Adaptive Testing Applied to Hierarchically Structured Objectives-Based Programs

Ronald K. Hambleton and Daniel R. Eignor
University of Massachusetts, Amherst

Objectives-based instructional programs (e.g., Glaser & Nitko, 1971) were introduced to provide instructional programs that would be maximally adaptive to the needs of individual learners. While the specific methods of implementation have varied widely, common to all has been the notion that a curriculum should be defined by a set of objectives. Another common theme of objectives-based programs has been that student progress should be measured by comparing student performance to standards of performance set on the objectives defining a curriculum; student progress was not to be measured by comparing the performance among students (Hambleton, 1974).

Criterion-referenced tests were introduced initially by Glaser (1963) and Popham and Husek (1969) to provide a way for collecting the kind of information needed to assess student performance relative to a set of objectives. More recently there have been numerous contributions to the emerging field of criterion-referenced testing technology (e.g., Glaser & Nitko, 1971; Hambleton & Novick, 1973; Hambleton, Swaminathan, & Algina, 1976; Millman, 1974; Popham, 1975).

Student mastery of objectives in a segment of a curriculum is often determined by an administration of a criterion-referenced test. "Mastery" is inferred when a student's test performance on a set of items measuring an objective exceeds some minimum performance level. The minimum performance level for mastery is often referred to as a cutting score or passing score.

In theory, criterion-referenced test scores can be made as reliable and valid as necessary by adding additional test items. Unfortunately, making a mastery--non-mastery decision on each of the objectives measured by a criterion-referenced test often requires a considerable amount of testing time. Therefore, it is usually impractical to consider lengthening tests, particularly to the length that would often be necessary to accomplish some desired goal for reliability and validity of test scores.

Some critics have argued that there is already too much criterion-referenced testing in objectives-based programs. On the other hand, some increase in testing time can be defended on the grounds that test response data is closely tied to the objectives defining the curriculum and that the data are used to monitor student progress. Nevertheless, it seems clear that research is needed on procedures offering potential for reducing testing time without reducing the quality of decision-making from test score results.

The use of Bayesian statistical procedures represents one promising method for reducing testing time and/or improving the quality of mastery decisions (Hambleton & Novick, 1973; Novick & Jackson, 1974; Swaminathan, Hambleton, & Algina, 1975). This method is particularly appealing because it requires no change from the most common methods of test administration. Improvements in decision making are attributable to the utilization of information ignored by non-Bayesian procedures. Bayesian procedures may use not only the *direct information* provided by an examinee's test score, but they also make use of *collateral information* contained in the data of other examinees and of *prior information* on other relevant data that are available on the examinee (e.g., test scores from other segments of the course).

In one simulation study Hambleton, Hutten, and Swaminathan (1976) compared several Bayesian estimation procedures with several classical procedures for assessing student mastery and making instructional decisions. They reported modest gains from use of the Bayesian estimation procedures. On the negative side, Bayesian statistical procedures are based on restrictive assumptions, and robustness of the procedures has not been studied extensively. Also, some individuals feel that the utilization of group information to influence individual mastery estimates is a contradiction of one of the fundamental postulates of objectives-based instruction, that is, each student is judged on his/her own merits; thus, mastery decisions should not depend on the performance of other students.

A second promising solution to the testing time problem is offered by adaptive testing. Adaptive testing has been defined as a strategy for testing in which the sequence and number of test items a student receives are dependent upon his/her performance on earlier test items. Of special interest in this paper is the application of adaptive testing to hierarchically structured objectives-based programs. When the hierarchy of objectives is specified, inferences can be made about student mastery of objectives in the hierarchy which have not been tested. If, for example, a student is tested and found to have "mastered" a particular objective, all prerequisite objectives can also be considered to have been mastered. If an examinee has "not-mastered" an objective, it can be inferred that all objectives to which it is a prerequisite are also unmastered.

A considerable amount of work on adaptive testing has been done since the late 1960s (e.g., Lord, 1970; Weiss, 1977; Wood, 1973). Much of the research has concentrated on adaptive testing schemes for improving the precision of measurement of examinee ability while decreasing the amount of testing time. A second problem area (not independent of the first) is one of classifying examinees into "mastery" and "non-mastery" states for a set of objectives that can be arranged hierarchically.

An additional problem of considerable importance is to find optimum adaptive testing strategies for assigning examinees to "mastery states" for the objectives in a learning hierarchy. (The expression "learning hierarchy" will be used in this report to refer to a set of objectives arranged into a hierarchy reflecting dependencies among the objectives.) A further problem concerns the amount of testing time that can be saved in comparison with testing examinees on all of the objectives included in a learning hierarchy.

There are two areas requiring attention from researchers before results of empirical studies on adaptive testing will be of much value. The first area is the construction and validation of criterion-referenced tests; the usefulness of adaptive testing is related directly to the validity of criterion-referenced tests. Certainly, unless the intended interpretations of the criterion-referenced test scores are validated, decisions or descriptions based on the test data are suspect. The second area is the construction and validation of learning hierarchies. If information about a hierarchy of objectives is to be used to influence the adaptive testing plan, it is essential that a learning hierarchy be produced and validated.

Three topics will be covered in this paper: (1) construction and validation of criterion-referenced tests, (2) construction and validation of learning hierarchies, and (3) adaptive testing research.

