COMPUTERIZED ADAPTIVE TRAIT MEASUREMENT: PROBLEMS AND PROSPECTS

Proceedings of a Symposium Presented
at the
1975 Annual Convention of the
American Psychological Association

NANCY E. BETZ
JAMES R. MCBRIDE
JAMES B. SYMPSON
C. DAVID VALE

with contributions by

ROBERT L. LINN
R. DARRELL BOCK

edited by

DAVID J. WEISS

RESEARCH REPORT 75-5

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

NOVEMBER 1975

Prepared under contract No. N00014-67-A-0113-0029
NR No. 150-343 with the Personnel and
Training Research Programs, Psychological Sciences Division
Office of Naval Research

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 75-5	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Computerized Adaptive Trait Measurement: Problems and Prospects	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Nancy E. Betz, James R. McBride, James B. Sympson and C. David Vale, with contributions by R. Darrell Bock and Robert L. Linn edited by David J. Weiss	8. CONTRACT OR GRANT NUMBER(s): N00014-67-0113-0029
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: P.E.:61153N PROJ.:RR042-04 T.A.:RR042-04-01 W.U.:NR150-343
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: November 1975
	13. NUMBER OF PAGES: ii+59
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This report is the proceedings of a symposium presented at the Annual Convention of the American Psychological Association, August 30, 1975.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

testing	sequential testing	programmed testing
ability testing	branched testing	response-contingent testing
computerized testing	individualized testing	automated testing
adaptive testing	tailored testing	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This symposium consisted of four papers and the comments of two discussants.

1. C. David Vale. Problem: Strategies of Branching through an Item Pool.

This paper describes a variety of strategies for adapting tests to the trait level of each individual on the basis of the testee's responses to previously administered items. Based on data from computer simulations, the various strategies are compared in terms of

the levels and shapes of information curves they provide under one particular set of conditions. Limitations of the data presented are discussed.

2. James R. McBride. Problem: Scoring Adaptive Tests.
Several approaches to scoring adaptive tests are described. Inapplicability of traditional number correct scores in adaptive testing, where different individuals answer different items, is discussed. The essentials of latent trait theory are summarized, and two scoring methods usable with that approach are explicated. These scoring methods--maximum likelihood scoring and Bayesian scoring--are compared using simulation data, on criteria of information, bias, and regression on ability. Limitations of these scoring methods are discussed.

3. James B. Sympson. Problem: Evaluating the Results of Adaptive Testing.
Six component elements of a testing procedure are described; it is suggested that proper evaluation of a testing procedure should be based on consideration of these elements as separable components. Classes of criteria for evaluating a testing procedure are differentiated into validating criteria, theoretical criteria, psychosocial criteria, and cost criteria. Within each of these categories, the various criteria are discussed and contrasted. Suggestions are made as to the appropriate applications of each of these criteria. The problem of using multiple criteria is briefly discussed and it is suggested that live-testing and simulation research be systematically combined. A number of specific recommendations are made concerning problems of evaluating the results of adaptive testing.

4. Nancy E. Betz. Prospects: New types of Information and Psychological Implications.
Several types of new information available from computerized adaptive measurement are described. These include individualized error of measurement, response consistency, improved response modes, response latencies, and new kinds of tests. Data from live computerized testing are presented showing that response consistency moderates test-retest reliability. The potential psychological advantages of computerized testing are discussed. Data are presented from two studies demonstrating the facilitating effect of immediate knowledge of results after each test item on ability test performance.

Comments by the discussants, Robert L. Linn of the University of Illinois and R. Darrell Bock of the University of Chicago, include a discussion of some of the limitations of the research presented, some differing interpretations, and suggestions for future research in adaptive testing.

Computerized Adaptive Trait Measurement: Problems and Prospects

Edited by David J. Weiss

INTRODUCTION

David J. Weiss . . . . . University of Minnesota

The research program which generated the ideas and findings reported in the four main papers in this symposium has been supported since early 1972 by the Personnel and Training Research Programs, Office of Naval Research, Washington, D.C. The continuing support of that office and the encouragement and guidance received from its Director, Dr. Marshall J. Farr, and its Assistant Director, Dr. Joseph L. Young have made it possible to develop a sustained research effort designed to answer very basic questions about the psychometric and practical utility of computerized adaptive testing. Prior to the ONR support, this research effort was supported for two years by grants from the General Research Fund of the Graduate School at the University of Minnesota. Special thanks are due to our project programmer, Louis J. DeWitt, who has worked on this research since 1970. Without his persistence in re-programming our testing system for four different computers, during the last five years, much of our research would have been impossible.

The focus of our research to date on computerized adaptive measurement has been in the area of ability testing. However, our methods and findings should be applicable, in general, to the measurement of any homogeneous trait. We have also based much of our research on the latent trait model. But latent trait theory is only one way of conceptualizing trait measurement, although a very useful one. There are alternative ways of approaching trait measurement which might be equally useful in adaptive testing, and which are used in the measurement of other psychological variables. Thus, the methods of adaptive testing should prove useful in the measurement of personality variables, interest traits, values, and in other aspects of human functioning which can be conceptualized as homogeneous traits.

We have approached research in adaptive testing in two complementary ways--live-testing and computer simulations. In the last three years we have administered over 7,500 tests to real testees by interactive computers. In addition, based on the latent trait model, we have simulated the responses of hundreds of thousands of testees on a wide variety of testing strategies. These simulations augment the findings from live-testing studies in very important ways, and provide further hypotheses for research using live computerized testing.

The four papers that follow will describe three major problems we have faced in our research on adaptive testing, and some of the prospects for improving testing that derive from computerized adaptive testing. The first problem that we faced is *how* to adapt a test to individual differences in trait level, during the process of testing. Mr. Vale, in the first paper, will describe what an adaptive test is, how various strategies of adaptive testing function, and will present some data comparing adaptive testing strategies.

The second major problem we faced was that of scoring adaptive tests. In adaptive testing, different testees answer different test items. Thus, since each test is unique to the individual, traditional scoring methods are inappropriate. Mr. McBride, in the second paper, will describe some of our research on scoring methods, and present some data comparing the characteristics of several scoring methods.

For over fifty years the paper and pencil test has been the predominant mode of testing. But now that computerized adaptive testing is a possibility, we have an alternative. Thus, there is a need, that was not evident in former years, to evaluate the characteristics of different testing strategies. Mr. Sympson, in the third paper, will discuss some of the problems we have faced in comparing different testing procedures, and suggest a systematic approach to the problem.

Computerized adaptive testing is not all problems, however. In addition to the problems we have faced in our research, we have become aware of the potential of this mode of testing for improving psychological measurement in other ways, beyond the purely psychometric benefits resulting from the adaptive process. In the fourth paper, Ms. Betz will describe some of these new prospects, and present data derived from live testing concerning some valuable new kinds of information resulting from computerized testing, and data on the potential psychological benefits of this mode of testing.

The four papers will describe only some of the problems we have faced in our research program, and some of the potential advantages of computerized adaptive trait measurement. The discussants will, we hope, help us to find some solutions to the myriad problems that we face in this new field of research, suggest other ways of approaching these problems, or suggest other kinds of research which might extend our findings.

Our first discussant, Robert L. Linn of the University of Illinois, was a pioneer researcher in adaptive testing. His research in the mid-1960's pioneered one important methodological approach to research in adaptive testing, and produced initial findings supporting the utility of adaptive or sequential testing. Our second discussant, R. Darrell Bock of the University of Chicago, is an internationally known authority on latent trait theory, the theoretical approach used in most adaptive testing research. His important theoretical and methodological developments in this area have permitted adaptive testing research to move forward efficiently. We appreciate Dr. Linn's and Dr. Bock's contributions to this symposium, and their helpful comments.

PROBLEM:

STRATEGIES OF BRANCHING THROUGH AN ITEM POOL

C. David Vale
University of Minnesota

The problem I am addressing has been the focus of much of the research in adaptive or tailored testing and provides, in fact, the major motivation for administering tests adaptively. The problem is: given a large pool of test items and a constraint to administer a relatively small number of them, what is the best way of selecting that small number of items? In this presentation, I am going to show some strategies that have been used for selecting items in the framework of their evolution from the simple conventional test to complex adaptive or tailored testing models. To clarify the distinctions between some of the models we will follow the progress of a hypothetical, low ability subject, Dennis Dull, through a test administered under each strategy and note how his items are selected. We will further examine differences between strategies in terms of the amount of information about ability which each strategy provides.

Assumptions

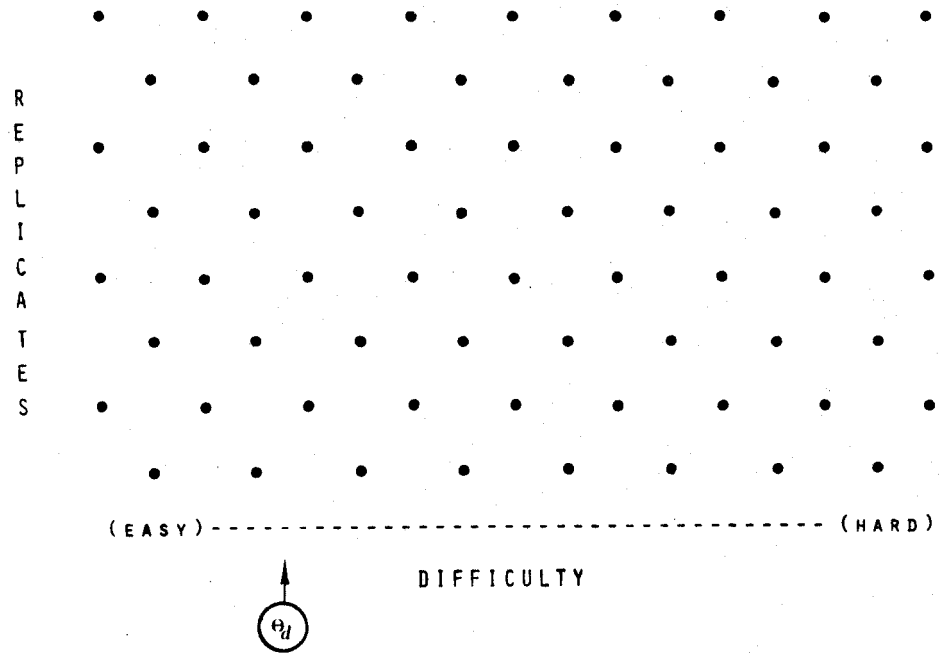
In order to make possible the analyses done for this presentation, some simplifying assumptions were made. First, it was assumed that a large pool of equally good items (i.e., items with equivalent discriminating power) was available to choose from. Second, it was assumed that these were free-response items and, hence, guessing was not possible. Third, it was assumed that all items were scored by a common technique, in this case, a Bayesian scoring procedure. Finally, to make comparisons between some strategies meaningful, it was assumed that a prior estimate of ability, correlating .5 with ability, was available.

Figure 1 shows schematically the item pool that will be used for testing with the various strategies. On the horizontal dimension are seventeen columns, each containing four items, ranging from very easy items at the left to very difficult items at the right. The vertical dimension represents replications of items at each difficulty level; all items in a column are equally difficult.

I will illustrate the various item selection strategies using eight items from this pool of 68. While an eight-item test is convenient for illustration, eight items are too few to allow some of the adaptive strategies to function well. Therefore, for evaluation of the strategies a 24-item test was used. Items for the 24-item tests were chosen in a manner analogous to the way items were chosen for the illustrated eight-item test. The results I'll present are from computer simulations (see Appendix for details of the simulation method; numerical results are in Appendix Table A-1).

Figure 1

Schematic Representation of the Item Pool Showing Dennis' Ability (Θ_d) in Relation to the Item Difficulties



Testing Strategies

Rectangular conventional test. One way to compose a test is to select a fixed set of items having a wide range of difficulties. Figure 2 shows such a rectangular conventional test. In this case, eight items equally spaced on the difficulty continuum were chosen from alternate columns ranging from the next to easiest to the next to most difficult columns. Dennis Dull, our low ability subject, produced the response record shown in Figure 2 with those items he answered correctly marked by a plus (+) and those he answered incorrectly marked by a minus (-). The items in this test could have been administered in any order but for clarity of presentation, we started at the left and worked toward the right.

The first item Dennis encountered was beneath his ability level and, knowing the answer, he responded correctly. The second item was a bit more difficult for Dennis but he still answered it correctly. The third item, being a bit above his ability level, was too difficult for Dennis and he answered it incorrectly. Similarly, the fourth through eighth items were even more difficult and he answered all of them incorrectly.

Figure 2

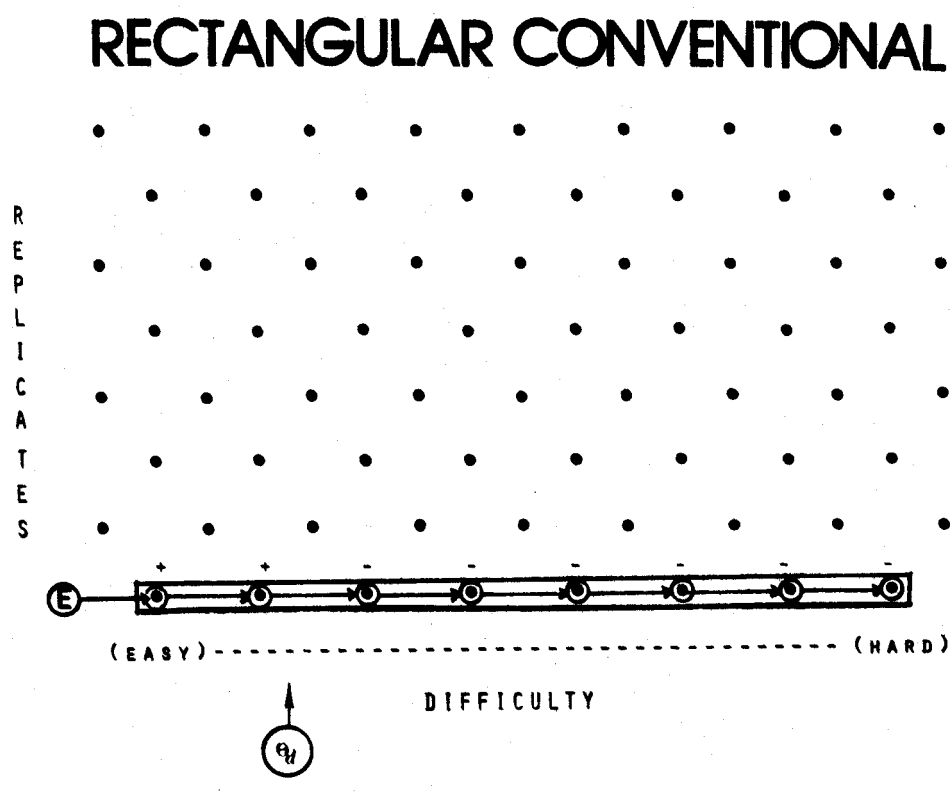


Figure 3

Information Curve for the Rectangular Conventional Test

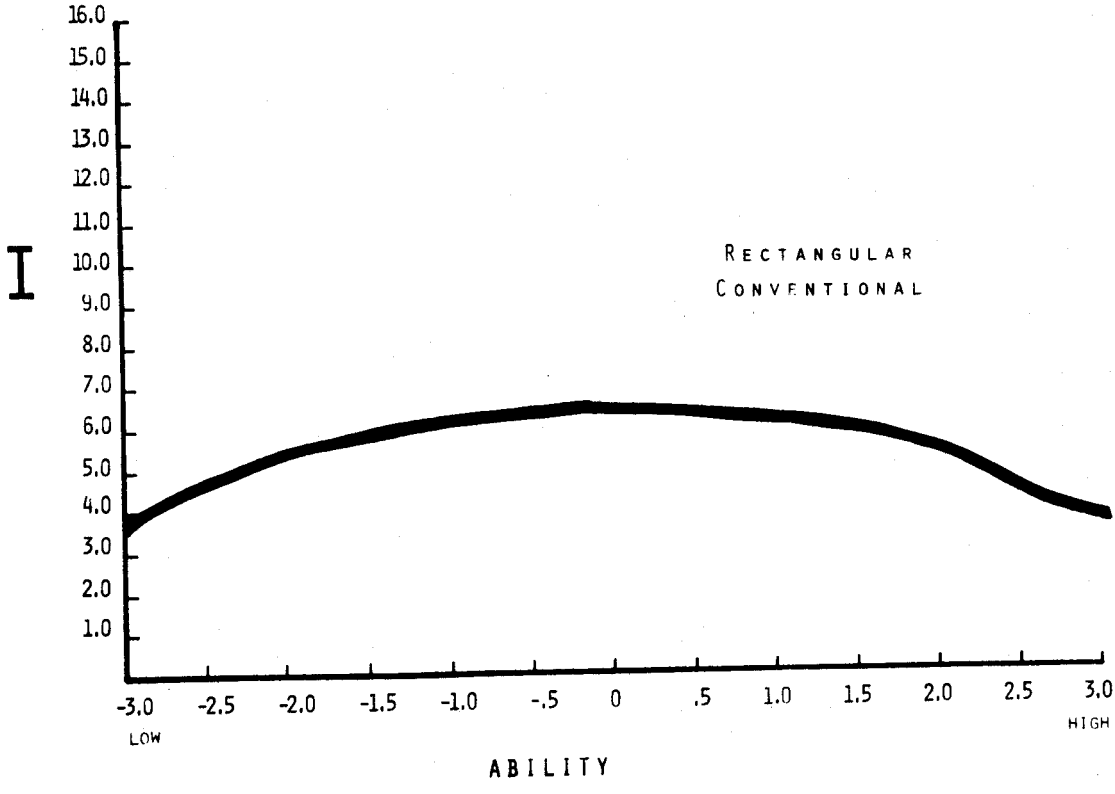
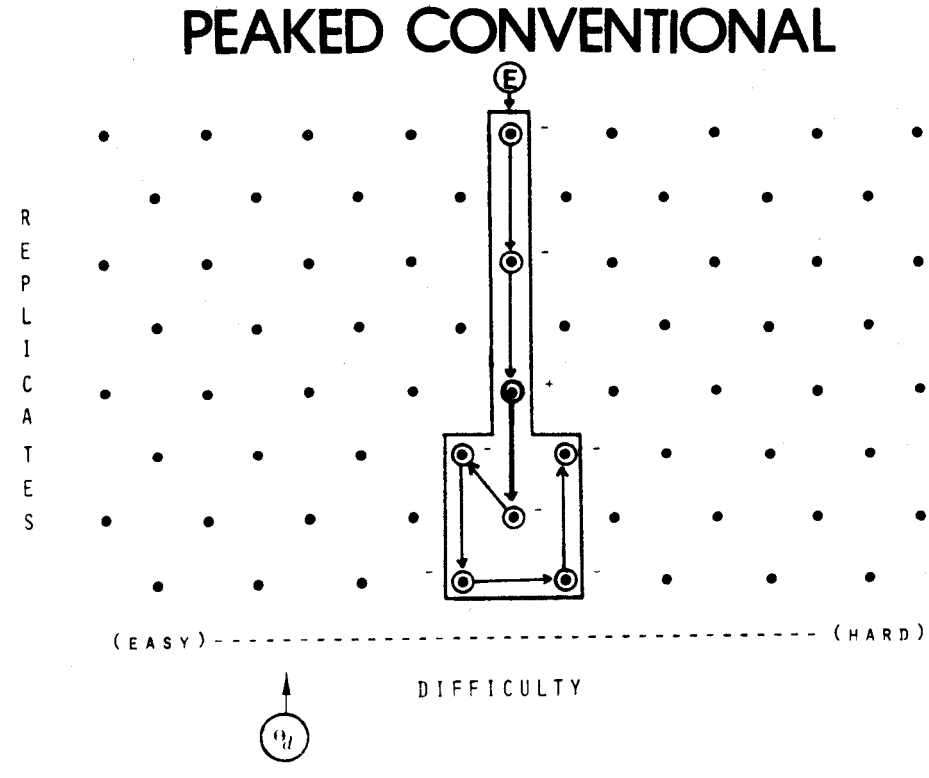


Figure 3 shows an information curve produced by the rectangular conventional test. Information can be thought of as related to the precision of measurement produced by a test at a given level of ability, or as how well a test can discriminate between two contiguous ability levels (see Lord, 1970, for a discussion of information curves). A good test produces an information curve that is high (i.e., provides precise measurement) and is flat (i.e., provides this high level of precision for all testees at all ability levels). Although not apparent from Figure 3, it will become obvious from comparisons with later information curves that the rectangular conventional test produces an information curve that is fairly flat but somewhat low. It can be seen, however, that even this information curve tapers off at the extremes indicating poorer measurement for testees where ability level is distant from the mean.

