PROBLEM:

EVALUATING THE RESULTS OF COMPUTERIZED ADAPTIVE TESTING

James B. Sympson
University of Minnesota

The problem of evaluating a testing procedure is not unique to adaptive testing. Thus, many of the comments I will make are applicable in a broader context than the one we are dealing with today. On the other hand, several considerations will be mentioned in connection with adaptive testing that do not exist under other circumstances. Before discussing methods of evaluation, we should first be more precise about what it is we wish to evaluate.

Elements of a Testing Procedure

A testing procedure can be conceived of as a composite process that has six component elements. These elements are: a theory of the trait being measured, a strategy of item selection, a medium for item administration, a medium for responding, a mode of item response, and a scoring procedure.

By "theory of the trait" I mean the entire network of hypotheses and deductions associated with the construct we are attempting to measure. This would include statements about the nature of the trait, its relationships with other traits, and its relationships with a variety of observable variables. The most fundamental theoretical hypotheses in current latent trait theories are hypotheses regarding the form of the item characteristic curve.

By "strategy of item selection" I mean the rule, or set of rules, that determines which items in a large item pool will be administered to a given testee. In non-adaptive testing procedures all testees are administered the same set of items. As Mr. Vale has illustrated, in adaptive testing the set of items administered to a testee depends on his/her responses at the time of testing.

By "medium for item administration" I refer to the method by which the test stimuli are delivered. Examples include: verbal medium, as in individual clinical testing; printed medium, as in paper and pencil testing; and electronic medium, as in computer-controlled testing via cathode-ray-tube terminals (CRTs).

By "medium for responding" I refer to the method by which the testee indicates his/her response. This includes vocal responses, written responses, and responses typed in at a teletype or CRT keyboard.

By "mode of item response" I refer to the type of response required from the testee. In many cases the testee either indicates which of several response alternatives is correct or agrees or disagrees with a statement. Other possibilities include free-recall responding and confidence weighting schemes.

By "scoring procedure" I refer to the rule, or set of rules, by which the testee's responses are converted to a summary statement (usually quantitative) about the testee's status on the trait dimension of interest.

This analysis of a testing procedure into component elements leads me to reject the idea of evaluating any such procedure as an undifferentiated composite. Rather, we should attempt to evaluate the individual effects of each element. Evaluation of these elements can best be achieved by comparing two testing procedures that differ in only one element. If one testing procedure is found superior to the other, we may attribute this superiority to the one element in which the procedures differ.

Unfortunately, this approach to evaluating the elements of a testing procedure is not always possible. Some item selection strategies cannot be implemented in any other medium than computer administration. Similarly, some scoring methods presume that items have been selected in a certain way during the testing process. Thus, it is not possible to implement every conceivable combination of testing procedure elements. This means that in some instances one must compare testing procedures which differ in two (or possibly more) elements. Under such circumstances the effects of the elements which differ will be confounded. Research in adaptive testing will progress best, however, when efforts are made to evaluate these elements of a testing procedure in terms of their unconfounded effects.

Classes of Evaluative Criteria

Many characteristics of a testing procedure can be subjected to evaluative scrutiny. Most of these characteristics can be considered as belonging to one of four classes of evaluative criteria. These classes are: validating criteria, theoretical criteria, psycho-social criteria, and cost criteria.

Validating criteria and theoretical criteria have one principal feature in common. They are based on the characteristics of scores generated by a testing procedure. This may be contrasted with psycho-social criteria, which involve consideration of the psychological and social effects of a testing procedure, and cost criteria, which involve consideration of economic costs and benefits. Validating and theoretical criteria differ in that the former serve to establish the construct validity of a measurement procedure (Cronbach & Meehl, 1955) while the latter do not. They also differ with regard to the type of research they are based upon. Validating criteria require empirical research while theoretical criteria are examined via either mathematical derivations or computer simulations.

Validating criteria. Most important for the evaluation of testing procedures is the role that theory plays in telling the researcher what to expect from an adequate measure of the trait. Given a theory of the trait to be measured, conclusions regarding the "proper" characteristics of measures of that trait may be derived. Evaluation in terms of validating criteria proceeds by determining the extent to which a testing procedure generates scores that possess these characteristics.

Validating criteria include: stability coefficients, internal consistency coefficients, alternate form correlations, correlations with other tests, correlations with non-test variables, characteristics of score distributions in specified subject groups, differences between score distributions generated by different subject groups, and statistical or graphical methods for assessing goodness-of-fit.

A theory of the trait to be measured should indicate how much stability over time to expect in assessing the trait. Testing procedures that generate scores with the expected degree of stability from test to retest should be evaluated more highly than procedures giving scores that do not conform to expectation.