Construction and Validation of Criterion-Referenced Tests

Since 1969 over 400 papers have been written on the topic of criterion-referenced testing. Unfortunately, this avalanche of papers reflects almost as many views and assumptions as there are contributors to the literature; and as a result, there has been substantial confusion in the field. However, with the important integrating work of Glaser and Nitko (1971); Millman (1974); Harris, Alkin, and Popham (1974); and Hambleton, Swaminathan, Algina, and Coulson (1978) much of the terminology has been standardized, many issues delineated, and many important technical matters resolved. It is now known how criterion-referenced tests are developed, test scores used, and test score reliability and validity assessed. Although many problems still remain, at least it is now possible to initiate a successful criterion-referenced testing program by drawing on a more than adequate testing technology.

A criterion-referenced test can be defined as a test constructed to permit the determination of an examinee's level of mastery relative to a "well-defined" behavior domain (Popham, 1975). Each behavior domain is keyed to an objective in the learning hierarchy. A criterion-referenced test is constructed to assess student mastery of each objective in a learning hierarchy. On many occasions, items from different behavior domains are included in the same test. In such cases, examinees receive scores based on their responses to items in each behavior domain. The definition of a criterion-referenced test is not unlike the definition offered by Millman (1974) for *domain-referenced tests*.

Preparation of Objectives

To operationalize the skills included in a learning hierarchy, the usual strategy is to write "behavioral objectives." Behavioral objectives have some desirable features, e.g., they are easy to write. Unfortunately, a behavioral objective usually lacks the clarity necessary to permit a clear determination of the domain of test items which measure the behaviors defined by the objective. If the proper domain of test items measuring an objective is not clear, it is impossible to select a representative sample of test items from that domain.

Current thinking among criterion-referenced test theorists is that objective statements should be expanded into domain specifications (Popham, 1975). Domain specifications are intended to reduce confusion among test users and item writers over the set of behaviors spanning the area defined by an objective. Improved clarification of domains of items measuring objectives will contribute substantially to the improvement of criterion-referenced test construction methods.

Popham outlined four steps for the development of domain specifications. First is the preparation of a general description; it could be a behavioral objective, a detailed description of the objective, or a short cryptic descriptor. Second, a sample test item is prepared. This step will help clarify the proper domain of test items and specify item format. In the third step, perhaps the most difficult, it is necessary to indicate the behaviors included in the domain. At this stage, the writer of a domain specification could list exemplary behaviors as well as behaviors which are not included in the domain specification if there might be some confusion over their status. In the fourth and final step, characteristics of response alternatives are specified.

The important aspect of these four steps is that they are *specific*. It is not necessary, however, that they specify a homogeneous set of behaviors. Domain specifications for objectives in a learning hierarchy will typically be more homogeneous than those written for other purposes. If domain specifications for objectives become too heterogeneous, it will become considerably more difficult to sequence objectives in a hierarchy.

Excellent examples of domain specifications are those prepared by Hively, Patterson, and Page (1968; see also Millman, 1974; Popham, 1975). Their work is based on two requirements: (1) All items which could be written from the domain to be tested must be written (or known) in advance of the final item selection process, and (2) a random or stratified random sampling procedure must be used in the item selection process.

One way to achieve these two requirements is through item forms analysis. An item form has the following characteristics:

1. It generates items with a fixed syntactical structure.
2. It contains one or more variable elements.
3. It defines a class of item sentences by specifying the replacement sets for the variable elements.

One of the obvious advantages of such a system is that the workload which would be required in writing the larger number of items needed to satisfy the two conditions above would be reduced.

The Hively et al. study is important in that it demonstrated that it was possible to develop and use item generation rules to construct a test. The study also underscored one of the major weaknesses of item generation procedures--the procedures are more easily employed with highly structured subject matter areas, such as mathematics. Although it is unlikely that the clarity of domain specifications offered by Hively and his colleagues can be produced in very many content areas, clarity is still necessary if criterion-referenced tests are to be useful for assessing mastery within an adaptive testing scheme.

The importance of having domain specifications cannot be overemphasized. What is typically desirable is to be able to interpret an examinee's test score as an estimate of an examinee's performance level in the larger domain of items measuring an objective. In addition, the scores are used to make instructional decisions (assign examinees to "mastery" and "non-mastery" states on each objective in the learning hierarchy). When the domain of items is unclear (or unspecified), only a "weak" criterion-referenced interpretation is possible, i.e., examinee test performance must be interpreted in terms of the items included in a test. Generalizations about student performance to a domain of behaviors, i.e., a "strong" interpretation, are not possible.

Item Writing

This step is not very different from writing test items for norm-referenced tests. There are a set of principles for item writing that should be followed; it is necessary, however, for item writers to attend to the domain specifications. Test items should "tap" behaviors in the domain of behaviors defined by domain specifications.

Item and Test Score Validity

Cronbach (1971), Messick (1975), and Linn (1977) have argued that to validate interpretations of criterion-referenced test scores (i.e., to determine what is being measured), it is necessary to proceed beyond a consideration of content validity. Until recently, it was thought that content validity considerations of a criterion-referenced test were sufficient. However, Messick (1975) stated:

> The major problem ... is that content validity ... is focused upon test *forms* rather than test *scores*, upon *instruments* rather than *measurements*. Inferences in educational and psychological measurement are made from scores, and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself it does not provide validity. (p. 960)

Content validity is a test characteristic. It does not change from one group of examinees to another. However, the validity of test score interpretations *will* vary from one testing situation to another; therefore, construct validation studies must be conducted. For example, if a criterion-referenced test is administered by mistake under highly speeded testing conditions, the meaning of the test scores obtained from the test administration will be different than if the test had been administered with more suitable time limits.