Peaked conventional test. Instead of choosing items with a wide range of difficulty, we could instead choose items peaked at the center of the ability range and administer them to all testees. Figure 4 shows such peaked conventional test. The four items from the median difficulty column and two from each of the adjacent columns were chosen for this test. Again, these items could have been administered in any order but we will begin at the top for clarity.

Figure 4

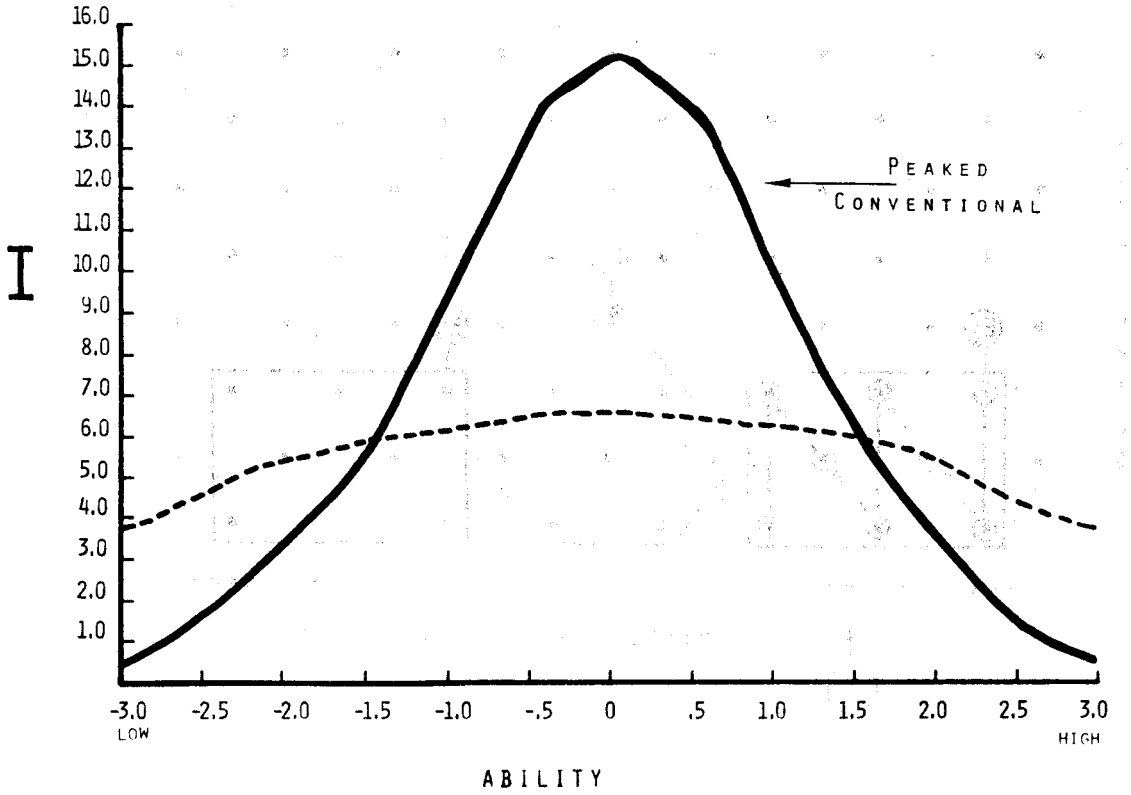


These items were intended for average ability testees and were all too difficult for Dennis. He answered incorrectly the first item, the second item, and most of the rest of the items. In fact, the only item he answered correctly asked for the definition of "Oedipal", a term he had picked up from his analyst.

The information curve for the peaked conventional test (Figure 5) shows graphically what Dennis felt as he took the test; the peaked conventional test provides good measurement for some testees but very poor measurement for others. As Figure 5 shows, the peaked conventional test produces precise measurement for individuals with abilities in the middle range but little information for extreme ability subjects. The peaked conventional test provides more information about ability than does the rectangular conventional test within the range of ±1.5 standard deviations of the ability range for which it was peaked but less outside of this range.

Figure 5

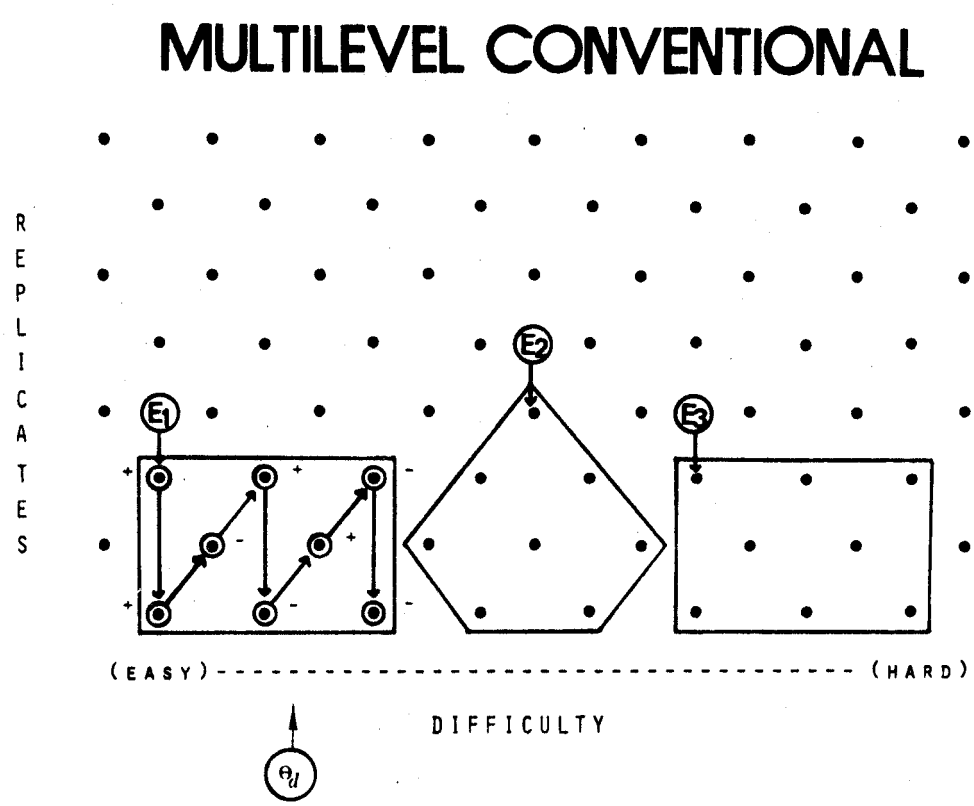
Information Curve for the Peaked Conventional Test



It seems that with a fixed set of items (i.e., a conventional test) we can please some of the people all of the time or all of the people some of the time, but we can't please all of the people all of the time. If, however, we could figure out a way to move a peaked ability test to the ability level of each person being tested, we could please all of the people all of the time and provide a high level of information at all ability levels. If a testee's ability were known before testing, we would construct a test made up of those items with difficulties closest to his ability level (i.e., items which he/she would be expected to answer correctly 50% of the time). But if we knew his ability beforehand, we would have no reason to administer the test at all.

Multi-level conventional tests. In practice we have, at best, a fallible prior estimate of the testee's ability level and may want to administer items more or less rectangularly distributed in a narrow range around that estimated ability level. Some achievement tests use a prior ability estimate, such as grade in school, to determine which section of a test a testee should take. Figure 6 shows such a test. Knowing that Dennis ranked at the 27th percentile in his grade school graduating class, if this were a

Figure 6

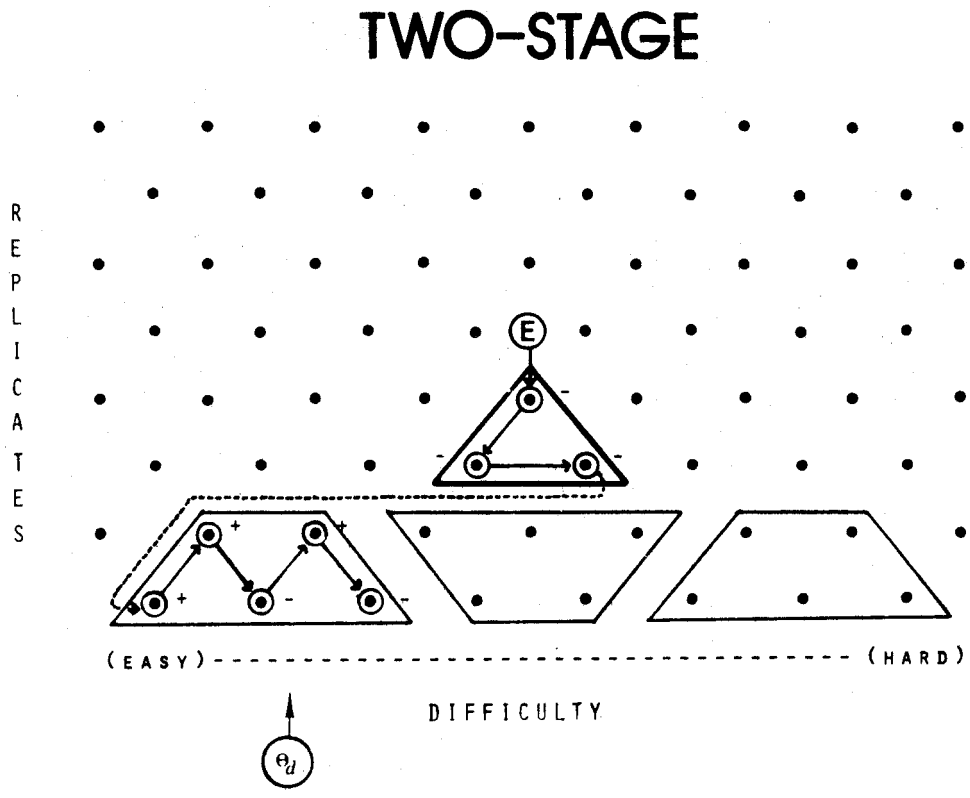


high school freshman achievement test, we might use this prior information to start Dennis at the easiest entry point (E_1). Or, if we had a testee with all A's in grade school, we might start him at the high entry point. Given a prior ability estimate, therefore, it is possible to adapt the test to the individual within the framework of a conventional test. But if prior information is not available, we have to use a test that tailors item difficulty in its absence. One possible strategy for doing this is the two-stage testing strategy (Angoff & Huddleston, 1958; Betz & Weiss, 1973, 1974) which is like the previous test but generates its own prior ability estimate.

Two-stage tests. In a two-stage test, a testee is first administered a short routing test and, on the basis of his score on that test, is branched to a measurement test of more appropriate difficulty. Figure 7 shows a two-stage

test. A testee takes a three-item routing test and one of three five-item measurement tests. Dennis answered all three of the routing test items incorrectly as they were too difficult for him. Since this suggested that his

Figure 7



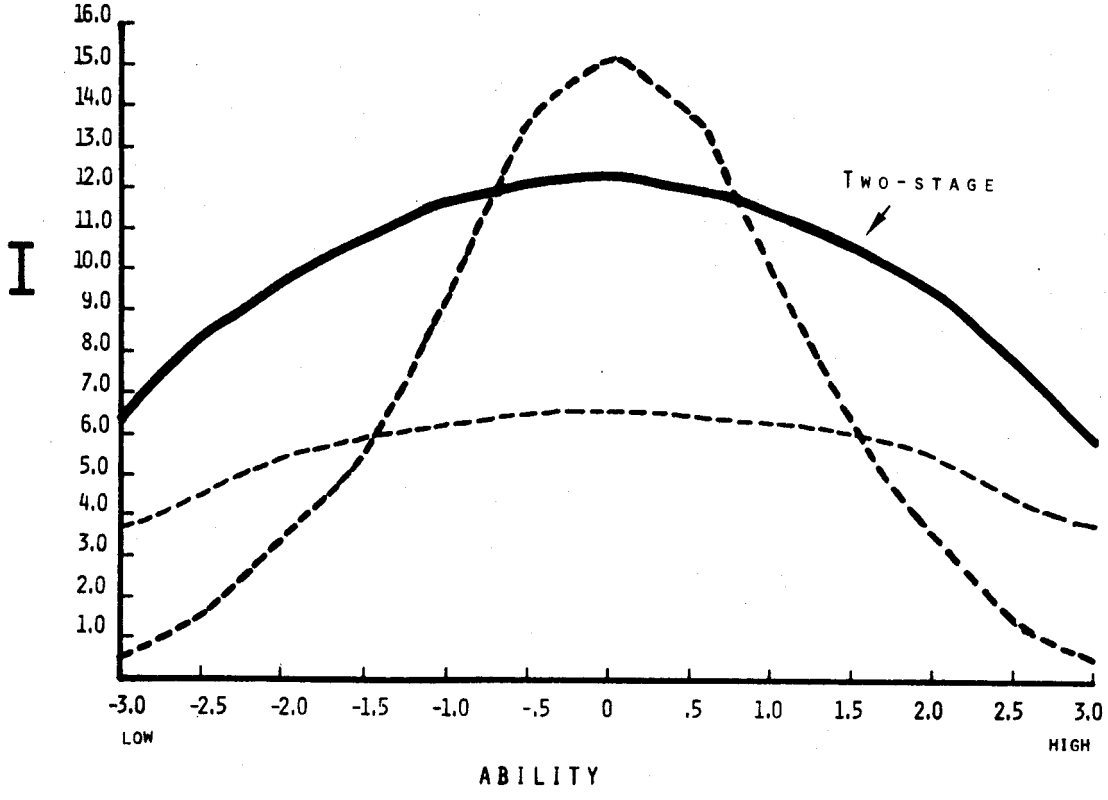
ability was low, he was branched to the easiest measurement test where he answered three out of the five items correctly.

As Figure 8 shows, this two-stage test yields an information curve that is at all points higher than that of the rectangular conventional test and higher than the information curve of the peaked conventional test except in the center. Thus, this two-stage test provides more precise measurement than the rectangular conventional test at all ability levels, and more precise measurement than the peaked conventional test at most ability levels.

One problem with the two-stage testing strategy is that if a testee's ability is between the difficulties of two adjacent measurement tests, there is no measurement test of appropriate difficulty. A solution to this problem

Figure 8

Information Curve for the Two-Stage Test



is available in the form of the continuous second stage two-stage test (Sympson, 1975), a variant of the previous two-stage test, shown in Figure 9. As in the standard two-stage test, the testee is first administered the three-item routing test. Then, on the basis of the score on that test, he is branched to a five-item measurement test. But instead of using one of three pre-structured measurement tests, a measurement test consisting of the most appropriate item and two adjacent items on each side is individually composed for the testee. Given our restricted circumstances, the information curve of the continuous two-stage test would be very similar to that of the standard two-stage test and will not be shown here.

Another problem inherent in the two-stage procedure is that of misrouting. The measurement test decision is based on a short and fallible routing test and thus may be incorrect. For example, had the word "Oedipal" occurred in the two-stage routing test, Dennis would have answered one out of the three items correctly and might have been branched to the middle measurement test containing items that were too difficult for him.

