APPLICATIONS OF COMPUTERIZED ADAPTIVE TESTING

Proceedings of a Symposium Presented
at the
18th Annual Convention
of the
Military Testing Association
October 1976

ISAAC I. BEJAR
JAMES R. McBRIDE
STEVEN M. PINE
JAMES B. SYMPSON
C. DAVID VALE

edited by

DAVID J. WEISS

RESEARCH REPORT 77-1

MARCH 1977

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

Prepared under the following contracts
with the
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research:

N00014-76-C-0243, NR No. 150-382
N00014-76-C-0244, NR No. 150-383
N00014-76-C-0627, NR No. 150-389

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Research Report 77-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Applications of Computerized Adaptive Testing		5. TYPE OF REPORT & PERIOD COVERED: Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): James R. McBride, James B. Sympson, C. David Vale, Steven M. Pine, and Isaac I. Bejar. Edited by David J. Weiss		8. CONTRACT OR GRANT NUMBER(s): See 18 below
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: See 18 below
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington, Virginia 22217		12. REPORT DATE: March 1977 13. NUMBER OF PAGES: 54
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 1976 Meeting of the Military Testing Association. Research reported herein was sponsored by the following contracts:
N00014-76-C-0243, P.E. 61153N, PROJ. RR042-04, T.A. RR042-04-01, W.U. NR150-382;
N00014-76-C-0244, P.E. 61153N, PROJ. RR042-04, T.A. RR042-04-01, W.U. NR150-383;
N00014-76-C-0627, P.E. 61153N, PROJ. RR042-04, T.A. RR042-04-01, W.U. NR150-389.

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

testing	branched testing	automated testing
ability testing	individualized testing	test bias
computerized testing	tailored testing	test fairness
adaptive testing	programmed testing	achievement testing
sequential testing	response-contingent testing	performance testing

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

This symposium consisted of five papers:

1. James R. McBride: A Brief Overview of Adaptive Testing

Adaptive testing is defined, and some of its item selection and scoring strategies briefly discussed. Item response theory, or item characteristic curve theory, which is useful for the implementation of adaptive testing is briefly described. The concept of "information" in a test is introduced and discussed in the context of both adaptive and conventional tests. The advantages of adaptive testing, in terms of the nature of information it provides, are described.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

2. James B. Sympson: Estimation of Latent Trait Status in Adaptive Testing Procedures

The role of latent trait theory in measurement for criterion prediction and in criterion-referenced measurement is explicated. It is noted that latent trait models allow both normed-referenced and criterion-referenced interpretations of test performance. Using a 3-parameter logistic test model, an example of sequential estimation in a 20-item adaptive test is presented. After each item is administered, four different ability estimates (two likelihood-based and two Bayesian estimates) are calculated. Characteristics of the four estimation methods are discussed. The information available in the items selected by the adaptive test is compared with the information available from comparable "rectangular" and "peaked" non-adaptive tests. The joint application of latent trait theory and adaptive testing is advocated as a useful approach to human assessment.

3. C. David Vale: Adaptive Testing and the Problem of Classification

The use of adaptive testing procedures to make ability classification decisions (i.e., cutting score decisions) is discussed. Data from computer simulations comparing conventional testing strategies with an adaptive testing strategy are presented. These data suggest that, although a conventional test is as good as an adaptive test when there is one cutting score at the middle of the distribution of ability, an adaptive test can provide better classification decisions when there is more than one cutting score. Some utility considerations are also discussed.

4. Steven M. Pine: Applications of Item Characteristic Curve Theory to the Problem of Test Bias

It is argued that a major problem in current efforts to develop less biased tests is an over-reliance on classical test theory. Item Characteristic Curve (ICC) Theory, which is based on individual rather than group-oriented measurement, is offered as a more appropriate measurement model. A definition of test bias based on ICC theory is presented. Using this definition, several empirical tests for bias are presented and demonstrated with real test data. Additional applications of ICC theory to the problem of test bias are also discussed.