In preparing criterion-referenced tests for use in adaptive testing, both the content validity of the test and the construct validity of the test scores must be established. (For a lengthy discussion of the validity question, see Hambleton, 1977.)

Content Validity

Content validity of a criterion-referenced test is studied by investigating two questions:

1. Is the domain specification clear?
2. Is there agreement that a set of items adequately samples the behaviors defined in the domain specification?

The first question can be studied by comparing test items generated by different item writers and analyzing the judgments of content specialists about items relative to the domain they were developed to measure. Three techniques for the collection and analysis of the judgments of content specialists have been described by Rovinelli and Hambleton (1977).

The second question is difficult to investigate, unless the domain of items is completely revealed in a domain specification. The question can be investigated by Cronbach's (1971) interesting, but somewhat impractical (at least for small-scale test development projects), duplication experiment. The judgments of content specialists are also useful.

The matter of technical quality of test items is handled at the test development stage. It could be done at the content validation stage also by asking content specialists to address the matter of technical quality of items in their review.

Construct Validity

According to Messick (1975), the definition of construct validation is "the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning"(p. 955).

Construct validation studies begin with a statement of the intended use of the test scores; a statement of use will provide direction for the kind of evidence that is worth collecting in a construct validation study. Following is a description of some kinds of investigations that can be conducted to study construct validity of a criterion-referenced test.

Item statistics. When items in a domain are expected to be relatively homogeneous, a test developer can compare estimates of item difficulty parameters, item discrimination parameters, or both. Since items from a homogeneous domain of items measuring an objective would be expected to have similar item parameters, estimates of the parameters can be compared in order to detect items that deviate from the norm defined by the remaining items. Such "deviant" items can be carefully scrutinized for flaws that may be reducing their validity.

Instructional experiments. One solution, albeit costly, is designing experiments to determine the construct validity of criterion-referenced test scores. Individuals are randomly assigned to one of two groups. One group receives instruction on content defined by a domain specification. The other group receives no instruction. If the treatment is effective, and this could be determined from past experience, higher test scores by the experimental group would support the construct validity hypothesis. Experiments on other objectives could also be conducted, but it would be desirable to alternate control and experimental groups so that no single group of examinees would be denied instruction.

Factor analysis. Factor analysis is a commonly employed procedure for the dimensional analysis of items in a norm-referenced test or scores derived from different norm-referenced tests. It could be used in construct validation studies of criterion-referenced test scores. Perhaps one reason for its lack of use is that the usual input for factor analytic studies is correlations. Correlations are often low between items in a criterion-referenced test or between criterion-referenced test scores and other variables because criterion-referenced test-item and test-score variability is often not very great.

The problem can be remedied by choosing a sample of examinees that includes a range of scores. For example, a test developer should choose a sample of examinees that includes both masters and non-masters of the material measured by the test of interest. The research problem, in the language of factor analysis, becomes a problem of determining whether or not the factor pattern matrix has a prescribed form. If the intercorrelations among a set of items in a criterion-referenced test are factor analyzed, as many factors would be expected to be obtained in the factor solution as there are objectives measured by the test. Test items should "load" only on the factor (or objective) that they were designed to measure. Items deviating from this pattern would be carefully studied for flaws.

Experimental studies of sources of invalidity. There are numerous sources of error that can reduce the validity of a set of criterion-referenced test scores, and the influence of many factors is involved. Some relevant questions are:

1. How clear were the test directions?
2. Was there confusion in using the answer sheets?
3. Was the test administered under speeded testing conditions?
4. Were the examinees motivated to do their best?

To the extent that any of these (and many other) factors influence test scores, the descriptive interpretation of the test scores is weakened.

Required are experimental studies of potential sources of error to determine their effect on test scores. Results of these studies can be used to further clarify domain specifications. For example, if it is discovered that item format influenced test scores, the item type to be used could be specified after it is determined which item produced the most construct valid test scores.

Item Selection and Test Length

Items should be selected to be representative of the domain of items measuring the objective of interest. A random or stratified random sample of items is essential to permit generalizations from examinee test scores on the sample of test items to the larger domain of items.

How many items are needed in a test to measure each objective? When using criterion-referenced tests to assign examinees to mastery states (i.e., "mastery" and "non-mastery"), 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 tests very long, but this is impractical.

Millman's binomial test model. Millman (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, it is possible to determine the probability of misclassification conditional upon an examinee's true score, an advancement score, and the number of items in the test. (The advancement score is the minimum number of items that an examinee needs to answer correctly to be assigned to a mastery state, as distinguished from the cut-off score,which 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, it is possible to determine the test length and advancement score that produce a desired probability of misclassification for a *given* true score. In Millman's model, the assumption is made that examinees attempt each item in the test measuring an objective with probabilities equal to their "true scores" for the objective. ("True score," "domain score," or "level of functioning score" is the proportion of items in the domain of items measuring an objective that the examinee can answer correctly.)

It follows, then, that the probability of an examinee with true score p obtaining a test score x on an n-item set of items measuring an objective is given by

$$\text{Prob}\ (x|p) = \binom{n}{x} p^x (1-p)^{n-x} \quad . \qquad [1]$$

Let n=5 and p=.60. The probabilities of an examinee obtaining a score of 4 or 5 are .26 and .08, respectively. To calculate the probability of a "false-positive" error when both the advancement score and cutting score are set at 80%, it is necessary to calculate the probability that the examinee's test score equals or exceeds 80%. For an examinee with true score equal to .6, the probability of a false positive error is .34. In the same way, the probability of misclassifying an examinee (false positive error or false negative error) can be calculated for any examinee with known true score on a test of n items with known advancement score and cutting score.

