objective. There are only three possible cut-off scores: 0, .50, and 1.00. If 1.00 is selected, some examinees who can meet the .80 standard will be assigned to a non-mastery state and therefore the estimate of the percent of masters will be too low. On the other hand, if, .00 or .50 are selected, some examinees who could not meet the .80 standard will be assigned to a mastery state and therefore the estimate of the percent of masters will be too high. Clearly, if the "actual" and the "true" cut-off score differ, biased results (in a known direction) will be obtained and the seriousness of the bias will be related to the difference of the two cut-off scores. The implication of this is clear: sample examinees, and administer each examinee a sufficient number of items to enable the cut-off score on the sample of test items to equal the desired cut-off score in the pool of test items measuring the objective (i.e., if the true cut-off score is .75, the number of test items administered must be a multiple of 4 so that the cut-off score set on the sample of items can also be set equal to the value, .75). Assuming the amount of test data to be collected is fixed in estimating the percent of masters of an objective in a population, it is not clear whether it would be better to use (1) short tests and many examinees or (2) longer tests and fewer examinees.

It is also possible to approach the problem of estimating the percent of examinees exceeding some standard of performance (students defined as "masters" of the objective) through the use of procedures presented in the previous section. Rather than getting reliable mastery decisions on a sample of examinees and inferring the true percent of masters, it is possible to use multiple matrix sampling to estimate the entire score distribution and infer the percentage of examinees that lie above a given minimum level of performance. It remains to be seen which of the two procedures described results in the most efficient and yet accurate results.

Hypotheses concerning the percent of students reaching mastery, parallel those presented in the previous section. The first relates to the comparison of the estimated percent of masters to some pre-established standard. A second hypothesis of interest concerns a comparison of percent of masters across different objectives. Finally, there may be interest in a comparison of percent of masters across two or more groups.

(4) Estimation of the Percent of Examinees Mastering a a Given Percent of Objectives

It is often of interest to represent the success of a group on an entire set of objectives. For example, statements such as the following can be extremely descriptive of the success of a program:

"Eighty percent of the group mastered at least seventy-five percent of the objectives." To do this efficiently, it is possible to present samples of examinees with item samples selected from a representative subset of objectives. Inferences sre drawn to all examinees, to the entire item domain and finally, to the entire set of objectives. This procedure does, however, hinge on the "representativeness" of the subset of objectives.

Reliability of Group Scores

Discussions of various approaches to reliability of criterion-referenced measurement are readily available in the literature (for example, see Hambleton, Swaminathan, Algina, & Coulson, 1978). The emphasis in the work to date, however, has been on the reliability of individual test scores and associated decisions. There are ample methods and guidelines to aid the practitioner in estimating the reliability of domain score estimatea and mastery decisions. Since group information is of most interest to program evaluators, the reliability of the group statistics (average domain score and percent of masters, for example) are of concern, rather than reliability of individual scores. Reliability then is the accuracy of estimation of the group derived estimates. In the estimation of domain scores, the accuracy is expressed in terms of the standard error of estimation presented in Equations [3] through [6]. In the estimation of proportion of masters (P), the degree of precision is given (approximately) by the formula $\sqrt{\frac{P(1-P)}{tn}}$.

The variables that affect the accuracy, are the number of tests (t), number of examinees per group (n) and the number of items per test (k).

Suppose the content component of a particular program is comprised of 50 well-defined objectives. Let us also suppose the program is serving 5000 students (N=5000). It is appropriate at this time, for the evaluator to fix the desired accuracy of estimates and choose values of k, t and n that will result in standard errors of estimate less than some desired value. As suggested by Shoemaker (1973a) one possible procedure is to place, in the equation for the standard error of the estimate, an acceptable value for the standard error, the equation can then be solved for t, the number of tests. The difficulty with this procedure is that initial estimates of σ^2_{EI}, σ^2_E, and σ^2_I must be substituted in the expression. Rough estimates, however, could be obtained through pilot testing, or from norms studies. According to the guidelines presented by Shoemaker (1973a) the total number of observations, the product tkn, is the most important variable to consider when attempting to achieve a particular level of accuracy. As the total number of observations is increased, the size of the standard error of estimates correspondingly decreases. Another point presented by Shoemaker that is particularly important in this application relates to the distributional nature of the data. "For normal normative distributions, increases in the number of items per test are most effective in reducing standard errors of estimate; for negatively-skewed distributions or positively-skewed distributions, increases in the number of tests are most effective." Since criterion-referenced test data tend to be skewed, the most effective way of decreasing the standard error of estimate is to increase t, the number of tests. After deciding upon values of t, k and n, it may be the case that when all objectives are considered,testing time becomes prohibitive. For example, suppose the following values are decided upon, t = 10, k = 5, and n = 500. If all objectives are tested in a like manner, each student must respond to a total of 250 items (50 objectives x 5 items/objective). It may be necessary in this case to reduce k, or to administer fewer objectives to each individual. There is no unique solution to the choice of t, k, and n. The choices very often depend on practical considerations.