Flexilevel test. There are two solutions to the misrouting problem: One is to route more; the other is to route less (i.e., not at all). An example of the latter strategy is the flexilevel test (Lord, 1971) shown in Figure 10. For this test the potential item set is the same as the potential measurement test item set of the continuous two-stage test. But rather than taking a routing test, each testee starts with the median difficulty item of the

Figure 9

CONTINUOUS TWO-STAGE

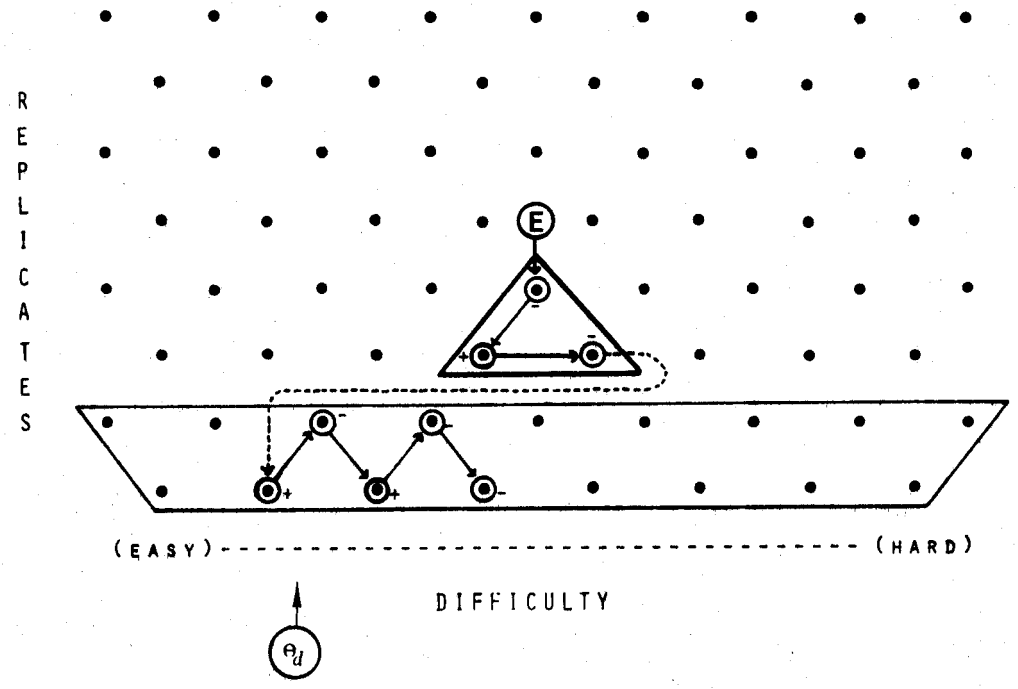
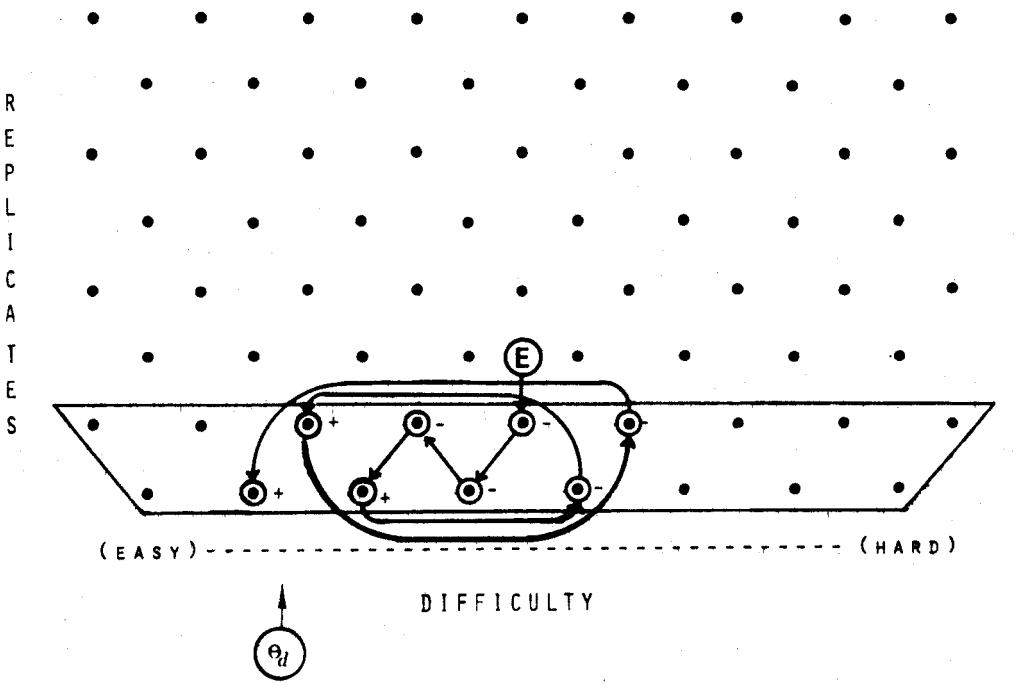


Figure 10

FLEXILEVEL

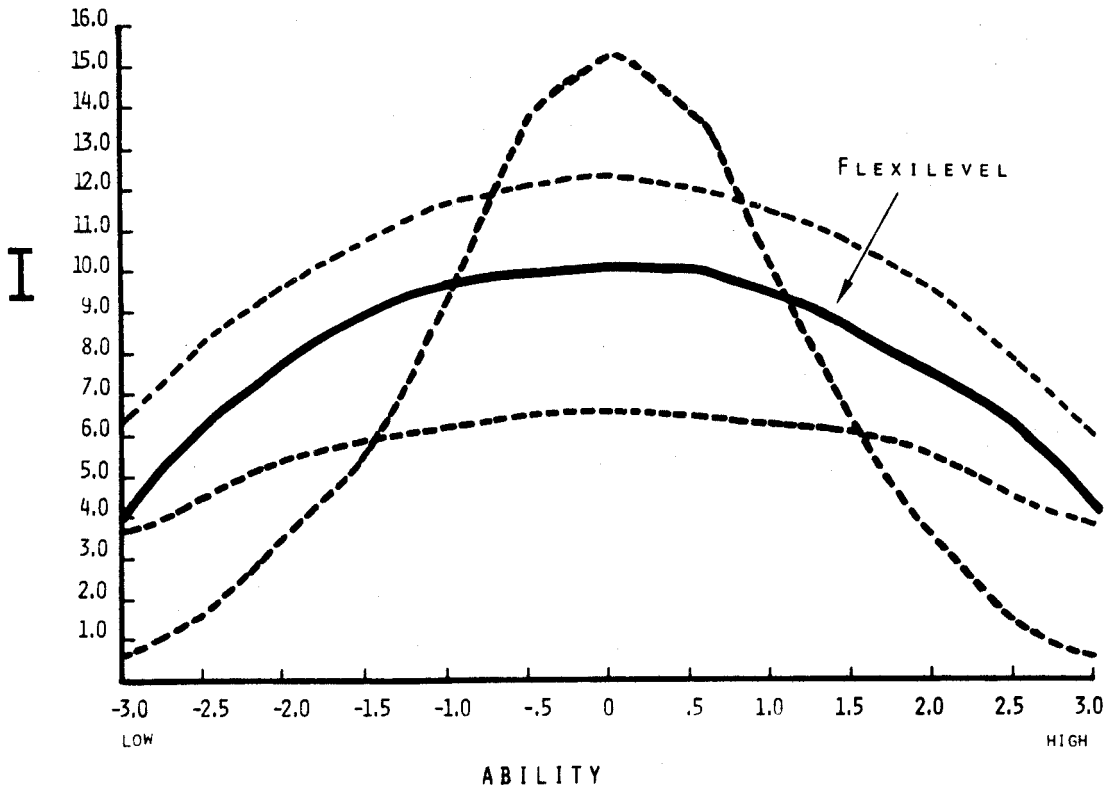


item set, and following each correct response is branched to the next more difficult unadministered item. Following an incorrect response, he is branched to the next less difficult unadministered item.

In Dennis' case, he answered incorrectly the first three items and was branched appropriately downward until he reached the third item below the median, an item slightly above his ability level. Knowing the answer, he answered that item correctly and was branched to the first item above the median which he answered incorrectly. He was branched to the fourth item below the median and continued oscillating between easy and difficult items until he had answered eight items.

Figure 11

Information Curve for the Flexilevel Test



The information curve for the flexilevel test is shown in Figure 11. Although the flexilevel test solves the problem of misrouting, the information it provides is always less than that provided by the two-stage test.

Three-stage test. Figure 12 shows an example of the other solution to the problem of misrouting, the three-stage test (sometimes referred to as the double-routing two-stage). In this strategy, an individual takes one routing test which routes him to a second routing test which routes him to a measurement test. Errors resulting from the first routing can be ameliorated by the second routing.

Figure 12

THREE-STAGE

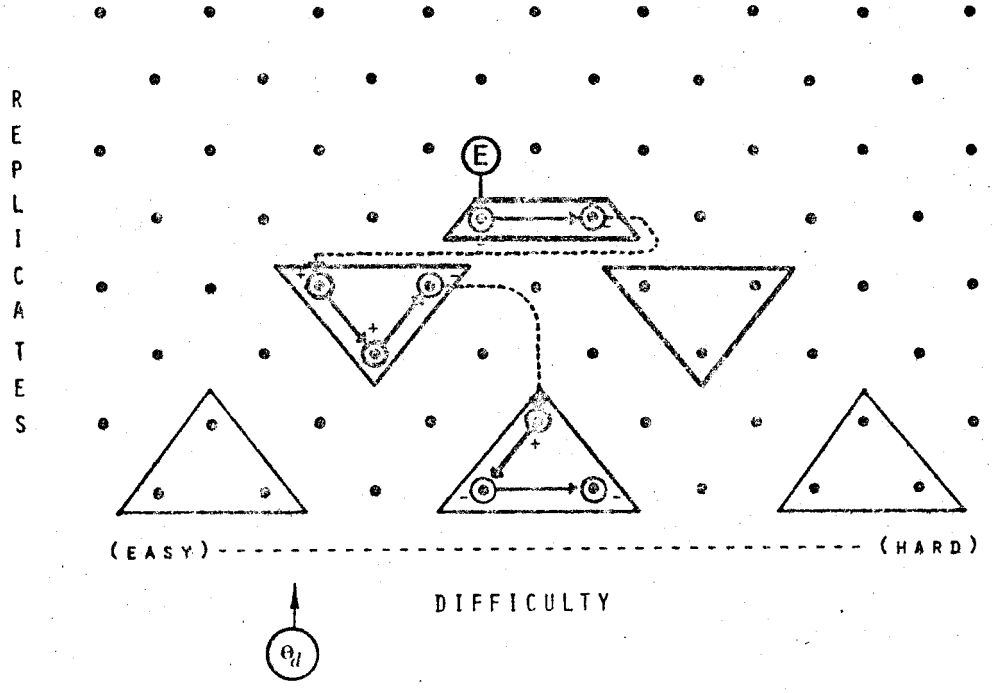
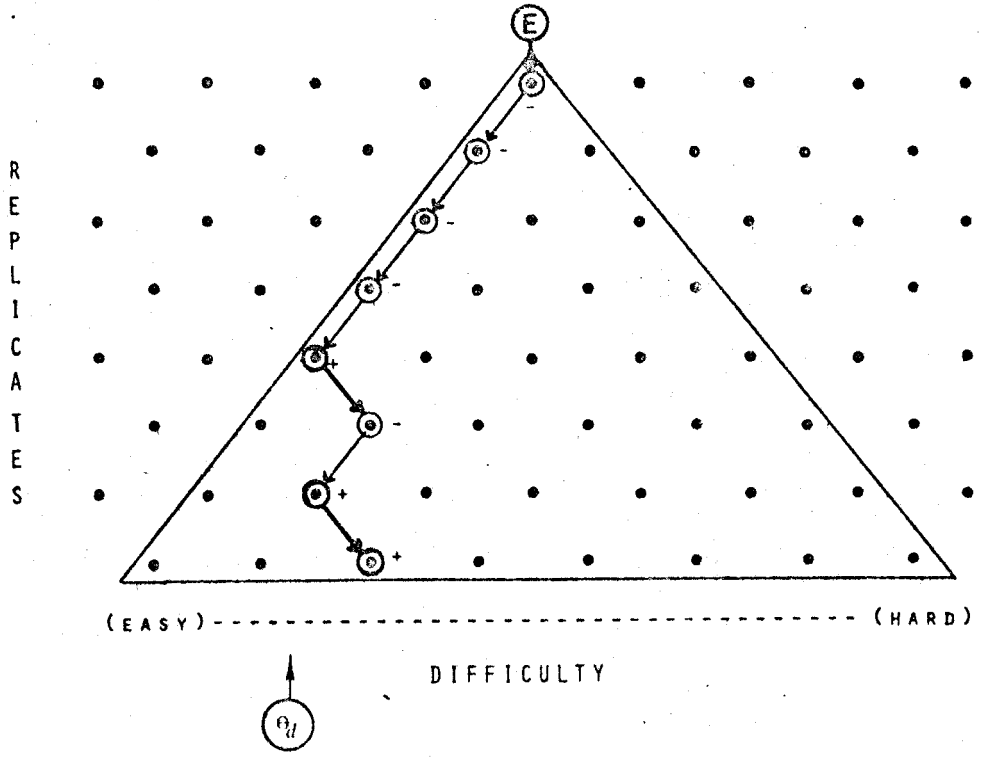


Figure 13

PYRAMIDAL

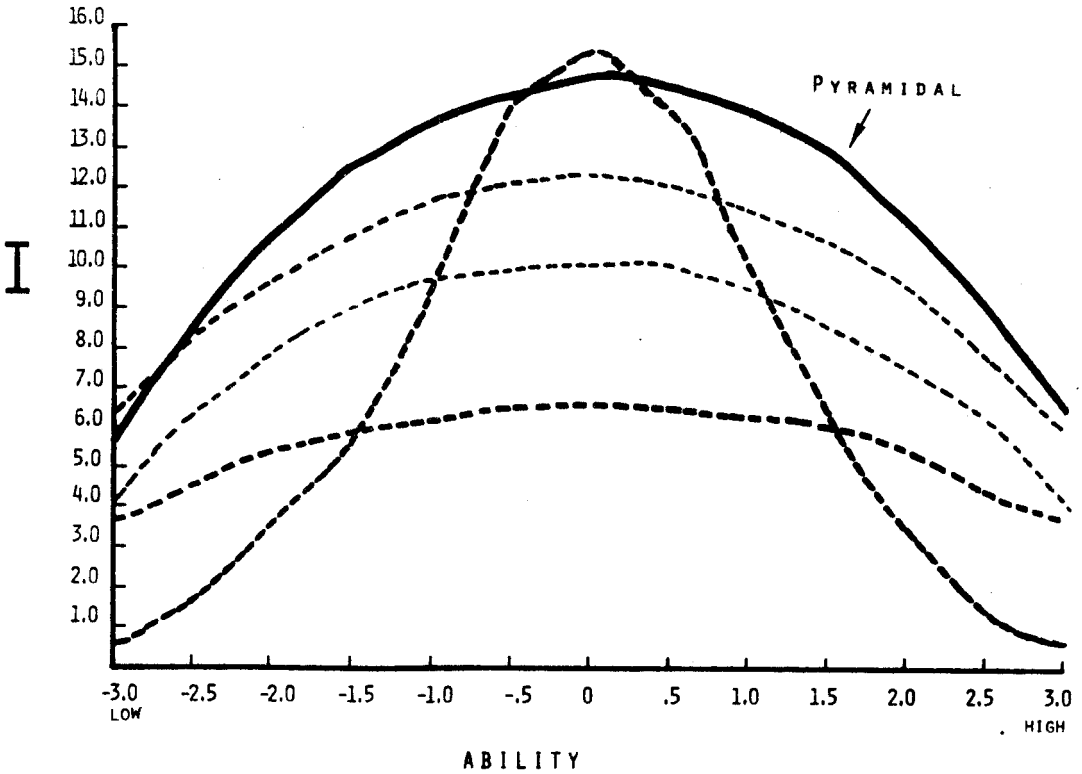


Pyramidal test. Carrying the idea of multiple routing to its logical extreme (i.e., using one item per stage) results, in this case, in the eight-stage test or, in the general case, the pyramidal test (Krathwohl & Huyser, 1956; Larkin & Weiss, 1974, 1975). As shown by Figure 13, in this strategy a testee starts with a median difficulty item and is branched after each item to a less difficult item following an incorrect response or to a more difficult item following a correct response.

The information curve for this test, shown in Figure 14, shows it to provide more information than any of the strategies discussed thus far except in the middle ability range where it is slightly surpassed by the peaked conventional test. It should be noted, however, that the information curve is far from flat. Less than half of the amount of information provided at the middle range of ability is provided at the extremes of this information curve, three standard deviations from the mean.

Figure 14

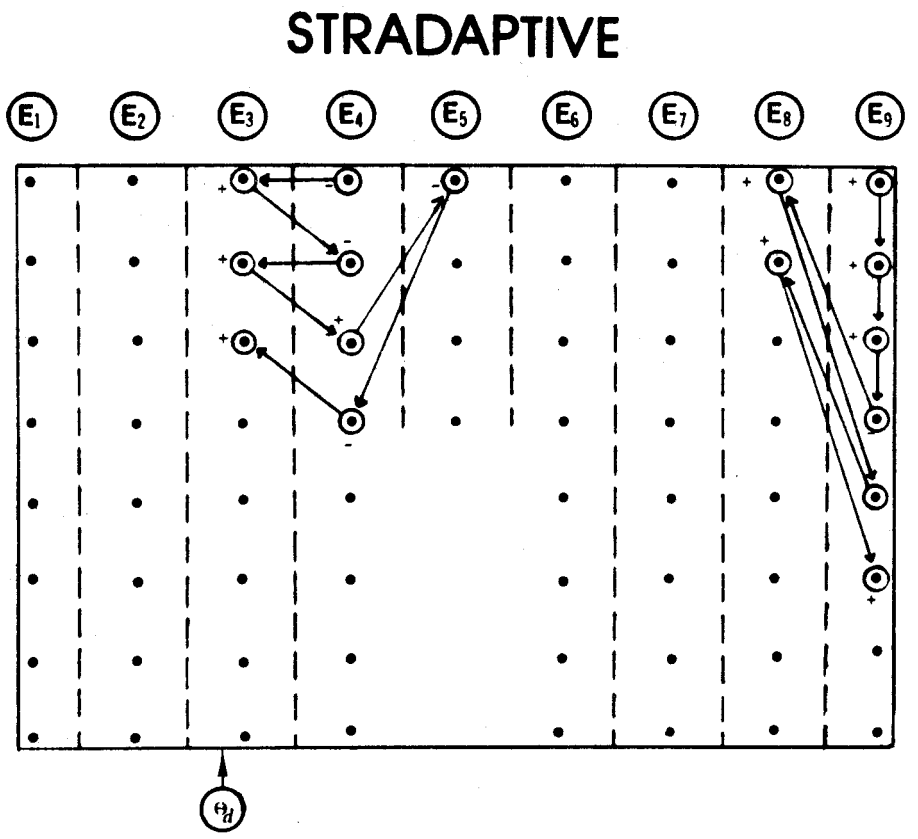
Information Curve for the Pyramidal Test



Stratified-adaptive test. The previously discussed adaptive tests have been developed for the situation in which prior ability information was not available and are not capable of using it when it is available. Now that we have reached the top of the pyramid, so to speak, we can make use of prior information by extending the pyramidal structure to allow entry at several points.

A direct extension is unable to handle branching for some extreme ability testees, however, so a modified extension of the pyramidal structure is used by the stratified-adaptive (stradaptive) testing strategy (Weiss, 1973) shown in Figure 15. Two changes beyond a direct extension are observed: 1) items are grouped into strata consisting of items of possibly slightly different difficulties; and 2) branching is between strata, with the item selected being the first unadministered item in a stratum.