Of course, in practice, examinee true score is unknown; in fact, it represents the characteristic of an examinee being estimated. Nevertheless, experience with Equation 1 can be very helpful in determining acceptable test lengths. Useful tables based on the binomial test model for various test lengths, true scores, advancement scores, and cutting scores have been reported by Millman (1973).

Adaptive sequential probability models. Millman's work is particularly helpful to classroom teachers and other criterion-referenced test developers. If, however, a computer terminal for test administration is available, more suitable solutions to the test length problem can be obtained. Ferguson (1969) used Wald's (1947) sequential probability ratio test to assign examinees to mastery states. Ferguson's procedure allows the tester to vary the test length for each examinee. Test length is varied to insure that a mastery--non-mastery decision for each examinee can be made so that the risks of a false positive or false negative error are below values set by the tester.

The tester specifies a minimum cutting score, p_0, that an examinee must meet to be assigned to a mastery state. Also, a second cutting score, p_1, where $p_1 < p_0$, is specified so that an examinee whose score is less than p_1 is assigned to a non-mastery state. Next, the tester specifies the probability of making a Type I (α) error (a false-negative error) and the probability of making a Type II (β) error (a false-positive error) that he/she is willing to tolerate.

If $\hat{p}$ is the proportion of items that an examinee has answered correctly, there are three possible mutually exclusive decisions:

1. If $\hat{p} \geq p_0$, classify the examinee as a master.
2. If $\hat{p} \leq p_1$, classify the examinee as a non-master.
3. If $p_1 < \hat{p} < p_0$, do not make a decision; administer more test items.

More formally, the sequential probability ratio test of strength (α,β) for testing

$$H_0 : p \geq p_0$$
$$H_1 : p \leq p_1$$

can be viewed as a Bernoulli-type experiment, which means results fit the binomial distribution. Here, p is the true score for the examinee. The risks of misclassification are as follows: The probability of a false-negative error ($\hat{p} \leq p_1$, $p \geq p_0$) should not be greater than α and the probability of a false-positive error ($\hat{p} \geq p_0$, $p \leq p_1$) should not be greater than β.

Once p_0, p_1, α, and β are specified for any particular objective, it is possible to prepare a table indicating the number of items that must be passed (or failed) at test lengths from 1 to n items in order to lead to mastery decisions (or non-mastery decisions). Formulas for preparing a table are offered by Wald (1947); applications of the formulas to adaptive testing are provided by Ferguson (1969).

Reliability

What form should reliability take for a criterion-referenced test? Suppose instructional decisions are made by comparing examinee test performance to a minimum proficiency level or cutting score for each objective. Of interest is the consistency of "mastery" and "non-mastery" decisions across parallel-forms (or repeated administration) of the criterion-referenced test. An intuitively appealing measure of agreement between the decisions made (for each objective) on the two administrations is $p_{11} + p_{22}$, where p_{ii} is the proportion of examinees placed in the i^{th} mastery state on each test administration. 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, designated κ (Swaminathan, Hambleton, & Algina, 1974), takes into account chance agreement and thus appears to be somewhat more appropriate for use with criterion-referenced tests. Coefficient kappa

(the expression for reliability or consistency of mastery decisions on an objective) is defined as

$$\kappa = (p_o - p_c) \; / \; (1 - p_c), \qquad [2]$$

where p_o, the observed proportion of agreement, is given by

$$p_o = \sum_{i=1}^{2} p_{ii}; \qquad [3]$$

and p_c, the expected proportion of agreement, is given by

$$p_c = \sum_{i=1}^{2} p_{i.} \; p_{.i} \; . \qquad [4]$$

Both $p_{i.}$ 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, kappa can be thought of as the proportion of agreement that exists over and above that which can be expected by chance alone. Kappa provides a useful indicator of decision-making consistency. If it is lower than desirable, test length should be increased. When Wald's sequential probability ratio test is used, reliability or consistency of decision making across a group of examinees for particular objectives can be replaced by the values of α and β chosen for each objective.

Construction and Validation of Learning Hierarchies

One of the most promising lines of research to come out of learning theory investigations over the past fifteen years is the concept of learning hierarchies. Starting with a study by Gagné (1962), the research in this area has progressed until at present there is a well-defined methodology for the development of learning hierarchies and statistical tests for validating posited hierarchies. The state of the field is developed sufficiently so that it can be easily linked to related fields, such as adaptive testing.

Gagné (1970) used the term "learning hierarchy" to designate the set of dependencies among component skills or objectives within a learning task. He also suggested that learning hierarchies might define optimal sequences for presenting learning events. Clearly, besides the instructional process, establishing learning hierarchies has some very useful implications for the testing process. The objectives identified in a learning hierarchy can be measured using the test development and validation methods outlined in the previous section. An adaptive testing scheme could be used to determine which criterion-referenced tests would need to be administered in order to determine an examinee's mastery states on objectives in a learning hierarchy.

Generating Learning Hierarchies

Most researchers investigating learning hierarchies develop a provisional ordering of the instructional objectives comprising the hierarchy before initiating any type of validation procedure. At present there are two possible ways of developing provisional orderings of objectives in a hierarchy. Passmore (1974) aptly named these two methods "introspection" and "statistical fishing."