5. Isaac I. Bejar: Applications of Adaptive Testing in Measuring Achievement and Performance

The paper reviews two relatively recent developments in psychometric theory, the assessment of partial knowledge and research in adaptive testing. It is argued that the use of non-dichotomous item formats, needed for the assessment of partial knowledge, and now made possible by the administration of achievement test items on interactive computers, should result in achievement test scores which are a more realistic and precise indication of what a student can do.

CONTENTS

Technical editing by Terryl Graham.

APPLICATIONS OF COMPUTERIZED ADAPTIVE TESTING

A BRIEF OVERVIEW OF ADAPTIVE TESTING

JAMES R. McBRIDE
U. S. Army Research Institute for the Behavioral and Social Sciences

This symposium will present some recent developments in adaptive testing which have applications to several military testing problems. The purpose of this overview is to provide a brief introduction to adaptive testing--what it is, what is needed to implement it, and why it is of interest.

"Adaptive" testing is one of a number of terms used to describe a procedure whereby the test items that comprise an individual's test are selected during the test itself. Some of the other terms used interchangeably with adaptive testing include tailored testing, branched testing, programmed testing, and individualized testing. The term "adaptive" was chosen because these tests adapt themselves to the examinee; different persons answer different items, with the items chosen sequentially to suit the individual examinee's performance.

Differential selection of test items may be accomplished in any number of ways. But, generally, in adaptive tests a more difficult item is administered following each correct answer, and an easier item following an incorrect one. Some methods of adaptive testing have been implemented in paper-and-pencil mode; for example, Lord's (1971) flexilevel adaptive test was designed specifically for paper-and-pencil administration. However, experience has shown that the instructions for paper-and-pencil adaptive tests are too complex for some examinees to follow successfully (Weiss & Betz, 1973, p. 23) A more satisfactory mode of administration is through use of an interactive computer terminal or similar device. Thus, Weiss (1976) chose to administer adaptive tests at a cathode-ray terminal (CRT); Bayroff, Ross and Fischl (1974) reported the Army's development of a computer-controlled slide projection terminal for adaptive testing; Waters (1977) designed and built a micro-processor terminal which directs the examinee through an adaptive sequence of test items read from a printed booklet.

Item selection strategies. Because adaptive tests are quite different from conventional tests in which all examinees must answer the same set of test items, adaptive testing poses some new psychometric problems. One problem is how to choose successive items from the pool of available items. This problem can be solved through an item selection strategy, which defines a formalized rule for item choice.

Numerous item selection strategies are possible. They vary from very simple two-branch rules to rules based on the optimization of rather complex mathematical functions (Weiss, 1974). Obviously, computerizing the item-selection process facilitates the use of the mathematical optimization procedures.

Scoring adaptive tests. Since different examinees take sets of test items which may differ in number, difficulty, and discriminating power, the traditional number correct score will not suffice to order people on most adaptive tests. Some scoring procedure is required which will consider not only *how many* items were answered correctly, but also *which* items were taken, and the *pattern* of right and wrong answers to those items. The scoring procedures most widely used in adaptive testing are based on various formulations of latent trait theory (e.g., Birnbaum, 1968; Lord, 1952, 1974; Rasch, 1960). All of these formulations provide statistical methods for locating examinees on a common scale, even though they responded to different sets of test items.

Item response theory. Because of the unique characteristics of adaptive tests--tailoring each test to the individual and locating all examinees on a common scale despite the different items constituting each test--traditional test theory is inadequate for use in adaptive testing. "Latent trait" or "item response" theory (Lord, 1952, 1976) provides an adequate theoretical basis for the development of adaptive testing.

Item response theory, also known as item characteristic curve theory, is a general term for theoretical formulations which account for examinees' responses to test items in terms of their status on an underlying attribute. In ability (or achievement) testing, the higher the attribute status, the larger is the probability of a correct response to any given item which measures the trait in question. Through appropriate scaling procedures, a response curve can be constructed for every such test item. This item characteristic curve (ICC) expresses the probability of a correct response as a mathematical function of the scaled trait and the item characteristics.