Figure 15



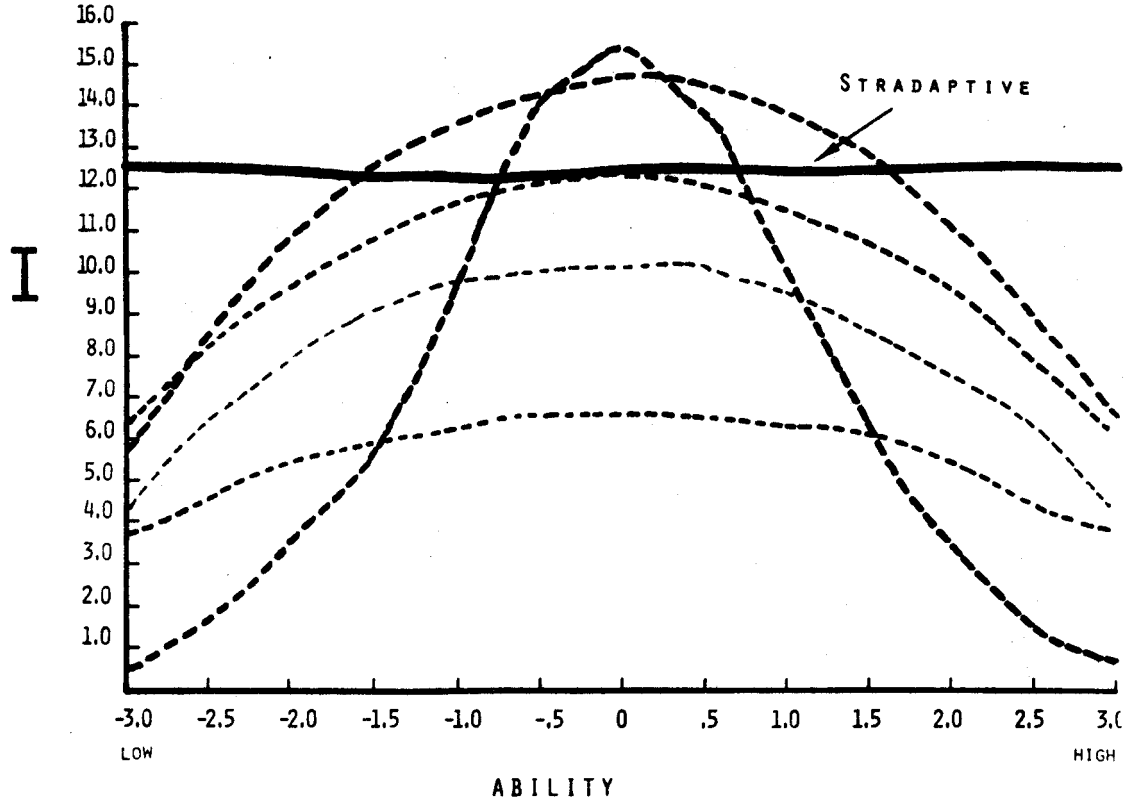
Dennis started at the fourth entry point. He did not correctly answer the first item in stratum four, was branched to the first item in stratum three, answered this item correctly, and alternated between these two strata until his fifth item. He correctly answered the fifth item, which was in the fourth stratum, and was branched to the first item in the fifth stratum. He did not know the correct answer to either this or the next item, and finished with his eighth item in the third stratum.

Branching to the first item in a stratum is of little value in a situation where all items are equally discriminating, but is useful when using a real item pool because all items will not be equally discriminating. This feature allows the most discriminating items to be put where they have the highest probability of being administered; at the top of the stratum. The information curve for the stradaptive test, shown in Figure 16, is almost

flat indicating that the stradaptive test provides very equiprecise measurement. Its level is surpassed by several other strategies in the center, however.

Figure 16

Information Curve for the Stradaptive Test



The previous adaptive strategies are all among the fixed branching strategies. The branching has been a function solely of the testee's performance at the immediately preceeding stage. The variable branching procedures calculate an ability estimate after each item and select as the next item the item best suited for an individual of that ability.

A Bayesian strategy. An example of the variable branching procedures is the Bayesian strategy (Owen, 1969), which is illustrated in Figure 17. On the basis of a prior ability estimate, which may be simply the mean ability of the population of testees, a first item is selected. On the basis of the response to that item and a prior ability distribution, which may consist simply of population parameters, a score is calculated and on the basis of that score, another item is selected. This procedure is repeated, each time selecting the one item in the pool which is closest in difficulty to the last ability estimate.

Figure 17

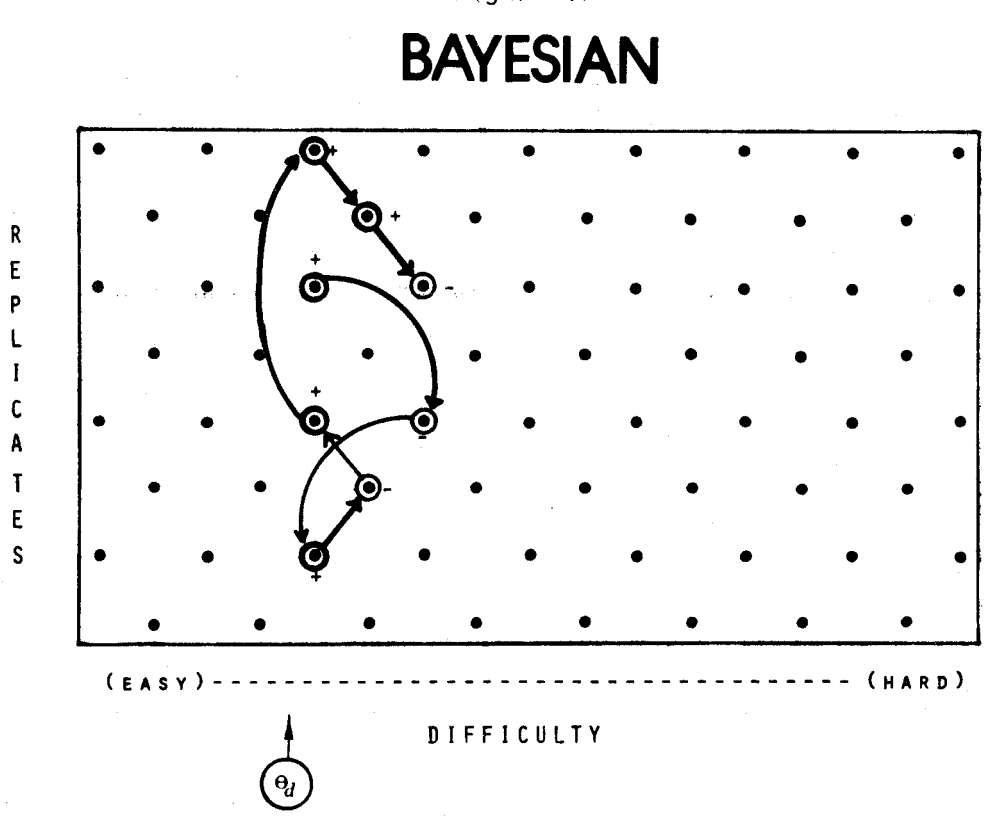
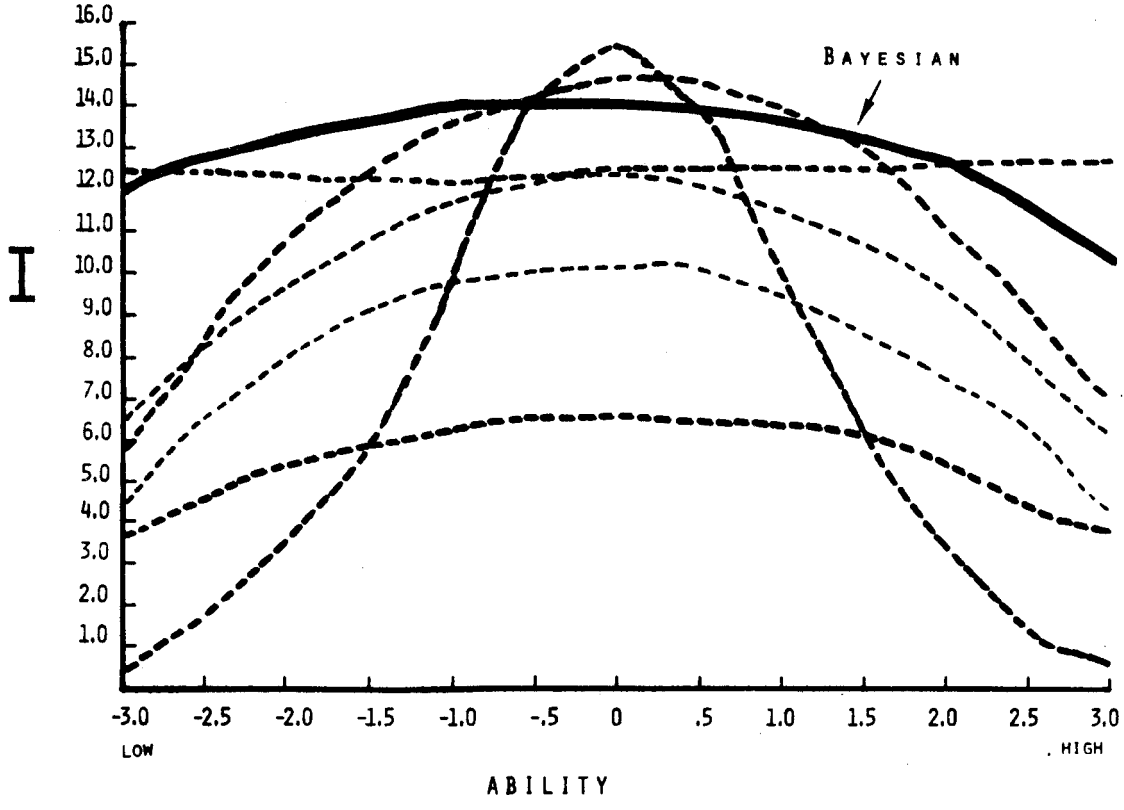


Figure 18

Information Curve for the Bayesian Test



The Bayesian test's information curve is shown in Figure 18. It is slightly higher than the stradaptive test's information curve and nearly as flat, although it drops more in the tails. The peaked conventional test and the pyramidal test still provide more information in the center of the ability distribution.

Limitations of the Results

In this presentation, I've attempted to give an idea why adaptive testing is needed, what some strategies of implementing adaptive testing are, and how these strategies compare in terms of the information they provide. If evaluation of adaptive testing were as simple as this presentation, however, our research would be unnecessary. This evaluation was very limited in a number of ways:

1. The information curves were calculated using a response model which may not accurately portray response tendencies of real subjects.

2. An unrealistic item pool containing equidiscriminating items was used. This would never be found in the real world.

3. Numbers of items per stage and peakedness of subtests were chosen arbitrarily and may not be optimal.

4. A common scoring technique was used which may not be optimal for all strategies. Mr. McBride will outline some of the alternative scoring procedures.

5. As you will see in Mr. Sympson's presentation, information curves are not the only way to evaluate the goodness of a testing strategy.

6. Strategies and scoring methods determined to be "best" in some situations may not be best in others.

The questions involved in adaptive testing are multifaceted and complex. The purpose of research in adaptive testing is to answer the questions necessary to decide when and how to use which kind of adaptive testing strategy. The illustrations provided here were designed simply to introduce the field and are, at best, limited in their generalizability.

PROBLEM:

SCORING ADAPTIVE TESTS

James R. McBride
University of Minnesota

The purpose of administering mental tests to people is usually to compare each person with some criterion, or to compare each person with others with regard to test scores. On a conventional test, where all examinees take a common set of test items, the test score is typically the number of items answered correctly, or some transformation of the number correct.

When all the items of a test are equivalent, having equal difficulty and equal intercorrelations, the number-correct score is a sufficient statistic for estimating ability level (Lord, 1953). It contains all the information in the pattern or vector of individual item scores. When the items in a test are not all equivalent, however, the simple number-correct score fails to convey all the information in the pattern of item responses. Instead a weighted linear composite of the item scores is needed (Solomon, 1961), where the weights are proportional to the item discriminating power. When guessing is a factor, the problem becomes even more complex.

In general, the number-correct score uses less than all of the information available in the test item responses. Further, the number-correct score provides only one more score category than the number of items in the test. For example, only twenty-one unique scores are available from a 20-item test. The shorter the test, the smaller the number of discriminations among persons which can be made. Still a third shortcoming of number-correct scores is the lack of comparability of scores from different tests of the same trait or construct, unless the tests are strictly equivalent.

In scoring adaptive tests, the comparability problem becomes even more pronounced. An adaptive testing strategy combines both an item selection procedure and a scoring method. Different persons in effect take different tests, and the different tests are intentionally non-equivalent across individuals. The sets of adaptive test items administered to any two persons may differ in difficulty and in item discrimination, as Mr. Vale has illustrated. Some adaptive testing strategies, such as the stradaptive and Bayesian ones, permit test length to vary as well. The number-correct score, and the weighted linear combination of the item scores, are both inadequate for scoring adaptive tests, except for certain special cases.

What is needed in adaptive testing is a general scoring method which will take account of the pattern of item responses, and of the difficulty and discrimination of the items administered, and which will yield scores which are directly comparable despite non-equivalence of the item sets. The scaling methods made possible by item characteristic curve properties in latent trait theory provide a class of solutions to the problem of adaptive test scoring. I will mention two of these methods. But first, a hasty introduction to latent trait theory.

Latent Trait Theory

For certain kinds of psychological variables, such as those measured by most ability tests, the construct or trait being measured is monotonically related to test score. At the dichotomous item level, this is tantamount to saying that the probability of a correct response increases with trait level. Trait level is assumed to vary continuously, but the metric for describing it is arbitrary.

An item characteristic curve describes the probability of a correct response $P(u_g=1)$ to a specific test item g as a function of level on the underlying trait.[9] The curve can be described as a function in several parameters, usually trait level (θ), item difficulty (b) and item discriminating power (a). Thus for a single item g, the probability of correct response, $P_g(\theta)$, can be expressed in terms of the three parameters:

$$P(u_g=1 \mid \theta) \equiv P_g(\theta) = F_g(\theta, b_g, a_g) \qquad [1]$$

Now if the forms and parameters of the item characteristic functions are known and if the convenient property of local independence can be assumed (or derived from other assumptions), then the probability of a pattern or string of item scores can be expressed as a compound function of the item characteristic functions. I.e.,

$$P(u_1, u_2, \ldots u_k) = \prod_{g=1}^{k} P_g(\theta)^{u_g} \left[1 - P_g(\theta)\right]^{1-u_g} \qquad [2]$$

Maximum likelihood scoring. For test scoring purposes, of course, we are not interested in estimating the probability of a pattern of item scores, but in estimating the trait level parameter θ from the item scores. This presumes that the item parameters b_g, a_g have been determined (or estimated) already, so let us say that they have been. Then for any pattern or vector of dichotomous item scores there is a likelihood function such as Equation 2. We may use as our trait-level estimate--or test score--the value of θ at which the likelihood function is maximal. That is, given a pattern of item scores, and the parameters of the items administered, trait level may be estimated by means of maximum likelihood techniques. More important, as long as all the item parameters are expressed with reference to a common metric and to a common norm group, trait level estimates in a common metric may be obtained from examinees' scores on different sets of items. For this reason, maximum likelihood scoring is especially appropriate for use with adaptive tests.

Although maximum likelihood scoring allows us to make direct comparisons of persons who took different sets of test items, the method is not without its shortcomings. For instance, the solution is indeterminate when an examinee answers every item correctly or every item incorrectly, in which cases the estimation procedure converges on plus or minus infinity. When items can be answered correctly by guessing, the same problem may occur with other item score patterns as well. Although adaptive tests, by virtue of their item selection processes, are less subject than conventional tests to item response patterns yielding infinite maximum likelihood score estimates, there is no guarantee that such patterns will not occur.

Bayesian scoring. A Bayesian sequential scoring method proposed by Owen (1969) avoids the problem of infinite estimates, yet provides comparable scores from different sets of test items, in the same kind of metric the maximum likelihood procedure employs. The Bayesian method is likewise a consequence of latent trait theory, based again on the properties of item characteristic curves. For simplicity let the item characteristic curves all be normal ogives, so that

$$P_g(\theta) = P(u_g = 1 \mid \theta) = \Phi[a_g(\theta - b_g)] \qquad [3]$$

Again we do not know the value of Θ, but we observed the item scores (1 or 0), and have previously estimated the parameters a_g and b_g of each item g. If we began by estimating that an examinee's trait level Θ was equal to the mean μ_o of a normal distribution, and that the variance of that distribution is σ_o^2, Bayes' theorem permits us to calculate the mean and variance of Θ posterior to observing his score on a single item. That is, using Bayes' Theorem and the parameters of the prior distribution we may proceed from $P(u_g=1|\Theta)$ to $P(\Theta|u_g=1)$ and from $P(u_g=0|\Theta)$ to $P(\Theta|u_g=0)$ which in turn permit us to evaluate expressions for $E(\Theta|u_g)$ and $var(\Theta|u_g)$, the expected value and variance of the posterior distribution of Θ, conditional on item score.

As proposed by Owen in the context of an adaptive testing strategy, the Bayesian estimation procedure never yields the troublesome infinite estimates. It is dependent, however, on the order in which the item scores are evaluated, since it involves updating the trait level estimate one item at a time. Several factors are capable of limiting the accuracy and validity of the resulting "final score" estimates. Guessing can introduce marked bias. Additionally, the Bayesian approach depends heavily on its "priors". An inappropriate choice of parameters for the initial prior distribution can result both in severe bias and some loss of validity (McBride, 1975) in the scores.

Choosing Among Scoring Methods

So, where does that leave us? We have a variety of scoring procedures available for adaptive tests. Two of these have been described above. Others are described by Lord (1970). Some are appealing by virtue of their simplicity, but either fail to provide adequate differentiation among examinees, or to rank examinees on a scale that permits comparing scores obtained on different tests, or both. Others are appealing because of their mathematical elegance, but are subject to distortions such as bias, or to absurdities such as infinite scores, or to invalidity due to inappropriate prior assumptions. Given that we are to use an adaptive test in some applied setting, how are we to choose among alternative scoring methods?

The answer is that there is no simple answer. The choice will depend on the test itself, on the setting in which the test is used, on the purpose to which the test scores are to be applied, on practical constraints such as scoring costs, and perhaps on other considerations as well. Using psychometric criteria, scoring methods can be evaluated in terms of a number of criteria, including information and bias.

Information. Suppose that trait level is distributed continuously, and measured in real numbers. We can talk of the regression of test scores on trait level, that is, a curve depicting the mean test score at any level of the trait. If the regression is linear, we know that its slope is constant, so that for any unit increase in trait level, there is a corresponding constant increase in mean test score. If the regression is non-linear, the increment in mean test score may or may not be linearly related to trait level.

Similarly, we may talk of the precision of measurement at any trait level in terms of the inverse of the standard deviation of test scores at that level. Like the slope, the precision may or may not be constant across trait levels. The "information" at any level of the trait is defined as the squared ratio of the slope at that level to the standard deviation of scores at that level. Information may be constant across trait levels, or may vary. If the information is constant, the test scores are making equivalent discriminations at all levels of the trait. If it is not constant, the test scores discriminate better at some levels of the trait than at others, and perhaps discriminate best at some one point (see Appendix for a further discussion of "information").

Bias. Just as precision and information are discussed in terms of trait level, we may speak of bias at any given trait level. Bias here is defined as follows:

$$\text{bias} = \left[E(X) \,\middle|\, \theta \right] - \theta \qquad [4]$$

where X is the test score. Bias, then, is the algebraic difference between the expected value of the test scores X at a given trait level Θ and Θ itself. As I mentioned earlier, the metric for Θ is arbitrary. So is the test score metric X. Since both are arbitrary, we should be more concerned about the form of the relation of bias to Θ than to the numerical values. Constant bias, or bias linear in Θ, is not usually a problem in psychological measurement. Non-linear bias, however, may be a problem in some applied settings.