Gagné (1962) developed the logical questioning technique called "introspection." The researcher takes the final objective and asks, "What would an individual need to know to display competence in this subject matter?" The same question is then applied to each of the behaviors specified by the last application of the questioning techniques. This procedure is continually used until simple behaviors are reached which cannot be linked further to other necessary pre-behaviors. The generated sequence can then be represented as a hierarchy, which at this point can be considered only as provisional. There will be general behaviors at the top of the hierarchy and more specific, subordinate behaviors near the bottom.

Almost all of the learning hierarchy research done to date has proceeded from a provisional hierarchy developed in this particular fashion. The method presupposes that the individual doing the questioning is well acquainted with the subject matter and is capable of relating necessary pre-behaviors to the considered terminal behavior. This is critical because, as will be discussed later, the methodology for validating hierarchies can only lead to acceptance or rejection of a posited hierarchical connection. It cannot generate hierarchical links that have not been provisionally specified. Thus, it is critical that an individual or individuals well acquainted with the domain be involved in the introspection process.

"Statistical fishing methods" for developing hierarchical clusters, which could be applied to learning hierarchy research, have been advanced but to date have seldom been utilized. This would appear to be for two reasons. First, the methodology requires the understanding of statistical procedures not usually directly in the researcher's command. Second, the available literature has not been directly related to the establishment of learning hierarchies, thus requiring the researcher to link two fields not presently related. For instance, much of the work done has used measures of similarity between examinees, while the work on hierarchies would require some type of measure involving objectives.

The available procedures of a statistical nature involve either the application of statistical clustering techniques, such as discriminant analysis and hierarchical cluster analysis (Tryon & Bailey, 1970) or application of hierarchical clustering schemes based on principles of numerical taxonomy (McQuitty, 1970; Johnson, 1967). The former methods either yield final clusters with no hierarchical structure or depend upon an a priori definition of the number of clusters. It would seem, therefore, that the methods related to numerical taxonomy would be preferred.

The iterative procedures available relate similarity measures on individuals at each iterative stage until, at the final stage, all individuals are related and belong to one overall cluster. Researchers must make a decision as to which iterative stage they want to view the hierarchical structure. The Baker (1972) article contains an excellent example of the use of Johnson's (1967) MAX procedure, in which the operation of the clustering algorithm can be seen. This example and others are based upon individual similarity measures, but the procedures appear to be adaptable to usage involving similarity measures on objectives.

Evaluating Provisional Hierarchies

Once a provisional ordering of objectives in a learning hierarchy has been advanced, the ordering is evaluated to determine whether the connections between objectives should be rejected or accepted. It is at this level of hierarchy development that extensive research has been done. An extensive review of methods for validating learning hierarchies was prepared by White (1973), who then (White, 1974a, 1974b) developed a new model for validating learning hierarchies that appears to be the best possible procedure to use at present. Following is a brief explanation of the steps involved, taken and adapted from White (1974a):

1. Define the top objective of the hierarchy.
2. Derive, using Gagné's method of questioning, the subordinate objectives, being careful not to include verbalized knowledge objectives.
3. Check the postulated hierarchy with subject matter specialists.
4. Subdivide the objectives, if necessary, so that clear definitions are obtained.
5. Using a sample of students, check that objectives in the hierarchy are distinct (see White, 1974a).
6. Prepare instructional materials for the objectives with the test items (two or more for each objective) to be administered following instruction on the objectives.
7. Have a suitable number of students (at least 150) work through the instructional materials.
8. Analyze the results to see whether the postulated connections can be rejected, using the statistical test developed by White and Clark (1973).
9. Remove from the hierarchy all connections for which the probability of a hierarchical connection is small.

White field-tested his procedure (White, 1974a) as did Linke (1975); both found the procedure to work well with the objectives being taught and tested. It is noteworthy that these investigations and those of Gagné and others (see White, 1973) all involve hierarchies in the areas of science and mathematics. Possible reasons for this fact may be twofold: One, the method of questioning about necessary prerequisite skills may not work well for the social science and language fields; perhaps some sort of statistical clustering method needs to be employed. Two, until recently, the problems of writing higher order objectives for the social science and language fields may have proven to be a hindrance. Whether the methodology outlined by White can be applied to areas other than math or science remains at present a research question of considerable importance.

Statistical Techniques for Validating Hierarchies

As discussed in the previous section, a statistical test is desirable for ascertaining whether two objectives are connected in a linear fashion. Researchers who did the early work in the field of learning hierarchies used a number of different indices, all of which suffered to a greater or lesser degree from the following two maladies: (1) The indices can have values that indicate a hierarchical connection, even when the objectives are independent, and (2) the procedures are deterministic and do not allow for statistical tests of fit. Noteworthy of the indices falling into this category are Gagné and Paradise's (1961) "proportion positive transfer," Resnick and Wang's (1969) use of Guttman's coefficient of reproducibility, and Capie and Jones' (1971) use of the phi coefficient.

White and Clark (1973) developed a statistical test that can be used when more than one test item is used to measure each objective. A hierarchical relationship is postulated. Then the number of examinees is observed who answer all the test items for the lower objective incorrectly and answer all the test items for the higher objective correctly. The connection is judged invalid when this number of examinees exceeds a critical value, specified for the probability of wrongly rejecting the null hypothesis that the connection is hierarchical. Examples of the procedure are discussed in White and Clark (1973) and White (1974a).