Every person can be characterized by his/her location on this scale. Every test item also has a location parameter (its threshold, or "difficulty") and perhaps its own rate parameter (proportional to the steepness of the ICC), analogous to its discriminating power. Some items also have a lower asymptote, or guessing parameter.

Knowing which items a person has answered; the difficulty, discrimination, and guessing parameters of those items; and whether the answers were correct or incorrect permits the use of the statistical techniques of item response theory to estimate the examinee's ability. The resulting ability estimate is a "test score" of sorts which has an error component like any other observed score. Unlike classical test theory, item response theory makes no assumption that measurement errors are independent of "true score", which is appropriate because this central assumption of classical test theory is untenable (Lumsden, 1976). Whether ability is defined as "true score" or as location on a latent continuum, errors of measurement can vary at different levels of the trait, reflecting in part the discrepancy between examinee trait level and the difficulties of the test items.

Information. Item response theory permits the evaluation of something closely akin to the standard error of measurement as a function of underlying ability, if the test item parameters are known. This is called the test information function (Birnbaum, 1968) which is inversely proportional to the standard error of estimating an examinee's location on the trait scale. If the information function of a

typical peaked conventional test (one whose items are all about equal in difficulty) were plotted, its test information function would likewise be peaked--very high over a narrow range of the trait, but diminishing in magnitude elsewhere. Such a test will discriminate very well over a narrow interval of the trait range; it will not discriminate as well outside that interval. The ability level at which the test information function is highest can be referred to as the test "center".

The information function of a "rectangular" conventional test (one whose item difficulties are uniformly distributed over a wide range) is fairly flat, but low over a broad interval on the trait scale around the test center. This test would measure about equally well over a much wider range than the peaked test, but other things being equal, would not discriminate nearly as effectively as does the peaked test at its center.

The design of conventional tests. A test measures best (most precisely) where its information function is highest (and hence its standard error is lowest). It is frequently desirable to have high measurement precision over most of the normal range of the attribute we seek to measure. This is tantamount to a high, flat information function. Conventional testing, however, presents a dilemma. A peaked test can be constructed which yields an information function with a high peak; or at the other extreme, a rectangular test can be built which has a low, flat information function. A test with a high, flat information function cannot be constructed for conventional test administration unless it is extremely long.

This problem can be referred to as a "bandwidth-fidelity dilemma", with apologies to Cronbach (1961), who described a different "bandwidth-fidelity dilemma". The designer of a conventional test can construct it to have high "fidelity"--high precision, low measurement error--over a *narrow* range of ability; or to have a broad "bandwidth"--equiprecision of measurement over a *wide* range of ability, at the expense of fidelity. In designing a conventional test, there is a tradeoff between broad bandwidth and high fidelity; the designer cannot have both.

Adaptive testing. Herein resides the most attractive feature of adaptive tests from a psychometric point of view: Because the test is adapted to the individual, the discrepancy between trait level and item difficulty can be made both small and fairly constant across the trait range. The result is a flat information function which is also generally high. Adaptive tests--and only adaptive tests--are capable of accurate, equiprecise measurement over a wide ability range. This should pay dividends in test reliability, criterion-related validity, and in the general utility of the test for a broad range of measurement and decision applications.

A properly designed adaptive test will have higher reliability than a conventional test of the same length. As a corollary to that, an adaptive test can achieve a specified level of reliability in substantially fewer items than can a conventional test, thus permitting the measurement of additional attributes in the time saved. Both improved reliability and additional measurements should result in an increment in predictive validity over that obtained using conventional tests.

In addition to the psychometric benefits accruing from the use of adaptive tests, there are psychological benefits to the examinees. Adaptive tests can have

positive effects on the test-taking motivation of examinees (Betz & Weiss, 1976b) and, for some testees, on their measured ability levels (Betz & Weiss, 1976a). By tailoring test difficulty to examinee ability, adaptive tests can reduce the effects of guessing among low-ability examinees and make any remaining effects relatively constant across ability levels.