Comparison of Maximum Likelihood and Bayesian Scoring

In choosing a scoring method for an adaptive test, it would be prudent to evaluate the information and bias characteristics of the resulting scores against the criteria dictated by the purpose of testing. These evaluations may be conducted by analytic methods for certain kinds of tests (e.g., Lord, 1970), but where real item pools are involved, Monte Carlo computer simulation methods may be necessary. An example of such a simulation follows (see Appendix for details of the simulation method; numerical results are in Appendix Tables A-2 through A-4).

This simulation study used the Bayesian sequential adaptive testing strategy designed by Owen (1969). Rather than accepting Owen's method for scoring the resulting patterns of item responses, however, we wanted to evaluate it in comparison with two alternative scoring procedures: 1) the maximum likelihood estimation procedure described above and 2) the number correct score.

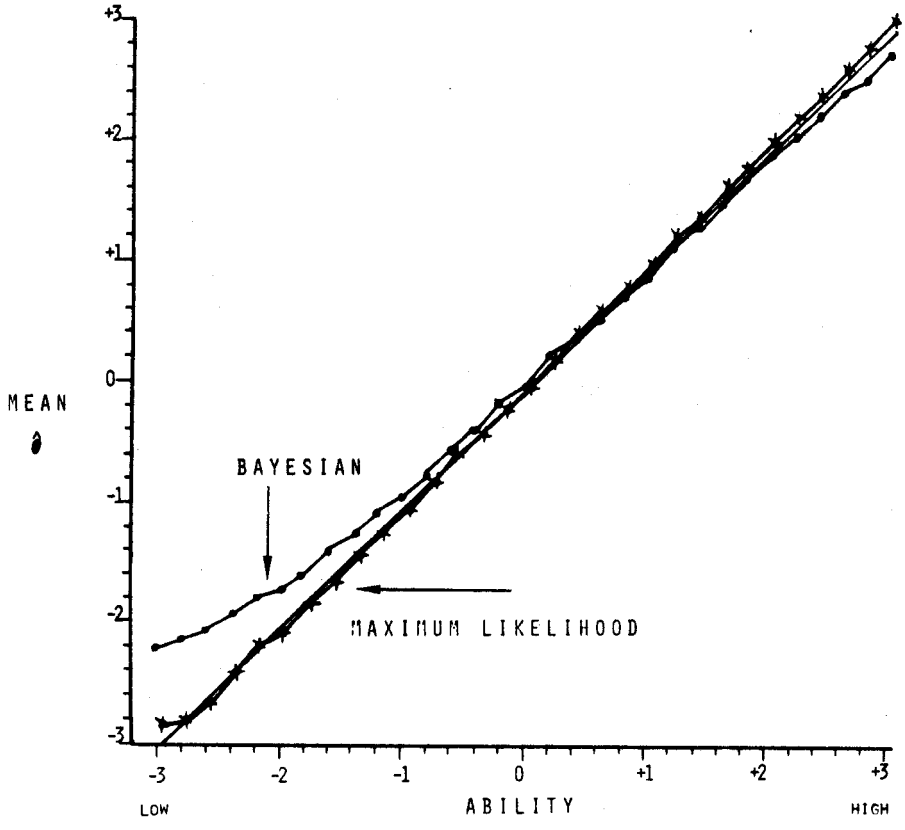
In order to generate data from which to compare the three scoring methods, we simulated administering a 20-item Bayesian sequential adaptive test to 3200 examinees of known ability--100 examinees at each of 32 trait levels (Θ) in

the interval $[-3.2 \leq \theta \leq +3.0]$. These trait level values can be thought of as standard deviation units. A pattern of 20 simulated item scores (1 or 0) was generated for each simulated examinee. Every such pattern was scored using each of the three scoring methods. For each scoring method, the mean and standard deviation of the 100 scores at each trait level θ were calculated.

Regression of scores on ability. The means are plotted against trait level θ in Figures 19 and 20. Figure 19 contains the mean scores for the

Figure 19

REGRESSION CURVES FOR BAYESIAN AND MAXIMUM LIKELIHOOD SCORING



Bayesian scoring method. Note that the estimated regression of Bayesian scores on θ is slightly non-linear. Its slope varies from one level to another, which has implications for the information in the scores. Figure 19 also contains the means for the maximum likelihood scoring technique. Note that the regression of these scores on θ appears almost linear. Figure 20 shows the mean number-correct score as a function of θ. For these scores the regression is somewhat non-linear.

Figure 20

REGRESSION CURVE FOR NUMBER CORRECT SCORE

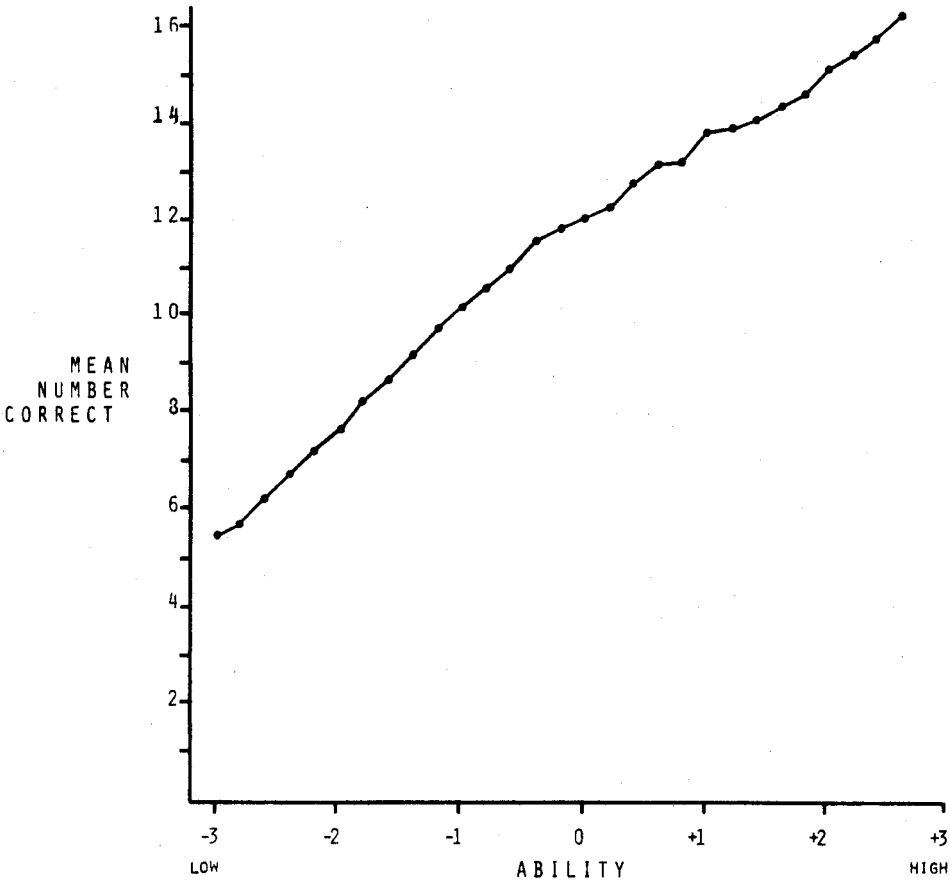
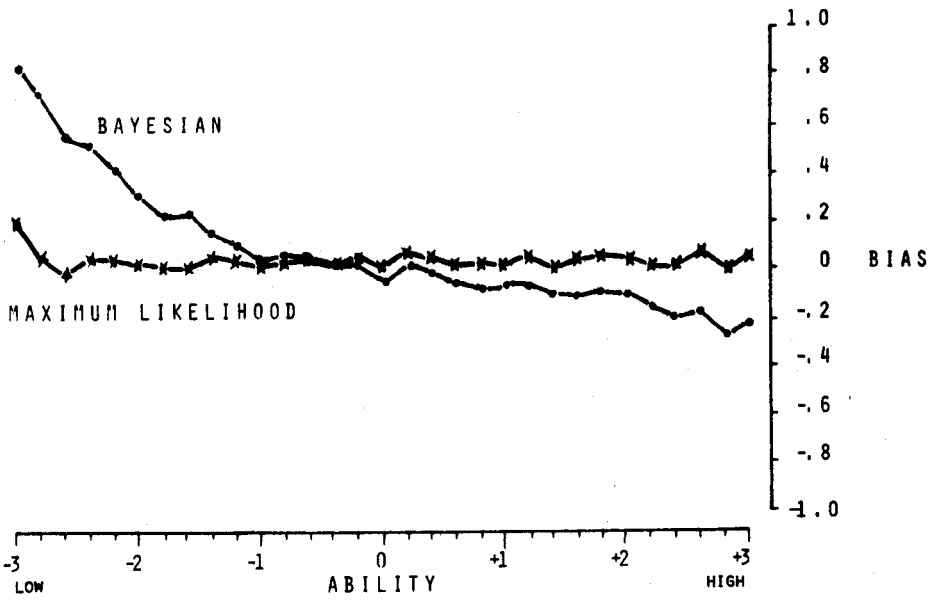


Figure 21

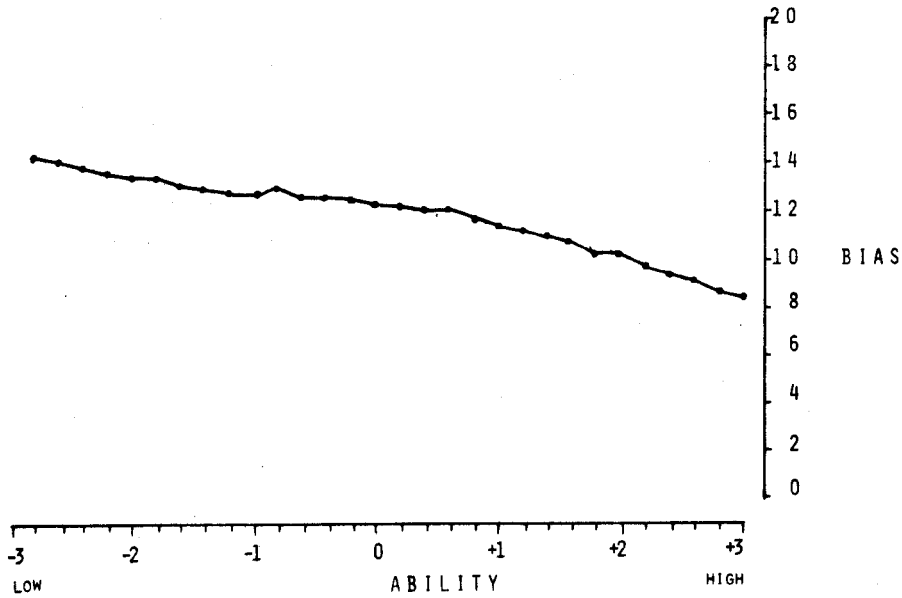
BIAS CURVES FOR BAYESIAN AND MAXIMUM LIKELIHOOD SCORING



Bias. Figure 21 contains bias plots for the Bayesian and maximum likelihood scores. Figure 22 is the bias plot for the number correct scores. In the trait interval shown, the maximum likelihood scores appear to be nearly unbiased estimators of trait level. The Bayesian scores are not so favorable in this regard. The bias is severe in the extremes of trait level, and is noticeably non-linear. The bias in the number correct scores follows a trend similar to that of the Bayesian scores.

Figure 22

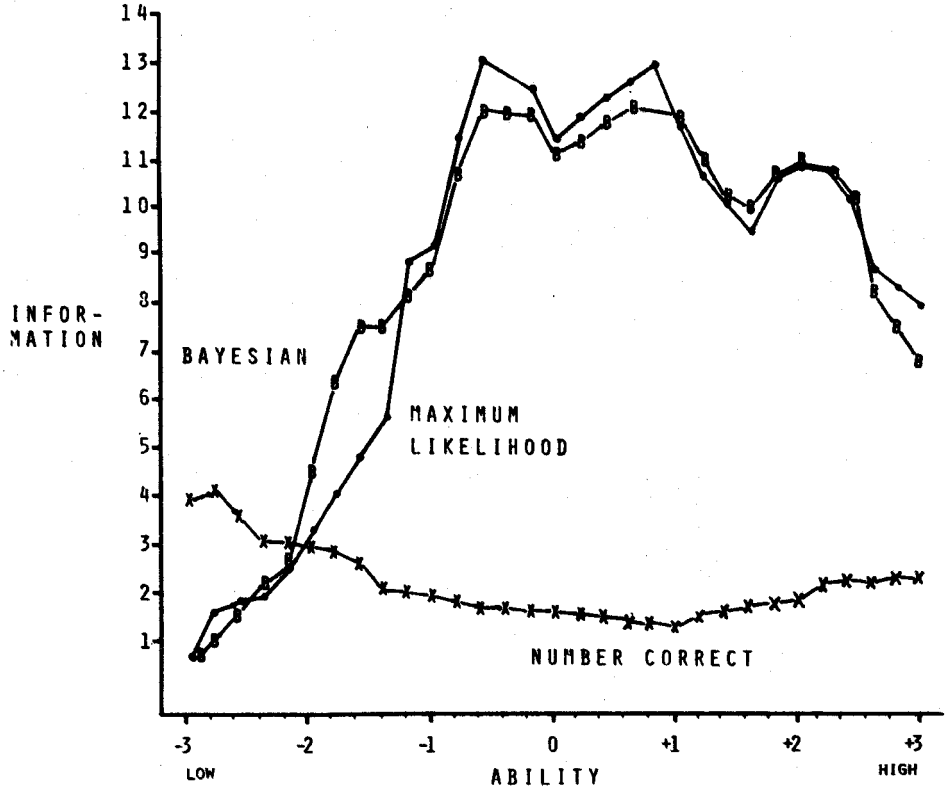
BIAS CURVE FOR NUMBER CORRECT SCORE



Information. So far we have looked at the regression of test scores on θ, and found that only for the maximum likelihood scores was the regression approximately linear. Similarly, the maximum likelihood scores appear far superior to the other two in terms of bias. Now let us look at the estimates of the information curves for the three methods. Figure 23 shows these for all three scoring methods. Both the Bayesian and the maximum likelihood curves are convex, rising from near zero at θ=-3 to a peak of 13 in the mid-range, then declining somewhat in the upper trait levels. The shapes of the curves are so similar, and their differences so small that it would be difficult to call either method superior in information in the θ range from -1 to +3. The number-correct information curve, on the other hand, is concave, and is clearly inferior to the other two except at the very low trait levels.

Figure 23

INFORMATION CURVES FOR THREE SCORING METHODS



Limitations of the Scoring Methods

Given the three scoring methods, then, which one should we select for use? The number correct score is obviously inappropriate except for ranking persons in the extreme low end of the trait level range. The similarity of the information curves for the two latent trait estimation techniques suggests that they are virtually interchangeable for ordering persons, other things being equal.

But of course other things are not equal. The maximum likelihood estimation method is about three times more expensive than the Bayesian one. On the other hand, the Bayesian method of scoring is subject to non-linear bias. If unbiased measurement were a goal of the test, the expense of the maximum likelihood procedure might be justified. If simple ordering of persons with respect

to trait level were all the tester required, the Bayesian scores seem preferable[1]. Other than that, no simple prescription is advisable.

I have mentioned only two true latent trait scoring methods. Numerous other scoring methods are available (e.g., Larkin & Weiss, 1974; Weiss, 1973), most of which lack the mathematical elegance of the Bayesian and maximum likelihood methods, yet may approach both in terms of information. All of these methods provide a sufficient range of scores to permit maximal discrimination among persons (if test length is sufficiently long), and many of them use all the information in the pattern of item responses. The two that I have illustrated above also permit comparisons of scores obtained on different tests of the same trait, although the bias in the Bayesian scores may make such comparisons hazardous. The point of this discussion has not been to prescribe an all-occasion scoring method, but rather to show that there is a choice, and to suggest computer simulation as a tool to facilitate a rational choice among alternatives in the face of shifting decision parameters.

---

[1]Test scores are usually used only to order persons relative to one another, or to classify them into two or more discrete categories. Technically, both Owen's scoring method and the maximum likelihood one are statistical estimation procedures. As such they are useful for actually estimating parameters θ_i characterizing persons i, on the basis of responses to a set of test items. For applied purposes requiring only the ranking or classification of persons, the test score information curves are of paramount interest. But there may be certain applications in which actual parameter estimates are important. For these applications the small-sample (where sampling is over items) bias characteristics of the estimation procedure have important implications for the utility of the resulting estimates.

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.

PROSPECTS: NEW TYPES OF INFORMATION AND PSYCHOLOGICAL IMPLICATIONS

Nancy E. Betz
University of Minnesota

Traditional psychometric theory and practice has largely failed to take advantage of the full variety and extent of information obtainable from responses to test items. Consequently, the most information usually extracted from a testee's responses to a series of items is a total test number correct score, or a score on some personality dimension or interest scale.

But patterns of test item responses are far richer in information and are far more complex to interpret than single number correct scores would imply. Computer-assisted testing procedures provide us with the capability of extracting much more and a greater variety of information about an individual or about the meaning of his/her score than have conventional testing procedures.

New Types of Information

Individualized errors of measurement. Probably one of the most important new types of information obtainable from computerized adaptive trait measurement procedures is a value indicating the accuracy of a given individual's score on a test--that is, a value indicating the degree of confidence we can place in a particular individual's test score. The traditional psychometric approach to this problem has involved the determination of a reliability coefficient characterizing a whole test--from that reliability coefficient we derive a standard error of measurement which we use to estimate the amount of probable error in a given individual's test score. However, this standard error of measurement represents the average expected error over all individuals in the group and, as Mr. Vale has shown, the error in a typical peaked conventional test is much greater for individuals whose ability levels deviate from the average. Consequently, the *average* expected error may be an overestimate or an underestimate of the amount of error in any one score.

Several of the adaptive testing strategies provide individualized estimates of score accuracy. For example, the Bayesian adaptive testing strategy provides, along with an ability estimate following each item administered, a value indicating the error of that estimate.

Figure 24 shows an example of ability estimates and errors obtained as successive items are administered to an individual in a Bayesian adaptive test. Note how the size of the error band around the ability estimate decreases as responses to successive items provide us with more information and a more stable estimate of ability.