While the statistical distribution theory is somewhat complicated, the practical application (based in part on the use of power functions) is straightforward. A cogent word of caution has been advanced by Passmore (1974), having to do with the relationship between sample size and the power of significance tests. Passmore has advised that valid connections may be rejected if standard hypothesis testing procedures are used with the large samples White has suggested. Passmore has further indicated that the relationship of White and Clark's power function to sample size considerations should be more fully explained to practitioners.

Recently, Dayton and Macready (1976) have developed a more general set of statistical procedures for validating hierarchies that subsume White and Clark's work as a component. This procedure can be used to test a whole hierarchy, whereas all previous methods test connections between pairs of objectives in a hierarchy. The model offers a χ^2 goodness of fit test of observed proportions data to expected proportions generated on the basis of the hypothesized hierarchical relationship. Furthermore, the model allows for the testing of arbitrary hierarchies, which include linear and branching patterns; and it allows for concept attainment models, in which the hypothesized pattern vector consists of ones or zeros to specify mastery or non-mastery of objectives in the hierarchy.

Dayton and Macready have offered two reasons why they feel that their probabilistic model is preferable to use when testing the connection between two objectives hypothesized to be hierarchical, rather than White and Clark's test. First, the estimation procedure of White and Clark is not maximum likelihood, while Dayton and Macready's method is; the Dayton and Macready method is thus more desirable from the point of view of sampling properties. Second,

using the Dayton and Macready procedure, a more general test of inclusion can be performed. Two sets of pattern vectors can be developed--one implying inclusion or a hierarchical connection and the other having the pattern vector for inclusion augmented by the discrepant pattern configuration, which demonstrates a non-hierarchical relationship. Then a χ^2 test can be performed on the difference between the expected proportions generated from the two solutions of the model. A significant χ^2 on the difference would indicate a non-hierarchical connection.

The model and statistical procedures developed by Dayton and Macready represent a significant development in the validation of learning hierarchies. In addition to the facts that the model and statistical test can be used to validate an entire hierarchy and that there are available computer programs to use, the procedure has one important advantage not offered by White and Clark's procedure. While an a priori hierarchy may yield an insignificant χ^2 when the expected proportions generated are compared to the observed proportions, examination of model parameters (along with their standard errors) can reveal places in the a priori hierarchy which are inadequately specified, thereby suggesting possible changes in the hierarchy. This is the important advantage offered by the Dayton and Macready model not offered by other available procedures.

The model does more than reject hypothesized hierarchies; it also suggests areas of the hierarchy where revision is necessary. Pushed to the limit, the procedure could be used as a pure discovery procedure from which a hierarchy could be built. This then would offer another method for building proposed hierarchies besides introspection and hierarchical clustering techniques. Dayton and Macready have mentioned that, at present, they have not utilized the procedure in this fashion. With the work of White, Dayton and Macready, and others (for example, Bart & Krus, 1973; Boozer & Lindvall, 1971; Macready, 1975), the statistical generation procedures and the necessary methods for generating and validating hierarchies now appear to be in a usable form.

Adaptive Testing Research

To date there have been only two investigations of adaptive testing to learning hierarchies (Ferguson, 1969; Spineti & Hambleton, 1977). The only other related study was by Vale (1977) and was an investigation of misclassification errors. Ferguson (1969) was concerned with classifying students as "masters" or "non-masters" on each objective in a learning hierarchy. The routing strategy was complex (involving the sequential ratio test described earlier) and required a computer to perform the actual routing. Ferguson found a 60% saving in number of items administered in the computerized administration using a variety of adaptive testing procedures. A test-retest of the adaptive testing procedure gave high reliability, with the reliabilities of the adaptive testing classifications higher than those of the paper-and-pencil conventional test approach.

An important consideration in the work of Spineti and Hambleton (1977) has been that the adaptive testing strategies under investigation be implementable without the aid of computer terminals. Such a restriction clearly sets this work apart from that of Ferguson and most of the other research on adaptive testing, with the exception of the self-scoring flexilevel testing work of Lord (1971). The primary effect of the restriction is that it

eliminates the possibility of using complex decision-making rules such as the one adopted by Ferguson (1969). The concern has been to study the effectiveness of a multitude of adaptive testing strategies that could be implemented in objectives-based programs without the aid of computers. Since few objectives-based instructional programs have access to computer terminals for testing, this restriction was imposed so that the results would be of maximum usefulness. A fixed number of items was required to assess mastery of each objective tested, items were scored right or wrong, and all items measuring a particular objective were assumed to have similar statistical properties. Examinee performance on the test items was assumed to be represented by the binomial test model (Lord & Novick, 1968).

The interactive effects of several factors (test length, cutting score, and starting point) on the accuracy of mastery classification decisions and the amount of testing time in adaptive testing schemes were investigated. Values of each factor were combined to generate a multitude of adaptive testing strategies for study with two learning hierarchies and three different distributions of true scores across the hierarchies. The study was conducted via computer simulation techniques. Therefore, there was no need to be concerned about problems raised earlier in this paper, i.e., the problems of developing and validating criterion-referenced tests and learning hierarchies.

Of the many learning hierarchies reported in the educational literature, two were selected for study. These were the learning hierarchies for hydrolysis of salts (Gagné, 1970) and addition-subtraction (Ferguson, 1969). The second one was selected so some of the results of this study could be compared with Ferguson's results. The two learning hierarchies are shown in Figures 1 and 2 and are referred to as Hierarchy A and B, respectively. It was found that it was possible to obtain an overall reduction of more than 50% in testing time by introducing an adaptive testing scheme. With Gagné's hierarchy (Hierarchy A), the adaptive testing strategies resulted, on the average, in an overall reduction of testing time of 59.2%. With Ferguson's hierarchy (Hierarchy B), there was a 53.2% reduction in testing time. It is likely that adaptive testing strategies with Hierarchy B were not quite as effective as with Hierarchy A because Hierarchy B had two terminal objectives, whereas Hierarchy A had only one. The difference highlighted the importance of the particular form of the learning hierarchy on the effectiveness of adaptive testing strategies.