On the presumption that all the items in a test tap the same latent dimension, one expects high reliability for tests of sufficient length. With non-adaptive testing procedures, coefficient alpha (Cronbach, 1951) or related indices provide a suitable index for estimating this test characteristic. However, in adaptive testing different subjects are administered different items and the calculation of such indices is not possible. This forces the researcher to rely on alternate form correlations to estimate the reliability of scores from adaptive test procedures. It should be noted that in latent trait theory measurement error is seen to vary as a function of status on the latent dimension. Thus, overall reliability indices are generally not as important in latent trait theory as they are in classical measurement theory.

An adequate theory of the trait will imply a pattern of correlations between scores generated by a valid testing procedure and scores on other tests. Similarly, an expected pattern of correlations with various non-test variables (e.g., age, grade average, etc.) will be specifiable. In evaluating a testing procedure, one should determine whether the anticipated correlational patterns emerge.

In some situations one can specify how the scores of one or more selected subject groups should be distributed. The testing procedure that generates score distributions with the anticipated characteristics is to be considered superior to one that does not.

Finally, if the theory of the trait includes hypotheses or deductions about the form of the relationships among certain variables, statistical and graphical approaches to assessing the goodness-of-fit of empirical data to a theoretical model can be utilized (see, for example, Bock & Lieberman, 1970).

Theoretical criteria. Theoretical criteria, while also based on the characteristics of scores from a testing procedure, cannot establish the construct validity of the procedure. These criteria _assume_ the validity of certain critical theoretical hypotheses. They do not provide a method for testing these hypotheses. Thus, theoretical criteria can only be used to establish the superiority of one testing procedure over another if the two procedures have equal prior claim to construct validity.

Theoretical criteria include: distributions of latent trait estimates, correlations with latent trait scores, information curves, relative efficiency curves, bias curves, standard error of measurement (SEM) curves. and various types of "robustness".

In general, the use of these theoretical criteria requires that some particular form of item characteristic curve be assumed and that true item parameters be specified. Once these requirements are met, the various theoretical criteria can be obtained through either mathematical derivations or computer Monte Carlo runs. These criteria cannot be used in live-testing studies where the testee's status on the latent trait is unknown.

Given a particular form for the item characteristic curve and the parameter values for an item pool, it is possible to conduct a computer simulation in which simulated subjects with known latent trait scores are administered items under various item selection strategies and scoring methods. Following simulated testing, the researcher can compare the frequency distribution of the latent trait estimates to the distribution of known latent trait scores and can correlate the two sets of values.

It might seem that the testing procedure which generates estimates correlating most highly with latent trait standing should be preferred to other procedures. However, the correlation between latent trait estimates and latent trait scores is a joint function of the distribution of the testee population and the measurement properties of the testing procedure. In many cases a change in the distribution of the input population can lead to a different ordering of the testing procedures. Criteria that reflect the measurement properties of a testing procedure, but are not dependent on assumptions about the population of testees, are desirable. The remaining theoretical criteria have this property.

The "information" available from a testing procedure at some particular level of the latent trait was defined by Mr. McBride as the squared ratio of the slope of the test characteristic curve to the standard deviation of test scores at that level. If we plot the amount of information available from a testing procedure as a function of status on the latent trait, we generate the information curve (Birnbaum, 1968, pp. 460-468) for the procedure. Both Mr. Vale and Mr. McBride have shown examples of such curves. Testing procedures with uniformly higher levels of information over the latent trait continuum will be evaluated most highly.

If, at a given latent trait level, we divide the information value for one testing procedure by the information value for another procedure, we have calculated the "relative efficiency" of the two procedures at that level. A plot of such values as a function of latent trait level is referred to as a relative efficiency curve. A desirable property of such curves is that while a monotone transformation applied to the latent trait continuum will alter the shape of each individual procedure's information curve, the relative efficiency curve will be unchanged by any such transformation (Lord, 1974).

If the expected value of the estimator of a testee's latent trait level is equal to its corresponding parametric value, the estimator is unbiased. If the estimator is biased, it may be informative to plot a bias curve that shows the direction and magnitude of the bias over latent trait levels. Mr. McBride showed examples of such curves earlier.

Another characteristic of a testing procedure that can be used as an evaluative criterion is the standard deviation of the latent trait estimator at each latent trait level. A plot of the values of these standard deviations as a function of latent trait level can be referred to as a standard error of measurement (SEM) curve. If a testing procedure generates unbiased estimates of latent trait scores, then SEM values for the procedure can be obtained by taking the reciprocal square root of the procedure's information values along the latent continuum. The main advantage of the SEM curve over an information curve is that the SEM values are expressed in the same units as the latent continuum while information values are expressed in arbitrary units. For an unbiased estimator, SEM values indicate the typical magnitude of measurement errors at each level of the latent trait.