Summary

This overview has presented a rather broad-brush introduction to adaptive testing. Hopefully, it has conveyed some conception of what adaptive testing is, of the rudiments of the test theory supporting it, and of the significant psychometric and psychological advantages that can accrue when a well-designed adaptive testing program is implemented in a mental-measurement setting. The four principal papers in this symposium will deal in more detail with some methods used in conjunction with adaptive testing, and with a variety of areas of application of adaptive tests which are relevant to the needs and problems of test users in the military.

ESTIMATION OF LATENT TRAIT STATUS IN ADAPTIVE TESTING PROCEDURES

JAMES B. SYMPSON
University of Minnesota

During the last few years, latent trait theory has become increasingly important as a theoretical foundation for the practice of psychological and educational assessment. This has been due to shortcomings inherent in classical test theory (Lumsden, 1976) and to recent developments in testing practice. In particular, when "adaptive" or "individualized" testing is desired, latent trait theory provides a particularly useful conceptual scheme for guiding test design and test scoring procedures.

Latent trait theories are characterized by a mathematical model that relates the probability of occurrence of a particular response class (e.g., a "correct" response) in the presence of a particular stimulus (e.g., a test item) to a person's position on one or more metric dimensions. The graph of the function that relates probability of a particular response class to a person's status on these dimensions can be referred to as a *response-characteristic surface*.

Both univariate and multivariate latent trait models have been proposed. The univariate models (e.g., Birnbaum, 1968; Bock, 1972; Lord, 1952; Rasch, 1960) assume that response probabilities are related to the relative positions of persons and stimuli on a single metric dimension. Multivariate models (e.g., Christoffersson, 1975; Samejima, 1974) allow for the possibility of several latent dimensions.

Latent Trait Theory and the Objectives of Measurement

When they first encounter latent trait theory, many people question its practical utility. For example, they often ask, "Why should I bother with an approach to testing that involves inferred latent traits if what I'm really interested in is either predicting some criterion accurately or achieving content validity and implementing criterion-referenced measurement?" In order to motivate an interest in latent trait estimation procedures, it will be useful to discuss briefly the issues raised by this type of question.

The "existence" of latent traits. The adoption of latent trait theory as a guide to test construction and test scoring does not require a belief in the "existence" of unobservable traits that control human behavior. Empirically, it is sufficient to inquire whether peoples' responses to test stimuli can be predicted accurately on the basis of such a model. The postulated dimensions of latent trait theory can be viewed as quantitative variables that are created by calibrating and scoring test items in a certain way. These variables can provide a convenient basis for designing testing procedures and may lead to increased predictive accuracy in scientific and practical applications.

This research is supported by contract N00014-76-C-0243, NR150-382, with the Personnel and Training Research Programs, Office of Naval Research.

Measurement for criterion prediction. In many situations, tests are developed and applied with the sole intention of predicting performance on a criterion of interest. The introduction of intervening variables (latent traits) might seem unnecessary when one is only interested in obtaining a high degree of relationship between test scores and criterion scores. However, estimates of latent trait status can themselves be viewed as a particular variety of test score. Such scores may or may not have higher predictive validity than more conventional test scores; this is an empirical question. But,even if predictive validity is not increased via the use of latent trait scores, it may still be advantageous to adopt a latent trait approach if the testing process can be made more efficient as a result (e.g., through adaptive testing procedures).

Moreover, test development for the purpose of criterion prediction is always based upon an implicit structural model. No one chooses items at random from all conceivable item domains. Test developers try out items with certain kinds of content and never consider using other kinds of content. They also attempt to generate items that have difficulty levels or endorsement rates (i.e., p-values) that are not too extreme in the population to be tested. This is done so that item-criterion correlations will not be unduly restricted. Such procedures suggest the existence of an implicit structural model.