The results of this study on the saving of testing time varied from 50% to 70% and compared favorably with the empirical results of Ferguson (1969). He reported a saving of testing time of 60% over conventional testing. The similarity of the results added validity to the appropriateness of the simulation procedures of the present study.

The reduction in testing time derived from the adaptive testing strategies was impressive; however, it would have meant little if the total number of errors of classification was substantially larger than with conventional testing. In fact, with Hierarchy A the adaptive testing strategies resulted in a slightly lower number of errors of classification than with conventional testing. The reverse was true with Hierarchy B; but, again, the differences were slight. These findings, along with the information on the comparisons

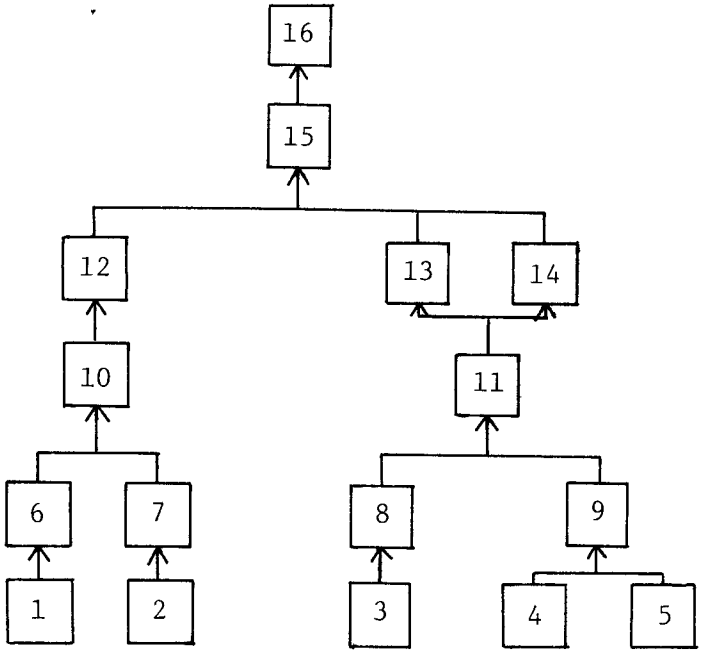

Figure 1. Gagne's Hydrolysis of Salts Hierarchy

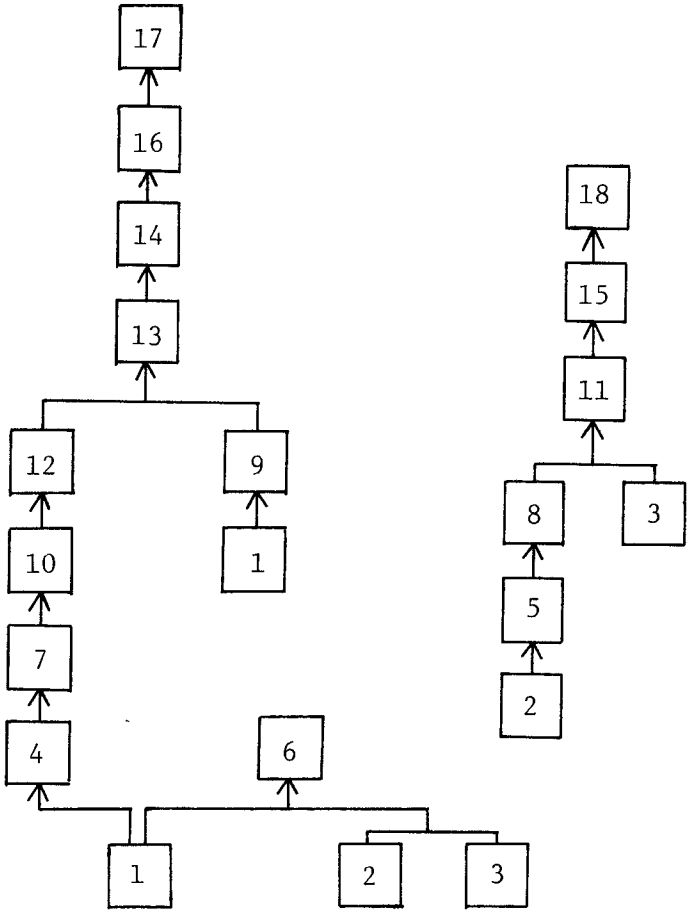

Figure 2. Ferguson's Addition-Subtraction Hierarchy

of testing time for conventional and adaptive testing, provide strong support for the speculations of many researchers. That is, by using an adaptive testing strategy in the context of learning hierarchies, there is much to be gained in terms of testing time efficiency without any significant loss in the accuracy of decision making.

It was dramatically clear from the numerous simulations that considerable saving in testing time was gained through implementing an adaptive testing strategy. Whereas the Ferguson adaptive testing strategies could only be implemented with the aid of computer-testing terminals, the Spineti-Hambleton testing strategies were simple enough to be implemented in a regular classroom with the aid of a "programmed instruction type" of test booklet.