Testing procedures may also be evaluated in terms of their "robustness". Several varieties of robustness can be considered. First, one can investigate the effects, on different testing procedures, of errors in estimating item parameters. Some procedures rely more heavily than others on the accuracy of item parameter estimates. Since we never have exact parameter values, testing procedures that are robust in the face of errors in the item parameter estimates should be evaluated more highly than procedures which are not. Similarly, testing procedures that are robust in the face of an error regarding the form of the item characteristic curve should be preferred. Finally, some testing procedures, such as Owen's Bayesian method (Owen, 1969), make assumptions regarding the form of the testee population. The researcher should determine, via either analytic derivations or Monte Carlo simulation, the robustness of such procedures when the stated distribution assumptions are in error.

Psycho-social criteria. Since my time is limited, I will not comment at length on evaluation in terms of psycho-social criteria. It will suffice for now to illustrate the kinds of questions that arise when evaluating the psychological and/or social effects of computerized adaptive testing. First, one might ask about the psychological effect on testee motivation of exposing the testee to a series of items adapted to the testee's standing on some ability or personality dimension. A testing method that maintains motivation at optimal levels should be evaluated more highly than methods which do not.

Another basis for comparison of testing procedures is their face validity in applied testing situations. While psychologists have previously encountered the problem of face validity with tests whose content did not "appear relevant" to the criterion behaviors being predicted, adaptive testing presents a new source of potential misunderstanding for the layman. Even when all the items in an item pool appear relevant to the casual observer, the fact that in adaptive testing different people answer different items may cause an observer (say, for example, a testee who has been rejected in his bid for a job) to wonder how different people can be fairly compared when they haven't been exposed to the same test questions.

Cost criteria. Some of the cost criteria that should be considered when evaluating testing procedures are: cost of the delivery system and/or materials needed to implement the procedure, the cost of generating and norming an item pool of the size required by the procedure, susceptibility to clerical errors by the testee during test administration or by office personnel in test scoring, susceptibility to time loss due to delivery system failure or misrouted documentation, and time and personnel costs associated with the administration, scoring, and interpretation of the test. While this list is not exhaustive, it does provide an indication of the variety of cost criteria to be considered.

The Problem of Multiple Criteria

At this point we have reviewed several varieties of evaluative criteria. The problem of how to integrate multiple, and possibly conflicting, criteria into an overall judgement about a given testing procedure remains. I would like to be able to resolve this problem for you, but cannot. The decision as to which criteria are most relevant will necessarily depend upon the particular circumstances in which the procedure is to be applied.

The one generalization that I am inclined to make is that the researcher should not rely exclusively on one criterion index, or class of criteria, to reach his/her conclusions. A balanced evaluation that utilizes both empirical and simulation studies is recommended.

Some Specific Recommendations

I would like to conclude with some specific recommendations regarding the conduct of studies to evaluate adaptive test procedures. First, one must be sure that the subject samples in empirical evaluation studies are representative of the groups one wishes to test ultimately. It is especially important that variability in the latent dimension not be artificially restricted. In both live-testing and simulation studies the sample sizes should be large enough to reduce sampling error to tolerable levels.

In live-testing studies it is essential that the test items have been carefully normed in a large and representative norming group. Bad item parameters can vitiate careful test construction efforts, especially with adaptive tests.

If one wishes to compare two testing procedures, insure that the procedures have access to items of equal quality. This means that item discrimination values in the two tests should be equated as closely as possible. Also, test length should be the same for the two methods. However, some adaptive testing procedures do not have a fixed test length and will require that this recommendation be ignored. In this situation the researcher should attempt to equalize the average test length for the two procedures.

In a study involving retesting on an adaptive test, some testees will receive new items if they alter any of their responses from test to retest. Thus, a comparison with any conventional non-adaptive test will require that an equal number of new items be administered in the conventional test in order to hold memory effects constant across the two test methods.

Finally, keep in mind that if you use correlation coefficients as criterion indices when comparing testing procedures (e.g., alternate form correlations, external validity coefficients, or correlations with latent trait scores) the comparison will be biased in favor of that procedure which measures best in the region of the latent trait continuum from which most of the testees are sampled. If one is interested in obtaining equally precise measurement at all points along the latent trait continuum, regardless of the distribution of the testee population, then the use of correlation coefficients as criterion indices is not recommended.