Conclusion

Program evaluators often find that it is important to evaluate programs with respect to the goals and objectives of the program and consequently they turn to criterion-referenced measurement. Criterion-referenced test scores can provide both descriptive and normative information. To date criterion-referenced test technology has been mainly directed toward information concerning individuals. In this paper, technical considerations associated with item selection and reliability assessment in relation to criterion-referenced tests constructed to provide group information were discussed. Hopefully, some of the ideas expressed in this paper will help to shape the technology for building tests and evaluating test scores in program evaluation studies.

References

Brandenburg, D. C., & Forsyth, R. A. Approximating standardized achievement test norms with a theoretical model. Educational and Psychological Measurement, 1974, 34, 3-9. (a)

Brandenburg, D. C., & Forsyth, R. A. The use of multiple matrix sampling to approximate norms distributions: An empirical comparison of two models. Educational and Psychological Measurement, 1974, 34, 475-486. (b)

Buros, O. (Ed.) The Eighth Mental Measurement Yearbook. Highland Park, NJ: The Gryphon Press, 1978.

Carver, R. The Coleman Report: Using inappropriately designed achievement tests. American Educational Research Journal, 1975, 12, 77-86.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Wenfeld, E. D., & York, R. L. Equality of educational opportunity, 2 volumes. (OE 38001; Superintendent of Documents Catalog No. FS 5.238:38001.) Washington, D.C.: U.S. Government Printing Office, 1966.

Ebel, R. L. The case for norm-referenced measurements. Educational Researcher, 1978, (December), 3-5.

Glass, G. V., & Ellett, F. S. Evaluation research. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual Review of Psychology. Palo Alto, CA: Annual Reviews, Inc., 1980.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1979. (2nd edition) (a)

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard setting. In R. Jaeger & C. Tittle (Eds.) Minimum competency achievement testing. Berkeley, CA: McCutchan Publishing Co., 1979. (b)

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

Lord, F. M. The use of true-score theory to predict moments of univariate and bivariate observed-score distributions. Psychometrika, 1960, 25, 325-342.

Lord, F. M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 2, 259-267.

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

Mehrens, W. A., & Ebel, R. L. Some comments on criterion-referenced and norm-referenced achievement tests. NCME Measurement in Education, 1979, 10, No. 1.

Perloff, R., Perloff, E., & Sussna, E. Program evaluation. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual Review of Psychology. Palo Alto, CA: Annual Reviews, Inc., 1980.

Popham, W. J. Educational evaluation. Englewood Cliffs, NJ.: Prentice-Hall, 1975.

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

Popham, W. J. The case for criterion-referenced measurements. Educational Researcher, 1978, (December), 6-10. (b)

Shoemaker, D. M. Allocation of items and examinees in estimating a norm distribution by item-sampling. Journal of Educational Measurement, 1970, 7, 123-128.

Shoemaker, D. M. Principles and procedures of multiple matrix sampling. Cambridge, MA: Ballinger Publishing Co., 1973. (a)

Shoemaker, D. M. A preliminary investigation of the theoretical sampling distributions of pooled estimates of moments in multiple matrix sampling. Southwest Regional Laboratory for Educational Research and Development Technical Memorandum, 1973. (b)

Shoemaker, D. M., & Knapp, T. R. A note on terminology and notation in matrix sampling. Journal of Educational Measurement, 1974, 11, 59-61.

Sirotnik, K. An analysis of variance framework for matrix sampling. Educational and Psychological Measurement, 1970, 30, 891-908.

Sirotnik, K. Some notes on the estimation formulas in matrix sampling. The Instructional Objectives Exchange, Los Angeles, California, Technical Paper No. 9, 1973.

Sirotnik, K. A. Introduction to matrix sampling for the practitioner. In W. J. Popham (Ed.), Evaluation in education: Current practices. San Francisco: McCutchan Publishers, 1974.

Worthen, B. R., & Sanders, J. R. Educational evaluation: Theory and practice. Worthington, OH: Charles A. Jones Publishing, 1973.