Figure 24

REPORT ON BAYESIAN TEST

X=CORRECT O=INCORRECT ?=NO RESPONSE
ERROR BAND PLOTTED IS + AND - STANDARD DEVIATION

	LOW ABILITY LEVEL HIGH										POSTERIOR ABILITY	
	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0		
STAGE	I----I	----I	----I	----I	----I	----I	----I	----I	----I	----I	EST	SD
0						E					-.85	1.87
1						X					.19	1.30
2						O					-.85	.94
3						X					-.36	.76
4						O					-.73	.66
5						O					-1.07	.65
6				O							-1.59	.50
7				X							-1.42	.44
8				X							-1.29	.40
9				X							-1.20	.39
10				O							-1.30	.36
11				O							-1.40	.36
12				X							-1.31	.35
13				X							-1.24	.33
14				X							-1.17	.33
15				O							-1.23	.32
16				O							-1.28	.32
17				O							-1.35	.30

17 ITEMS WERE ADMINISTERED

In Bayesian adaptive testing, we can either fix the number of items administered, thus allowing the error of the ability estimate to vary across individuals, or we can administer different numbers of items to different individuals with the intention of terminating the test when an acceptably small degree of measurement error has been achieved. Thus, the Bayesian ability estimate is far more interpretable than are conventional test scores because we can obtain an estimate of the amount of probable error in each individual's score.

Response consistency. Another type of information obtainable from some adaptive testing strategies is something that we have called the *consistency* of an individual's response pattern. Consistency refers to how reliably or consistently an individual is interacting with an item pool.

In personality assessment, response inconsistency is usually assessed using various types of validity scores. The notion of inconsistency in, for example, pair comparisons or forced choice formats, is operationalized as the number of circular triads. If a person's response pattern contains too many circular triads, we infer that something besides the trait of interest is influencing the person's responses and declare his test protocol invalid.

In ability measurement, we would expect that an individual should, in general, respond correctly to items below, or easier than, his/her ability level, and incorrectly to items above, or more difficult than, his/her ability level. If a person answers most easy items correctly and most difficult items incorrectly, we would say that he is responding consistently--that is, his response pattern seems to be influenced primarily by his position on the underlying trait continuum. However, if a person answers many easy items incorrectly and many difficult items correctly, he is responding inconsistently, indicating that something besides the trait of interest is influencing his responses.

In an ability test, response inconsistency may be caused by such extraneous variables as guessing, partial knowledge, or adverse psychological conditions such as test anxiety or lack of motivation to do one's best on the test. Whatever its cause, response inconsistency may reduce the reliability and/or validity of a given test score. And, knowing the degree of consistency of an individual's response pattern may be important when we intend to use that score in making practical decisions.

We have operationalized the notion of response consistency in the stradaptive testing strategy. As you may recall from Mr. Vale's presentation (see Figure 15), in the stradaptive test, items are organized into a series of levels or strata according to their difficulty. A correct response to an item in one stratum leads to the administration of the most discriminating item remaining in the next more difficult stratum. An incorrect response leads to the administration of the most discriminating item remaining in the next less difficult stratum.

Figure 25 shows a relatively consistent response pattern on the stradaptive test along with 10 ability scores and five consistency scores. This person entered the stradaptive test at stratum 5, based on some prior information. Stratum 5 items were too easy for him and he answered items correctly until, at item 4, he had been branched to stratum 8, which contained very difficult items. Notice that he consistently responded incorrectly to the stratum 8 items, which were too difficult for him, and correctly to the stratum 6 items, which were too easy for him. The items in stratum 7 seem most appropriate in difficulty, and he answered about half of them correctly and the other half incorrectly.

The consistency of this individual's response pattern was reflected in his relatively low consistency scores. Score 11, defined as the standard deviation of the difficulties of the items encountered by this person, was .59. Further, in the stradaptive test, items are administered until a termination criterion is reached. The consistency of this individual's response pattern enabled him to meet the termination criterion after only twenty items had been administered.

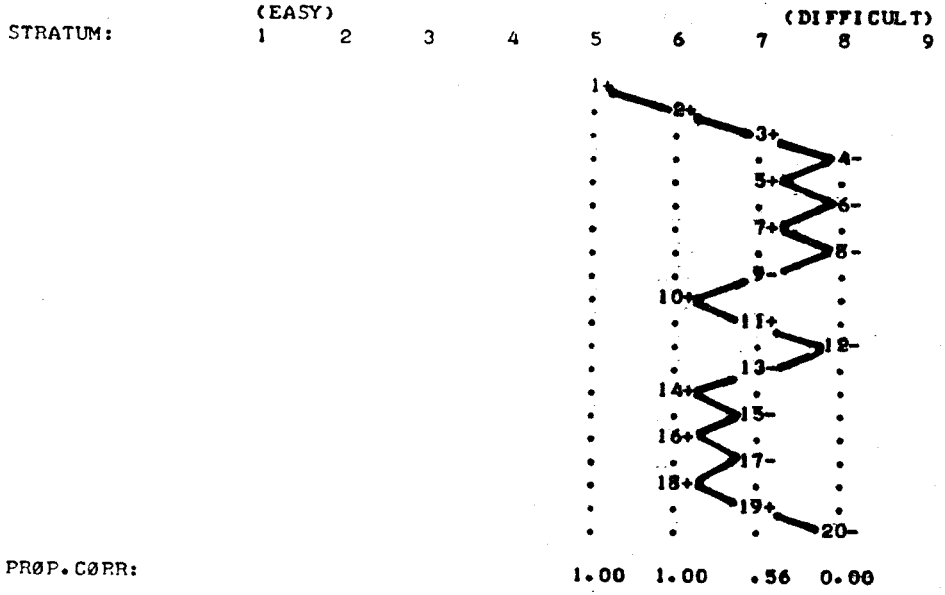
Contrast the response pattern of this consistent examinee with the one shown in Figure 26. The response pattern shown in Figure 26 was far less consistent and ranged over a larger number of strata, and thus a larger range of item difficulty. For example, this person answered some relatively easy items at stratum 5 incorrectly (e.g., items 8 and 26) and answered some difficult items at stratum 8 correctly (e.g., items 1 and 17). By responding inconsistently, it took many more items before the termination criterion was reached, and the individual's consistency scores are higher, reflecting a less consistent response pattern.

Figure 25

Report on a Stradaptive Test for a Consistent Testee

REPØRT ØN STRADAPTIVE TEST

ID NUMBER: DATE TESTED: 73/07/12



TØTAL PRØPØRTIØN CØRRECT= .550

SCØRES ØN STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= 1.49
2. DIFFICULTY ØF THE N+1 TH ITEM= 1.44
3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= 1.49
4. DIFFICULTY ØF HIGHEST STRATUM WITH A CØRRECT ANSWER= 1.33
5. DIFFICULTY ØF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= 1.33
7. INTERPØLATED STRATUM DIFFICULTY= 1.37
8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= .88
9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN CEILING AND BASAL STRATA= 1.28
10. MEAN DIFFICULTY ØF ITEMS CØRRECT AT HIGHEST NØN-CHANCE STRATUM= 1.28

Consistency Scores

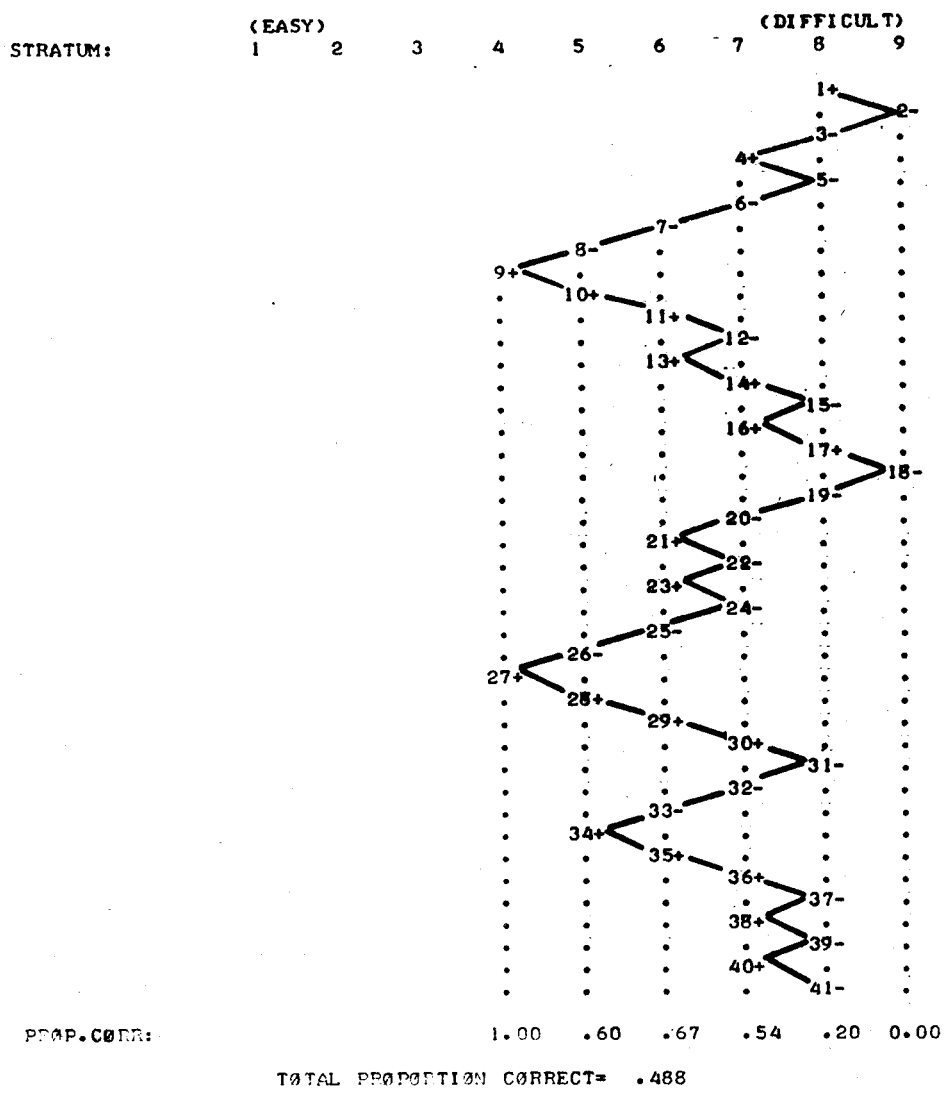
11. SD ØF ITEM DIFFICULTIES ENCØUNTERED= .59
12. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY= .46
13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY BETWEEN CEILING AND BASAL STRATA= .18
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 1.36
15. NUMBER ØF STRATA BETWEEN CEILING AND BASAL STRATA= 1

Figure 26

Report on a Stradaptive Test for an Inconsistent Testee

REPØRT ØN STRADAPTIVE TEST

ID NUMBER: DATE TESTED: 73/07/02



TØTAL PRØPØRTIØN CØRRECT= .488

SCØRES ØN STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= 1.89
2. DIFFICULTY ØF THE N+1 TH ITEM= 1.01
3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= 1.53
4. DIFFICULTY ØF HIGHEST STRATUM WITH A CØRRECT ANSWER= 2.01
5. DIFFICULTY ØF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= 1.33
7. INTERPØLATED STRATUM DIFFICULTY= 1.36
8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= .72
9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN CEILING AND BASAL STRATA= .76
10. MEAN DIFFICULTY ØF ITEMS CØRRECT AT HIGHEST NØN-CHANCE STRATUM= 1.24

Consistency Scores

11. SD ØF ITEM DIFFICULTIES ENCØUNTERED= .86
12. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY= .74
13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY BETWEEN CEILING AND BASAL STRATA= .50
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 2.64
15. NUMBER ØF STRATA BETWEEN CEILING AND BASAL STRATA= 3

Consistency and stability. We hypothesized that the ability test scores of individuals who are responding consistently should be more reliable than those of individuals who are responding inconsistently. To study this hypothesis, we used test-retest stability as an indication of score reliability, and divided a group of 200 subjects into five groups on the basis of their consistency scores on the first stradaptive test administration in a test-retest design. Within each group, we calculated the test-retest stability of the obtained ability estimates. Table 1 shows the results obtained for the consistency score defined as the standard deviation of the difficulties of all items encountered.

Table 1

STRADAPTIVE AND CONVENTIONAL TEST TEST-RETEST CORRELATIONS AS A FUNCTION OF CONSISTENCY SCORE 11 ON INITIAL TESTING

		STATUS ON CONSISTENCY SCORE 11				
		VERY HIGH	HIGH	AVERAGE	LOW	VERY LOW
MEAN CONSISTENCY SCORE		.517	.625	.706	.815	1.038
NUMBER OF TESTEES IN INTERVAL		27	30	41	43	29
STRADAPTIVE ABILITY SCORE:	1	.940	.849	.847	.768	.652
	2	.875	.721	.799	.778	.751
	3	.956	.813	.878	.826	.708
	4	.934	.840	.846	.731	.664
	5	.896	.722	.793	.756	.741
	6	.950	.798	.886	.820	.704
	7	.970	.844	.902	.851	.758
	8	.981	.927	.915	.853	.869
	9	.983	.939	.907	.899	.889
	10	.951	.792	.882	.822	.718
CONVENTIONAL TEST		.979	.890	.918	.826	.878

As the table shows, the highest test-retest stability was found in the most consistent group of examinees for all ten ability scores. The clearest pattern is that for ability score 1, where the scores in the most consistent group had a test-retest stability of .94, while the scores in the least consistent group had a stability of .65. The stabilities in the intermediate groups decreased with decreasing consistency. Note also that the stability for the most consistent examinees on scores 8 and 9 was .98, indicating very high stability of the obtained ability estimates. These results suggest that the use of consistency scores as moderator variables may provide us with additional information concerning the accuracy of longitudinal predictions from test scores.

Thus, such indices as estimates of the degree of accuracy of a given individual's test score or the consistency of a test response protocol add greatly to our capacity to meaningfully interpret a test score, and to the utility that the score will have in practical decision-making contexts.

Additional new kinds of information. Computerized trait measurement can provide us with additional types of information. For example, the computer can provide precise control over a subject's usage of confidence weighting procedures or probabilistic responding, which can be used to assess partial knowledge. When confidence weighting has been used in a paper and pencil format, it has frequently been found that some examinees fail to assign probabilities to the response alternatives in accordance with the test instructions. In computer-administration, however, the examinee is informed immediately when he has not assigned probabilities according to the rule. Thus computerized test administration can eliminate the problem of invalid test protocols.

Computerized testing also has the capability of providing us with exact response latency data for each item administered. Response latency data have a variety of potential uses. For example, it might be used in conjunction with confidence weighting procedures to aid in the identification of guessing behavior. In the area of personality assessment it could be useful in identifying the presence of random responding or response sets. Finally, the measurement of response latencies may lead to further understanding of the speed versus power components of ability.

Perhaps the most potentially important and fruitful area of research using computerized testing lies in the study of human problem-solving and reasoning abilities. Traditionally, psychometricians have asked *how many* problems a person could solve and have left it to the experimentalists to investigate the nature or the "how" of the problem-solving process. But knowledge of the process of problem-solving should be a part of our theories of human abilities and could contribute substantially to the construct validity of such theories.

One approach to the study of problem-solving abilities using computerized test administration would involve a within-problem branching sequence in which a series of interdependent questions are organized into a problem-oriented structure. For example, one response at a given point in the structure might result in the testee's arriving at a correct solution by an entirely different pathway than would a different response at that given point. We could study the amount and type of information the testee needs to solve a problem, the efficiency with which he goes about it, and the different problem-solving systems or pathways utilized by different individuals.

The time now seems right, therefore, for using the computer to integrate the measurement and the study of intelligent behavior. Limiting the information we obtain from test-taking behavior to whether an item was answered correctly or incorrectly is wasteful of much potentially significant and useful information and is now no longer necessary, thanks to the availability of computer-assisted testing procedures.

Psychological Effects

In addition to the variety of new information obtainable from computer-assisted testing procedures, it also has the potential to improve the psycho-

logical environment of testing. In the past, psychometricians have paid considerable attention to the characteristics of tests administered to groups, for example, their reliability and validity. But we have forgotten that it is an *individual* who takes a test, not a group. Highly valid and reliable tests can be rendered useless for the measurement of an individual if, for one reason or another, he is not performing to his fullest capacity. For example, substantial amounts of error in the test score of an individual may result if that person's performance is hindered by high levels of test anxiety or if he is not motivated to do his best or to respond truthfully to test items.

Anxiety, motivation and frustration. In the area of ability measurement, tests are typically geared to the ability level of the average member of a group. Such tests will be a rather different experience for examinees of differing ability levels. The low ability individual receives a series of items which are too difficult for him or her and may react by becoming threatened, anxious, or frustrated--the test may seem hopeless and he may simply stop trying. The high ability individual, on the other hand, receives items which are too easy for him--this person may find the task boring and unchallenging and, in a fashion similar to that of the low ability examinee, may simply stop trying to do his best. It is only for the average ability examinee that the items are likely to be sufficiently difficult to be challenging and yet not so difficult as to seem hopeless.

Adaptive testing procedures, however, tend to maintain an appropriate level of item difficulty for each individual. We don't yet know whether or not difficulty levels appropriate to each individual's ability level are the best ones for keeping motivation at high levels and anxiety and frustration at low levels. But at least adaptive testing procedures should keep the relative degree of item difficulty constant across ability levels and should thus have less tendency to arouse differential levels of motivation, anxiety, or frustration in individuals of different ability levels.

Feedback. Computerized test administration also makes it very easy to provide the examinee with feedback, immediately after each item response, as to the correctness or incorrectness of that response. A number of writers (e.g., Bayroff, 1964; Ferguson & Hsu, 1971; Zontine, Richards & Strang, 1972; Strang & Rust, 1973) have suggested that immediate knowledge of results, or feedback, may have positive motivating effects on some examinees and, therefore, may increase the likelihood that they will perform to their fullest capacities. Knowledge of results has long been considered important in the area of learning and instruction and has been built into methods of programmed and computer-assisted instruction. Further, the constructors of individually-administered intelligence tests, for example, Binet, Terman & Wechsler, all stressed that some form of encouragement by the examiner was essential in keeping the examinee motivated and performing to his fullest capacity, although this encouragement was *not* to include knowledge of results *per se*.

Since the effects of immediate feedback on performance on objective tests of ability have been only rarely studied, we have incorporated immediate feedback into some of our research designs.

Feedback and race.[2] In one study, both a conventional test and a pyramidal adaptive test were administered by computer to a group of inner-city high school students. The group was racially mixed, consisting of both black and white students. Tests were administered such that half the group received the conventional test first, while the other half received the pyramidal test first. Within each order of test presentation, half the group received feedback and the other half did not.

The results of the 3-way ANOVA for the conventional test scores are shown in Table 2, using number correct as the dependent variable. The only significant main effect was for race, with the overall performance of blacks being significantly lower than that of whites.

Table 2

3-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
ORDER	1	105.76	1.36	.25
RACE	1	2,013.26	25.84	.00*
FEEDBACK	1	81.74	1.05	.31
RACE X ORDER	1	161.54	2.07	.15
ORDER X FEEDBACK	1	28.74	.37	.55
RACE X FEEDBACK	1	170.40	2.19	.14
ORDER X RACE X FEEDBACK	1	599.40	7.69	.01*
ERROR	82	77.92		

However, the 3-way interaction among order, race, and feedback was highly significant. Figure 27 shows the means for the 3-way interaction. The left side of the graph shows the group means under feedback conditions, while the right side shows the means under no-feedback conditions. Note that the performance of whites was uniformly better than that of blacks *except* under feedback conditions when the conventional test was given first. In this case, the performance of blacks was not significantly different from that of whites.