Conclusions

The application of adaptive testing to learning hierarchies is substantially different from other applications and therefore includes some unique problems. First, while in any testing project there is concern about the generation of an item pool, the problem of developing an appropriate item pool for each criterion-referenced test is especially difficult; and the technology for accomplishing the task is quite new and not well understood. However, the "quality" of the item pool for a criterion-referenced test will be directly related to the overall success of an adaptive testing scheme. The problem must, therefore, be given careful attention.

Second, there is the unique problem of developing and validating learning hierarchies. Because of the inter-relationship between adaptive testing schemes and a learning hierarchy, the success of any adaptive testing scheme will depend on the "validity" of the learning hierarchy under investigation. In validating learning hierarchies, there are psychological as well as statistical problems involved. For example, several researchers have reported that while examinees may learn material in the sequence defined by a learning hierarchy, they may forget the information learned in any order. Thus, students may be able to perform a terminal objective although they have forgotten several of the prerequisite skills. The implications of this phenomenon for the validation of learning hierarchies and adaptive testing research are not clear. Third, classification problems, as opposed to measurement problems, are of interest. There has been relatively little research on using adaptive testing schemes to classify examinees into two or more categories.

There are adequate technologies to develop and to validate both criterion-referenced tests and learning hierarchies. Further refinements and advancements to the technology will take place as more researchers work in the area and encounter implementation problems. It will be especially interesting to observe the development of learning hierarchies and criterion-referenced tests in areas besides science and mathematics.

In adaptive testing research on learning hierarchies, it is important to distinguish between computer-assisted and non-computer-assisted test administration procedures. When computers are available for test administration, the possibility of using latent trait theory models and methods should be consid-

ered. To date, it is unclear how this application of latent trait theory could be accomplished. However, latent trait theory is now an established part of adaptive testing methodology for solving measurement problems. One possible problem may be that there usually are not sufficient items measuring any single objective to obtain a satisfactory latent trait ability estimate. This problem requires study as does the problem of optimal routing methods.

With non-computer-administered tests, some field studies need to be initiated. The design of such a study would involve developing a programmed instruction booklet which would include (1) test items designed to measure specific objectives in a learning hierarchy, (2) a self-scoring device, and (3) 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.

References

Baker, R. B. Numerical taxonomy for educational researchers. Review of Educational Research, 1972, 42, 345-359.

Bart, W. M., & Krus, D. J. An ordering-theoretic method to determine hierarchies among items. Educational and Psychological Measurement, 1973, 33, 291-300.

Boozer, R. F., & Lindvall, C. M. An investigation of selected procedures for the development and evaluation of hierarchical curriculum structures. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center, 1971.

Capie, W., & Jones, H. L. An assessment of hierarchy validation techniques. Journal of Research in Science Teaching, 1971, 8, 137-147.

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

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

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

Gagné, R. M. The acquistion of knowledge. Psychological Review, 1962, 69, 355-365.

Gagné, R. M. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Gagné, R.M. The conditions of learning (2nd ed.). New York, NY: Holt, Rinehart, & Winston, 1970.

Gagné, R. M., & Bassler, O. C. Study of retention of some topics in non-metric geometry. Journal of Educational Psychology, 1963, 54, 123-131.

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75, (14, Whole No. 518).

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

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

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

Hambleton, R. K. Validation of criterion-referenced test score interpretations. A paper presented at the Third International Symposium on Educational Testing, University of Leiden, The Netherlands, 1977.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

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

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

Hambleton, R. K., Swaminathan, H., & Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement. New York: Wiley, 1976.

Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.). Problems in criterion-referenced measurement (CSE Monograph Series in Evaluation, No. 3). Los Angeles, CA: University of California, Center for the Study of Evaluation, 1974.

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

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

Linke, R. D. Replicative studies in hierarchical learning of graphical interpretation skills, British Journal of Educational Psychology, 1975, 45, 39-46.

Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1977.

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

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

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

Macready, G. B. The structure of domain hierarchies found within a domain-referenced testing system. Educational and Psychological Measurement, 1975, 35, 583-598.

McQuitty, L. L. Hierarchical classification by multiple linkage. Educational and Psychological Measurement, 1970, 30, 3-20.

Messick, S. A. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current practices. San Francisco, CA: McCutchan, 1974.

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

Passmore, D. L. Sequencing learning events in performance-based instructional systems. Paper presented at annual meeting of the Rocky Mountain Educational Research Association, Albuquerque, New Mexico, 1974.

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

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

Resnick, L. B., & Wang, M. C. Approaches to the validation of learning hierarchies. Proceedings of the Eighteenth Annual Regional Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1969.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch, 1977, 2, 49-60.

Spineti, J. P., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objectives-based instructional programs. Educational and Psychological Measurement, 1977, 37, 139-158.

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. 87-98.

Tyron, R. C., & Bailey, D. E. Cluster analysis. New York, NY: McGraw-Hill, 1970.

Vale, C. D. Adaptive testing and the problem of classification. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977. (NTIS No. ADA038114).

Wald, A. Sequential analysis. New York: Wiley, 1947.

Weiss, D. J. (Ed.). Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977. (NITS No. ADA038114).

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. (a)

White, R. T. A model for the validation of learning hierarchies. Journal of Research in Science Teaching, 1974, 11, 1-3. (b)

White, R. T. Indexes used in testing the validity of learning hierarchies. Journal of Research in Science Teaching, 1974, 11, 61-66. (c)

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

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

Acknowledgements

The project reported herein was performed pursuant 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.