Trying certain types of items, and not others, implies that certain types of inter-person differences exist and are related to criterion performance, while others are not. More generally, any conceptual scheme for classifying test items implies a corresponding set of response variables that can be generated when the items are administered. In selecting items for criterion prediction the test developer indicates the response variables that are thought to be related to the criterion.

A concern about item difficulties and endorsement rates implies that the probability of a given response to an item is a function of status on the relevant response variable(s). If such probabilities were not a function of status on the response variables, an item would have the same p-value in every conceivable population and there would be no need to match item difficulties to the population that is to be tested.

A latent trait approach to test construction and scoring provides a formal vehicle for elaborating structural models and encourages the test developer to make structural assumptions explicit. When structural models are explicitly stated, they can serve to guide test construction efforts and aid in the interpretation of empirical results.

Content validity and criterion-referenced measurement. The testing situation never constitutes the entire behavioral domain of interest. The implicit objective in pursuing content validity and in implementing criterion-referenced measurement is to make more accurate inferences about a person's potential for performance in a hypothetical task domain (Cronbach, 1971, p. 452; Glaser & Nitko, 1971, p. 653). This hypothetical task domain, though it is not observable in its entirety, is carefully defined in terms of performance objectives or item content. Test items are generated that represent the domain, and responses to these items are used as a basis for making inferences about domain performance.

Some individuals protest such a view and argue that in criterion-referenced measurement the test stimuli are the criterion tasks of interest and that no further task domain is intended or implied. However, unless all the tasks that are required on the job are included in the test, inferences are necessarily being made about a larger task domain from a sample of person-stimulus interactions drawn from the domain.

What is the nature of the hypothetical task domain in achievement testing? Such task domains can be described in terms of a multidimensional structural model. Whenever test stimuli can be clustered with regard to common content or process and arranged in a learning hierarchy within each cluster, there is a definite possibility that a latent trait approach to achievement testing will be useful.

Norm-referenced and criterion-referenced interpretations of test performance. In recent years, the distinction between norm-referenced and criterion-referenced measurement has been widely discussed. An important fact to keep in mind is that this distinction properly applies to the type of information available from test scores, not to test content or the testing procedure itself (Hambleton & Novick, 1973, p. 162). This is important because estimates of latent trait status can provide information about both inter-person differences (norm-referenced interpretations) and intra-person response probabilities (criterion-referenced interpretations) for tasks drawn from a task domain.

An estimate of an individual's latent trait status can be converted to a centile rank or standard score relative to any norm group previously tested using the latent trait procedure. This same latent trait estimate, when considered in conjunction with the latent trait parameters of a test item (i.e., a task sample) that has been previously calibrated, allows generation of the probability of occurrence of a given response class (e.g., a "correct" response) in the presence of the item. (That is, one can determine the probability that a person will complete a given task successfully, even though the person has never attempted the task.) The fact that latent trait theory can provide both norm-referenced and criterion-referenced interpretations of test performance indicates that the current schism between psychological and educational testing may be narrowed considerably in the years to come.

Estimating Latent Trait Status

In order to exploit the wide range of potential applications of latent trait theory, it is necessary to understand procedures for estimating latent trait status of individual testees. Four methods for obtaining estimates of latent trait status are described below. In addition, it will be shown that the accuracy of such estimates can often be improved through the use of adaptive testing procedures.

The latent trait model to be described is one in which only two response classes are considered, a *keyed* response and a *non-keyed* response, and the probability of occurrence of each response class is a function of a single latent dimension. This model might be applicable to a test that has been constructed to maximize internal consistency (Nunnally, 1967, pp. 254-268) and in which items are scored dichotomously. The model would not be suitable for tests that involve a multidimensional item structure, but the principles of latent trait estimation that are discussed can be generalized to such cases.