Further analysis of this result suggested that it was due to motivational effects. If it can be replicated it suggests the possibility that under optimal conditions of test administration the performance differential between racial groups might be substantially reduced.

Feedback, ability level and testing strategy. In a second study, either a conventional test or a stradaptive test was administered with or without feedback in two groups of subjects. One group was a "high ability" group (College of Liberal Arts) and the other a relatively "low ability" group (General College) based on average college admission test scores and high school grades.

[2]These data were analyzed by Ms. Clara DeLeon.

Figure 27

MEAN SCORES FOR BLACKS AND WHITES COMPLETING THE 40-ITEM CONVENTIONAL TEST FIRST AND SECOND, BY FEEDBACK CONDITION

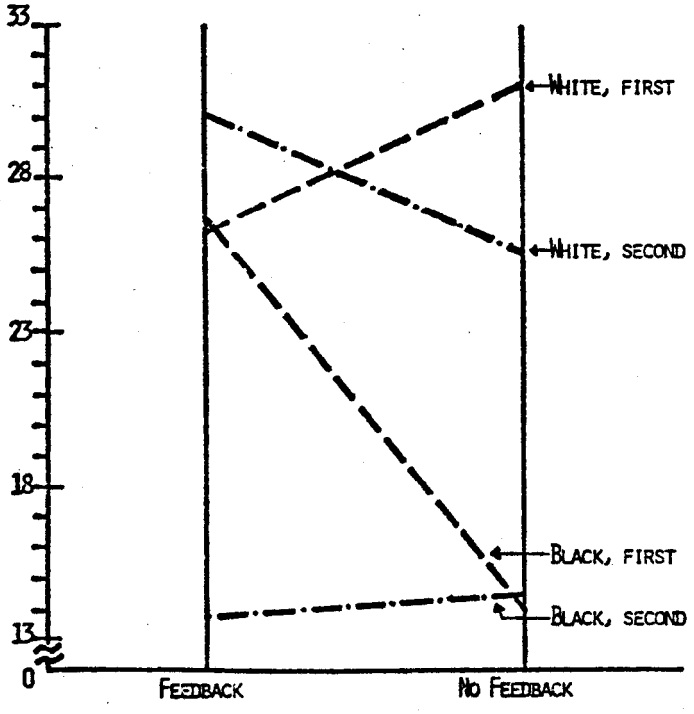


Table 3

MEAN NUMBER CORRECT ON 50-ITEM CONVENTIONAL TEST FOR TWO SUBJECT GROUPS WITH AND WITHOUT FEEDBACK

	FEEDBACK		NO FEEDBACK		TOTAL	
GROUP	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	30.47	57	27.10	117	28.83
GENERAL COLLEGE	28	22.54	28	20.71	56	21.62
TOTAL	88	27.94	85	25.00	173	26.50

TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	1945.29	21.67	.001*
FEEDBACK	1	354.28	3.95	.046*
GROUP X FEEDBACK	1	22.45	.25	.999
ERROR	169	89.77		

Table 3 shows the mean number-correct scores on the conventional test according to whether feedback was or was not given. The analysis of variance indicated a significant main effect for feedback, and analysis of the means indicated that in both subject groups, the provision of feedback resulted in significantly higher test scores. For example, in the College of Liberal Arts group, the mean number correct under feedback conditions was over 30, while that under no-feedback conditions was only 27. A difference of 3.5 score points on a 50-item test could be highly influential in a practical decision about an individual.

The results for the conventional test showed that feedback had a positive effect on test performance, but when we looked at the stradaptive test, the results were quite different. Table 4 shows maximum likelihood scores on the stradaptive test under feedback and no feedback conditions. Note that there is no significant effect for feedback.

Table 4

ABILITY ESTIMATES FOR STRADAPTIVE TEST FOR TWO SUBJECT GROUPS WITH AND WITHOUT FEEDBACK

	FEEDBACK		NO FEEDBACK		TOTAL	
GROUP	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	-.66	62	-.62	122	-.64
GENERAL COLLEGE	28	-.96	27	-.81	55	-.89
TOTAL	88	-.76	89	-.68	177	-.72

TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	2.27	1.75	.184
FEEDBACK	1	.24	.19	.999
GROUP X FEEDBACK	1	.10	.07	.999
ERROR	173	1.29		

However, in trying to interpret these apparently conflicting results, it is necessary to remember that in the stradaptive test, almost everyone answers about half the number of items administered correctly; thus, the feedback should be about half negative and half positive. In the conventional test, however, high ability examinees receive mostly positive feedback while low ability examinees receive mostly negative feedback. Further, the stradaptive test maintains item difficulties at levels appropriate to each examinee's ability so it is perhaps a less stressful and more positive experience, particularly for "low ability" testees.

Further analysis revealed that the levels of motivation reported by examinees who took the stradaptive test were uniformly higher than the levels reported by those who took the conventional test. These data suggest that an adaptive test led to higher levels of motivation whether or not feedback was given. Thus, particularly for the low ability testees, an adaptive test may have the same motivational effects that giving feedback on a conventional test seems to have.

Implications. The results I have presented here are obviously not conclusive. Replications and further studies are certainly necessary. But given the current concern with test fairness and bias, it seems that we should pursue further the effects of various conditions of test administration upon performance. Adaptive testing and immediate knowledge of results may be able to provide testing conditions more conducive to allowing each individual to demonstrate his/her fullest capacities in test performance. And, since computerized adaptive trait measurement can provide us with important additional information of a variety of types, it has promise of supplementing the paper and pencil tests which have dominated psychological testing for the last 50 years.

DISCUSSION

Robert L. Linn . . . . . University of Illinois

I'd like to start by commending David Weiss and the group of people involved in the Psychometric Methods Program at the University of Minnesota for the continued high quality work on issues related to adaptive testing. This work is praiseworthy in several regards but, in my view, most notably for its continued and systematic nature and for the use of multiple approaches combining theoretical, simulation and empirical techniques. These aspects give the work a cumulative quality that is too often missing.

Thanks largely to the continued work on problems in adaptive (or tailored) testing by Fred Lord and the work at the University of Minnesota there is by this time a pretty good understanding of the potential value of adaptive testing techniques, at least under idealized conditions. The best of the adaptive testing procedures provide the promise of measurement that is nearly of equal precision throughout a wide range of ability with only a small loss compared to a peaked conventional test at an ability level equal to the difficulty location of the peaked test. David Vale's results support this conclusion and indicate that several techniques have relatively good potential.

There are a couple of general questions, however, that need to be kept in mind in drawing conclusions from results such as Vale's. One of these issues is implicit in the limitations noted by Vale at the end of his paper. That is, will the results based on an overly simple model generalize to real items and real examinees? Items do differ in discriminating power. Furthermore, items with equal discriminating power are not apt to be uniformly distributed over item difficulty. Also, multiple-choice items are the backbone of most standardized testing and such items generally require another item parameter for the lower asymptote. For these reasons I'd like to see more simulation studies that are based on estimated item parameters (preferably with three parameters per item) for actual pools of items.

Fred Lord has recently released an ETS Research Bulletin which not only shows some very promising work of this type but includes an offer to make available item parameter estimates based on the three-parameter logistic model for 690 items from some fifteen forms of the SCAT and STEP tests. I think that the exploitation of this pool of items and parameter estimates in future simulation and empirical studies could be a great help in moving our understanding of adaptive testing forward.

A second type of question that needs to be addressed in considering the implications of results such as were presented by Vale and by McBride is whether the gain is worth the extra effort and precision. This is a pragmatic question and the answer will undoubtedly depend on a number of considerations. Important among these considerations, however, is the purpose of the testing. The procedures considered are of value where accurate measurement over a wide range is important. For many testing purposes equi-precision is

not very important. For example, in selection for a particular institution precision is needed near the cut point, not over a broad range, and here the peaked test does very well.

On the other hand, there are situations where precision over a wide range is needed. Some of these were discussed by Wood in his review article in the Review of Educational Research. Examples are for tests used by a wide variety of institutions for many purposes such as a college admissions exam. Another area is where there is an interest in plotting trends (growth) over extended periods of time. In the latter situation, however, the comparisons to conventional procedures might be fairer if the possibility of using prior information was allowed not only for the adaptive procedures but for the conventional procedures. For example, Vale showed nice results for the stradaptive test with different starting levels. Why not use prior information to select different peaked or other conventional tests? This is done in crude form all the time on educational achievement test batteries that have, so-called, vertically equated tests appropriate for different grade levels. However, current vertical equating is not based on an adaptive testing model, and there is reason to believe that current vertical equating procedures are rather inadequate.

The remainder of my comments are mainly on the paper by Nancy Betz and, to a lesser extent, the one by James Sympson. I think there is a need for considerably more research of the type reported by Betz under the heading of Psychological Effects. Feedback effects are not a necessary part of a computer-administered test but are an obvious possibility. The possible effects of feedback on the measurement characteristics of the instrument are many and mostly unknown. One might postulate that an adaptive test would be a less frustrating experience for low ability examinees because they would encounter fewer difficult items. On the other hand, many tests are arranged with easy items toward the beginning of the test and progressively more difficult items later in the test. Thus low ability examinees might have a less frustrating experience as the result of the very easy items early in the conventional test than on an adaptive test with a single middle difficulty entry point.

The three-way interaction of race by feedback by order obtained by Betz is a tantalizing result. It clearly is one that is of sufficient potential importance to require replication. Assuming that the result can be replicated then many questions will need to be addressed, with the primary one being--Is the same trait measured under feedback and no feedback conditions?

In the second feedback study reported by Betz it was unclear to me why there was no group difference on the adaptive test. Doesn't this suggest a problem with the adaptive test?

The focus on feedback vs. no feedback is an example of looking at one of the components that Sympson wants to have separately evaluated. The logic of separating the components is good but there is also a possibility of interaction which requires evaluation of the composites.

I'd like to mention two other types of testing problems where adaptive procedures may be of value. One is in instructional uses of tests where frequent measures are needed for short-term dichotomous decisions. Adapting

test length to the examinee can yield savings in testing time. The second problem area is in multidimensional measurement problems where allocation of testing time for various dimensions might be adapted to the individual.

In summary, I would mainly like to encourage more work which uses as a base parameters that mirror existing item pools, using these both for simulation work and for corollary empirical work, and more efforts on psychological effects.

R. Darrell Bock . . . . . University of Chicago

As these excellent papers were being presented I made a few notes that I'll discuss in turn. Any discussant has to face the question of how much of the difficulties of the subject he is going to let the speaker assume away. In the case of David Vale's presentation, I find myself very reluctant to let him assume away the item heterogeneity and the possibility of correct responses due to guessing.

My experience has been that sets of items that are supposed to be homogeneous are often surprisingly heterogeneous. The Ravens Matrices Test, for example, is usually considered homogeneous and certainly scored as such. But David Thissen, one of the students at Chicago, has a paper to appear in Educational Measurement in which he reports an item analysis of the Ravens A, B, and C sections. He found that the discriminating powers of the items estimated in this procedure indicate that the main source of discrimination is a subset of items in Section B. They define the well-determined dimension underlying the test and the other items contribute little to it.

This is typical of the dilemma that may confront the test constructor: he has items whose discriminating powers vary a great deal, so that if he were to throw away items in order to obtain a set that is homogeneous in discriminating power, he would be in the embarrassing position of throwing away what appeared to be the best items.

I am uncertain how to deal with this problem. Perhaps it is actually risky to use these highly-discriminating items because their source of discrimination may be something peculiar to the particular data. The discriminations may be valid for the calibrating population, but not generalized to the population to which the test will be applied. If this is the case, we may be well advised to regard these highly-discriminating items with suspicion and to remove them, or at least to adjust downward the discriminating powers when estimating latent scores for subjects. The issue is difficult to resolve, but any proposals for test construction that assume them away cannot be considered ready for implementation.

The other matter is that of guessing. I am not enthusiastic about any solution to this problem that assumes all subjects are guessing in the same way. Willingness to guess is very much a personality characteristic that cannot be suppressed even by explicitly instructing all subjects to guess. Some will guess, but others will omit items rather than mark them randomly. If the item analysis procedure is based on the assumption that all subjects guess, then it will have to, in effect, assign random responses to omitted items. But such a practice

may so seriously degrade the information that the test gives about the subject that it is misinformation and is better ignored than included in the scoring procedure.

A better strategy is one in which there is an evaluation of the probability that a given subject is in fact guessing in his response to a given item. This is essentially the strategy taken by Michael Waller in dissertation work at Chicago recently reported in an Educational Testing Service Research Bulletin. On the basis of a provisional estimate of a subject's ability and a provisional estimate of the difficulty of the item, Waller sets up an objective rule for deciding whether or not that particular response should be deleted from the next stage of estimation of the item parameter and latent ability. This is very similar to the approach to data analysis advocated by Tukey, in which an observation is trimmed or censored if it is sufficiently improbable that it could have arisen from the main population being sampled. My preference would be to regard all item response data as potentially contaminated by random responding and to take steps in the analysis to distinguish, insofar as possible, between informative and non-informative item scores.

Turning now to McBride's paper, I found myself having difficulty accepting at face value the bias curves comparing the Bayesian procedure and the maximum likelihood estimate of test score. In order to obtain biases in the Bayes estimates like those shown in Figure 21, McBride must be assuming a normal prior distribution of ability, which in effect restricts the Bayes estimate, especially at the ends of the distribution. In that case the word "bias" seems unduly prejudicial--both the Bayes and maximum likelihood estimates are valid inferences from the data starting from different assumptions. To plot curves of these estimates conditional on ability, as in the graph, is unfair to the Bayes estimate since, in effect, it takes into account the assumed probability density at each point on the trait continuum.

I also think it is somewhat misleading to show the information supplied by the ordinary test score based on an item-sequential test administration procedure in which all subjects are expected to obtain the same score (but from items of differing difficulty). The information of such scores, which is itself an expected value, is actually zero everywhere. This seems a little too trivial to plot on a graph. The non-zero values shown presumably reflect the imperfect working of the sequential procedure.

In Sympson's paper, I certainly agree with him that there is a need, before plunging headlong into latent trait assumptions and models, to give considerable thought to the plausibility of the assumption that the behavior in question is under the control of an unobservable and continuous latent trait. There may be good reason to do so, but some sort of theoretical justification is needed.

Consider, for example, a vocabulary test. We could estimate vocabulary size in terms of a sample of the number of words that a person has available in his personal lexicon. That would be a perfectly objective, direct way of describing the trait. How does one then justify switching from that intuitively direct concept to an abstract conception of a latent verbal ability? A possible justification that I can think of is that, if vocabulary is to serve as an index of cognitive development generally, we might wish to think in terms of capacity for

acquisition of vocabulary as a developing latent trait of which personal lexicon is a consequence. If so, it is a measure of that continuously developing capacity that we're trying to capture, and the vocabulary itself is just a symptom of that growth. Another justification might be that the latent trait estimate has greater generality and power to predict and account for behavior in other areas. If so, some of that generality and power should be demonstrated and not merely assumed as is so often the case in theoretical presentations of the subject.

Concerning specific criteria for evaluating items, I wish that Sympson had not chosen to omit discussion of some of the preliminaries. I think that, at an early stage in working with an item domain, it is advisable to look at some form of factor analysis of the item intercorrelations. In the past there have been objections to this because of the type of correlation coefficient used. Phi coefficients introduce spurious "difficulty" factors and should be avoided. Tetrachoric correlation coefficients may give a non-positive definite matrix of correlations and thus rule out rigorous factor analysis with a statistical test of the number of factors (although an approximate analysis goes through without difficulty).

But recently, in the March 1975 issue of Psychometrika, Anders Christoffersson has published a technique for a general factor analysis of dichotomous data that overcomes all of these objections and is reasonably practical computationally. His procedure could be used at a first level to verify that the item domain is unidimensional, or to classify items according to dimension. Once the set of items is narrowed down to a unitary domain under the control of a single latent trait, then the psychometric procedures for estimating item parameters and trait values will provide good statistical tests of whether or not the latent trait model holds. There are tests that would distinguish between the homogeneous case where a Rasch model would apply and a heterogeneous case where a more general model--the normal ogive or logistic model--would be required. The technical procedures for making these decisions are in fact available in the form of likelihood ratio tests of alternative models.

Finally, a comment on Nancy Betz's paper. Bob Linn has already made a number of points about that paper, so I will pick up just one aspect of it. I strongly support the point of view that traditional testing is too limited in terms of the sources of information that it exploits in order to assess abilities. But I am also concerned that, in an effort to expand these sources by a shift to computer terminal test administration, we will cut off a very important area of item content, namely the graphics. Much graphic material--half-tone or color pictures, for example--cannot be presented even on CRT displays under computer control. But these visual and non-verbal ways of communicating information should nevertheless be part of the evaluation of ability.

In recent years it has become increasingly clear that cognition is by no means limited to verbal skills. It appears that, in most people's minds, there is going on simultaneously with verbal and logical reasoning based on semantic mediation, a kind of analogical, spatial, non-verbal reasoning capable of solving concrete problems without the aid of the semantic device. The evidence for this is in work that shows the relative specialization of the left and right hemi-

spheres of the brain--the left to semantic processes and the right to spatial or configural processes. Some of this work has been summarized by Jerre Levy in the Proceedings of the 32nd Annual Biology Colloquium.

This work should remind us that if we restrict ourselves to the written or spoken word, we may end up measuring just half the brains of our subjects! Rather than do that, we should demand some type of computer-controlled equipment that is capable of handling visual displays as well as verbal displays. We could imagine stand-alone equipment based on a small mini-computer and some sort of random-access slide file. Slides would be selected under computer control and projected as the item stimuli. We might also want to have auditory display, but I think for most purposes a random access slide file would be sufficient. Prototypes of such equipment already exists, but I do not believe they are available off-the-shelf at non-prohibitive prices.

As education becomes increasingly centered on the individual, it is not unreasonable to assume that a school of any size ought to have some such special equipment for individual evaluation of students. Students could then be sent to the facility for individualized testing under control of the mini-computer at any time that a question arises about their educational progress. If schools had such facilities, the possibility would be open for much more frequent individual testing and monitoring of student progress. That's the direction I would like to see educational testing move in the next decade.

REFERENCES

Angoff, W.H. & Huddleston, E.M. The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test. Princeton, New Jersey, Educational Testing Service, Statistical Report SR-58-21, 1958.

Bayroff, A.G. Feasibility of a programmed testing machine. U. S. Army Personnel Research Office Research Study 64-3, 1964.

Betz, N.E. & Weiss, D.J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973. (AD 768993)

Betz, N.E. & Weiss, D.J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD A001230)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Bock, R.D. & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J. & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Ferguson, R.L. & Hsu, T. The application of item generators for individualizing mathematics testing and instruction. Report 1971/14, University of Pittsburgh Learning Research and Development Center, 1971.

Krathwohl, D.R. & Huyser, R.J. The sequential item test (SIT). American Psychologist, 1956, 2, 419.

Larkin, K.C. & Weiss, D.J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 783553)

Larkin, K.C. & Weiss, D.J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975. (AD A006733)

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-76.

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

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

Lord, F.M. The relative efficiency of two tests as a function of ability level. Psychometrika, 1974, 39, 351-358.

McBride, J.R. Adaptive testing research at Minnesota--Some properties of a Bayesian sequential adaptive mental testing strategy. Paper presented at the Conference on Computerized Adaptive Testing. Washington, D.C., June, 1975.

Owen, R.J. A Bayesian approach to tailored testing. Princeton, N.J.: Educational Testing Service, Research Bulletin RB-69-92, 1969.

Solomon, H. Studies in item analysis and prediction. Stanford, California: Stanford University Press, 1961.

Strang, H.R. & Rust, J.O. The effects of immediate knowledge of results and task definition on multiple-choice answering. The Journal of Experimental Education, 1973, 42, 77-80.

Sympson, J.B. An empirical investigation of a continuous second-stage two-stage testing strategy. Research Report 73-X, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1975 (in preparation).

Weiss, D.J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973. (AD 768376)

Zontine, P.L., Richards, H.C., & Strang, H.R. Effect of contingent reinforcement on Peabody Picture Vocabulary Test Performance. Psychological Reports, 1972, 31, 615-622.

APPENDIX

Technical Information

Data Generation and Analysis . . . . . C. David Vale

The data for the information curves presented were obtained from computer simulations of responses of hypothetical testees to hypothetical items administered under the different testing strategies. In a computer simulation, the computer first simulates a hypothetical testee, by specifying an ability level. Then, that hypothetical testee is "administered" items, and responds according to some mathematical model. For this presentation, assuming no guessing and item discriminations of a=1.0, the response model (i.e., algorithm) was as follows:

First, the probability of a correct response, given the "testee's" ability, was calculated from the following equation:

$$P_i(\Theta)=\Phi(\Theta-b_i) \qquad [5]$$

where $P_i(\Theta)$ = Probability that testee with ability Θ will answer item i correctly.

$\Phi(x)$ = the unit normal distribution integrated from $-\infty$ to the standard deviate, x.

Θ = the ability level of the testee.

b_i = the difficulty of the item.

After the probability of a correct response was determined, a random number was generated from a rectangular distribution between 0 and 1. If this number was greater than the probability of answering the item correctly, the item was considered answered incorrectly; otherwise it was considered correct.

Using data generated in this way, information curves were constructed by calculating information values at several points along the ability continuum from the following formula, suggested by Birnbaum (1968):

$$I_x(\theta) = \left[\frac{\frac{\partial}{\partial\theta} E(X|\theta)}{\sigma_{x|\theta}}\right]^2 \qquad [6]$$

where $I_x(\Theta)$ is the information about Θ provided by score x.

The numerator of Equation 6 may be viewed as a scaling function, converting the score, x, into an ability metric. It is also the partial derivative of the score with respect to ability evaluated at that level of ability, indicating the relative rate of change of the two variables. The denominator is simply the conditional standard deviation of the score, or the dispersion of the score, x, evaluated at a fixed level of ability (i.e., imprecision of measurement).

To calculate information values, 1000 response records were generated at each of 15 equally spaced levels of ability ranging from -3.5 through 0.0 to 3.5. For the middle 13 points, partial derivatives of the score means were calculated with respect to ability at each level of generating ability by taking the derivative of the second degree Lagrangian interpolation polynomial fitted to three successive points. This technique finds the first derivative of the second degree polynomial best fitting the point of interest and the two adjacent points. Because points on each side of the point of interest were needed to estimate the polynomial, the endpoints (i.e., -3.5 and 3.5) were not considered in calculating the information values. When the derivatives were obtained, they were divided by the standard deviation of the scores at the level of ability on which the derivative was centered and then squared to yield the information at that point.

Table A-1 presents the raw information values, for each of the seven strategies, from which the information curves were calculated. To draw the curves, these values were "smoothed" by fitting a cubic polynomial regression curve to them and then plotting the regression function (except for the peaked conventional test's curve which was smoothed by averaging the two sides).

Data Generation and Analysis . . . . . James R. McBride

Item response simulation. Every item g had its discrimination parameter a_g equal to 1.25, and its guessing parameter c_g equal to .20. Each simulated examinee i was characterized by one of 32 discrete values of the trait Θ. One hundred examinees were simulated at each of thirty-two points in the interval [-3.2, +3.0]. For each examinee i, the probability of a correct response to the current simulated test item was calculated by evaluating the 3-parameter logistic function:

$$P(u_g = 1 \mid \theta_i) = P_g(\theta_i) = c_g + (1 - c_g)\left[1 + \exp\{-1.7\, a_g\,(\theta_i - b_g)\}\right]^{-1} \qquad [7]$$

The resulting numerical value $P_g(\Theta_i)$ was compared with a random number R_{gi} from a distribution rectangular in the interval [0,1]. An item score of 1 was assigned if $P_g(\Theta_i) > R_{gi}$; otherwise a score of 0 was assigned. The item difficulty parameter value b_g was determined as described below.

Generation of item response vectors and scores. Owen (1969) gave the Bayesian sequential procedure for selecting test items, and scoring the resulting tests, used in this study. Before administering the initial test item to any examinee i, the method assumes a normal prior distribution on Θ_i, with parameters μ_o, σ_o^2. For every examinee in this study, $\mu_o = 0$ and $\sigma_o^2 = 1$. The parameters μ_g and σ_g^2 were updated after each item response u_g (u_g=1 or 0; here g=1, 2, ... 20). Thus, before administering item g, there was an assumed normal prior distribution on Θ_i, with parameters $\mu_{g-1}(i)$, $\sigma^2_{g-1}(i)$.

Table A-1

Raw Information Values Used To Construct Information Curves

Ability Level	Strategy						
	Rectangular Conventional	Peaked Conventional	Flexilevel	Pyramidal	Two-Stage	Stradaptive	Bayesian
-3.0	3.676	.636	4.582	6.425	6.411	12.301	10.983
-2.5	5.024	1.432	5.222	8.074	7.556	12.943	12.867
-2.0	5.737	3.165	7.253	9.757	8.839	11.806	12.988
-1.5	6.011	5.520	9.257	12.517	10.001	11.966	14.354
-1.0	6.286	9.251	9.510	13.643	11.133	12.055	13.941
-0.5	6.621	13.689	9.485	14.394	11.494	13.030	13.917
0.0	6.301	15.276	11.584	14.525	13.593	11.939	13.118
0.5	6.726	13.521	8.819	14.677	12.530	12.189	14.352
1.0	6.476	8.859	9.024	13.868	10.602	12.120	13.446
1.5	5.923	5.749	7.734	12.038	9.019	12.416	13.753
2.0	5.901	3.775	7.186	10.379	9.038	12.612	12.659
2.5	4.745	1.620	6.307	8.801	8.208	12.866	11.490
3.0	3.617	.654	4.081	6.605	6.392	12.612	10.683

The updated parameters $\mu_g^{(i)}$, $\sigma_g^2(i)$ were those of the posterior distribution on θ_i. After each step g the next item to be administered was selected from those in the item pool so as to minimize $E(\sigma_{g+1}^2)$. The Bayesian "test score" (which is an estimator of θ_i) assigned was the latest posterior value μ_g. In this study, twenty items were administered to each examinee i, so in each case the Bayesian test score $\hat{\theta}_i^B = \mu_{20}^{(i)}$, the posterior mean after twenty items were administered.

The method of updating μ_g and σ_g^2 is contingent on the item response. For a correct response to item g the updated prior was set equal to the posterior value μ_g, such that

$$\mu_g = \mu_{g-1} + \{1 - c_g\} \left| \frac{\sigma_{g-1}^2}{\sqrt{a_g^{-2} + \sigma_{g-1}^2}} \right| \left| \frac{\phi(D)}{c_g + \{1 - c_g\} \Phi(-D)} \right| \qquad [8]$$

and its variance is

$$\sigma_g^2 = \sigma_{g-1}^2 \left[1 - \left| \frac{1 - c_g}{\sqrt{1 + a_g^{-2} \sigma_{g-1}^{-2}}} \right| \left| \frac{\phi(D)}{A} \right| \left| \frac{(1 - c_g)\, \phi(D)}{A} - D \right| \right] \qquad [9]$$

Following an incorrect response the corresponding expressions are

$$\mu_g = \mu_{g-1} - \left| \frac{\sigma_{g-1}^2}{\sqrt{a_g^{-2} + \sigma_{g-1}^2}} \right| \left| \frac{\phi(D)}{\Phi(D)} \right| \qquad [10]$$

and

$$\sigma_g^2 = \sigma_{g-1}^2 \left[1 - \left| \frac{\phi(D)}{\sqrt{1 + a_g^{-2} \sigma_{g-1}^{-2}}} \right| \left| \frac{\frac{\phi(D)}{\Phi(D)} + D}{\Phi(D)} \right| \right] \qquad [11]$$

In equations 8 through 11:

ϕ (D) is the normal probability density function;

Φ (D) is the cumulative normal density function;

$$D = \{b_g - \mu_{g-1}\} / \sqrt{a_g^{-2} + \sigma_{g-1}^2}$$

$$A = c_g + \{1 - c_g\} \Phi(-D)$$

In this simulation study, at every step g, the item difficulty b_g was set equal to the current prior mean, which is μ_{g-1}. Thus at every step g the value D=0. This considerably simplifies evaluating Equations 8 through 11. Although it is an artifice, it was done here in order to "purify" the results, whose generality would be restricted if a typical finite item pool were used.

Table A-2. Sample means ($\bar{X}$) and conditional variances (S^2) of test scores from three different methods of scoring a simulated 20-item Bayesian adaptive test.

	Scoring Method					
	Owen's Bayesian		Maximum Likelihood		Number Correct	
Θ	$\bar{X}$	S^2	$\bar{X}$	S^2	$\bar{X}$	S^2
-3.2	-2.197	.203	-2.997	1.095	5.04	2.418
-3.0	-2.174	.212	-2.799	.516	5.54	2.028
-2.8	-2.118	.168	-2.747	.419	5.72	1.642
-2.6	-2.057	.118	-2.603	.442	6.30	2.110
-2.4	-1.887	.126	-2.352	.404	6.72	2.242
-2.2	-1.775	.126	-2.162	.421	7.35	2.428
-2.0	-1.756	.074	-2.015	.192	7.68	1.518
-1.8	-1.590	.071	-1.796	.197	8.46	1.848
-1.6	-1.373	.073	-1.601	.288	8.64	2.290
-1.4	-1.263	.067	-1.366	.096	9.31	1.754
-1.2	-1.097	.094	-1.167	.135	9.79	2.506
-1.0	- .960	.072	- .996	.100	10.23	1.777
- .8	- .762	.073	- .772	.092	10.65	2.088
- .6	- .569	.065	- .556	.072	11.07	2.005
- .4	- .386	.060	- .366	.070	11.68	2.198
- .2	- .189	.086	- .147	.097	11.89	1.678
0	- .034	.071	.010	.082	12.11	1.918
.2	.211	.089	.278	.096	12.30	1.910
.4	.389	.082	.458	.087	12.81	1.814
.6	.549	.068	.619	.071	13.27	1.697
.8	.729	.085	.816	.092	13.21	2.046
1.0	.940	.070	1.031	.079	13.87	1.673
1.2	1.141	.086	1.248	.096	14.01	1.630
1.4	1.290	.099	1.411	.114	14.15	1.528
1.6	1.506	.082	1.637	.096	14.51	1.090
1.8	1.705	.087	1.860	.104	14.68	1.638
2.0	1.887	.067	2.056	.080	15.32	1.258
2.2	2.042	.068	2.224	.086	15.52	1.010
2.4	2.209	.080	2.420	.101	15.88	1.126
2.6	2.442	.068	2.687	.090	16.43	1.065
2.8	2.536	.096	2.797	.120	16.67	1.341
3.0	2.784	.081	3.060	.108	17.07	.965

Let $\hat{\theta}_i^{ML}$ be defined as the maximum likelihood score (or estimator). This test score was calculated after each examinee's vector of twenty item scores was generated, and was based on the pattern of item scores, and the likelihood function

$$L = P(u_1, u_2, \ldots u_{20} | \theta) = \prod_{g=1}^{20} P_g(\theta)^{u_g} [1 - P_g(\theta)]^{1-u_g} \qquad [12]$$

Newton-Raphson iteration was used to solve for the approximate value of θ at which L was maximal.

The number-correct score for person i was simply the sum of the item scores u_g:

$$X_i = \sum_{g=1}^{20} u_{gi} \qquad [13]$$

Unlike the Bayesian and maximum likelihood scores, X_i is not strictly an estimator of θ_i. However, since number correct scores are usually used to estimate ability, it is not inappropriate to treat X_i as an estimator of θ_i, in order to study the bias in X_i. Therefore let $\hat{\theta}_i^X = X_i$.

Estimation of test score regressions, and information curves. Let the symbol X refer generally to any test score (or trait level estimate) $\hat{\theta}$. For each of the three scoring methods used in this study, the conditional mean and variance of the 100 scores X at each of the thirty-two θ levels were recorded. These data are listed by scoring method in Table A-2, and were used to generate the plots shown in Figures 19 and 20.

For each scoring method, the regression of X on θ was estimated by fitting a third-degree polynomial to the thirty-two observed mean scores, using a least-squares regression program to calculate the regression coefficients. The coefficients obtained are listed by scoring method in Table A-3; the regression equations were of the form

$$\hat{X} = a_0 + a_1\theta + a_2\theta^2 + a_3\theta^3 \qquad [14]$$

The slope of each regression equation was estimated simply by evaluating the first derivative of each, at every θ-point sampled in the interval $-3 \leq \theta \leq +3$.

Calculation of the test score information at each of the sampled θ-points was based on Equation 6. Numerical values shown in Table A-4 were derived by dividing the square of the regression slope by the smoothed score variance estimate at each θ point in that same interval.

That is, $I_X(\theta)$ was estimated for each scoring method by

$$\hat{I}_X(\theta) = \left[\frac{\text{SLOPE}}{\hat{\sigma}_{X|\theta}}\right]^2 \qquad [15]$$

Table A-3. Estimated coefficients of the third-degree polynomial equations for the regression of test score X on simulated trait level θ, for three different scoring methods.

Coefficient	Scoring Method		
	Owen's Bayesian	Maximum Likelihood	Number Correct
a_0	-.0302	.0217	12.0410
a_1	.9528	1.0237	1.7623
a_2	.0338	.0055	- .1020
a_3	-.0150	- .0046	.0192
Variance accounted for:	.9996	.9997	.9976

In estimating $\sigma^2_{x|\theta}$ for Equation 15 the method of "moving averages" was used to "smooth" the sample estimates at each θ-point in the interval [-3, +3]. The resulting estimates, $\hat{\sigma}^2_{x|\theta}$, are listed in Table A-4, along with the slope and information value estimates, for each θ-point in the interval. Figure 23 is a plot of the information estimates listed in Table A-4.

Table A-4. Slope of the estimated regression of test scores on trait level, the smoothed estimated conditional variance $\hat{\sigma}^2_{x|\theta}$ of test scores, and the estimated information $\hat{I}_x(\theta)$ in the test scores, for each of three methods of scoring a 20-item Bayesian adaptive test.

	Scoring Method								
	Owen's Bayesian			Maximum Likelihood			Number Correct		
θ	Slope	$\hat{\sigma}^2_{x\|\theta}$	$\hat{I}_x(\theta)$	Slope	$\hat{\sigma}^2_{x\|\theta}$	$\hat{I}_x(\theta)$	Slope	$\hat{\sigma}^2_{x\|\theta}$	$\hat{I}_x(\theta)$
-3.0	.341	.194	.58	.866	.677	1.11	2.894	2.029	4.13
-2.8	.407	.166	1.00	.884	.459	1.70	2.786	1.927	4.03
-2.6	.469	.137	1.61	.901	.422	1.92	2.683	1.998	3.60
-2.4	.528	.123	2.27	.917	.422	1.99	2.584	2.260	2.95
-2.2	.583	.108	3.15	.932	.339	2.56	2.490	2.063	3.01
-2.0	.634	.090	4.47	.946	.270	3.31	2.401	1.931	2.99
-1.8	.682	.073	6.37	.959	.226	4.07	2.316	1.885	2.85
-1.6	.726	.070	7.53	.971	.194	4.86	2.236	1.964	2.55
-1.4	.767	.078	7.54	.981	.173	5.56	2.161	2.183	2.14
-1.2	.804	.078	8.29	.990	.110	8.91	2.090	2.012	2.17
-1.0	.837	.080	8.76	.999	.109	9.16	2.024	2.124	1.93
- .8	.867	.070	10.74	1.006	.088	11.50	1.962	1.957	1.97
- .6	.893	.066	12.08	1.012	.078	13.13	1.905	2.097	1.73
- .4	.916	.070	11.99	1.017	.080	12.93	1.853	1.960	1.75
- .2	.935	.072	12.14	1.021	.083	12.56	1.805	1.931	1.69
0	.950	.082	11.01	1.024	.092	11.40	1.762	1.835	1.69
.2	.962	.081	11.43	1.025	.088	11.93	1.724	1.880	1.58
.4	.970	.080	11.76	1.025	.085	12.36	1.690	1.807	1.58
.6	.974	.078	12.16	1.025	.083	12.66	1.661	1.852	1.49
.8	.975	.074	12.84	1.024	.081	12.95	1.636	1.805	1.48
1.0	.972	.080	11.81	1.021	.089	11.71	1.616	1.783	1.46
1.2	.966	.085	10.98	1.017	.096	10.59	1.601	1.610	1.59
1.4	.956	.089	10.27	1.012	.102	10.04	1.590	1.416	1.79
1.6	.943	.089	9.99	1.001	.105	9.54	1.584	1.419	1.77
1.8	.925	.079	10.83	.998	.093	10.71	1.582	1.329	1.88
2.0	.905	.074	11.07	.990	.090	10.89	1.585	1.302	1.93
2.2	.880	.072	10.76	.980	.089	10.79	1.593	1.131	2.24
2.4	.852	.072	10.08	.970	.092	10.23	1.605	1.067	2.41
2.6	.821	.081	8.32	.958	.104	8.82	1.622	1.177	2.24
2.8	.785	.082	7.51	.945	.106	8.42	1.643	1.124	2.40
3.0	.746	.082	6.79	.931	.108	8.03	1.669	1.153	2.42