The Three-parameter Logistic Model

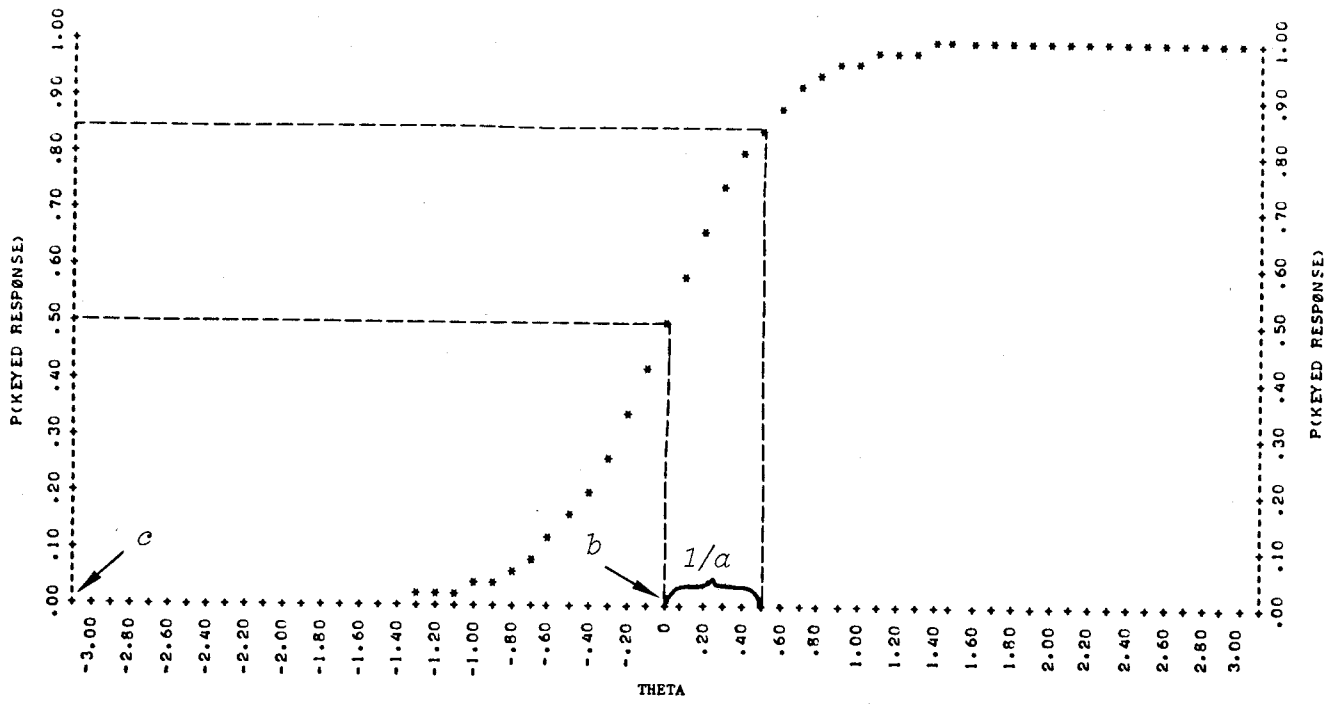
This latent trait model has been investigated extensively by Birnbaum (1968). The function rule that relates probability of a keyed response to the parameters of the model is given in Equation 1.

$$P_g(\theta) = c_g + (1-c_g)[1+\exp(-1.7a_g(\theta-b_g))]^{-1} \qquad [1]$$

The quantity $P_g(\theta)$ is the probability of a keyed response to item g, with parameters a_g, b_g and c_g, by a person whose location on the latent trait continuum is given by the quantity θ (theta). The exponential operator (exp) indicates that the quantity in parentheses is an exponent of the constant $e \doteq 2.71828$.

Figure 1 shows a graph of the function $P_g(\theta)$ in the interval from $\theta=-3.00$ to $\theta=+3.00$ for an item having $a_g=2.0$, $b_g=0.0$, and $c_g=.00$. This graph was generated by evaluating $P_g(\theta)$ at 61 points along the theta continuum. The irregularities visible in Figure 1 result from rounding $P_g(\theta)$ to the nearest .02 for plotting purposes.

Figure 1
Response Characteristic Curve (a=2.0, b=0.0, c=.00)

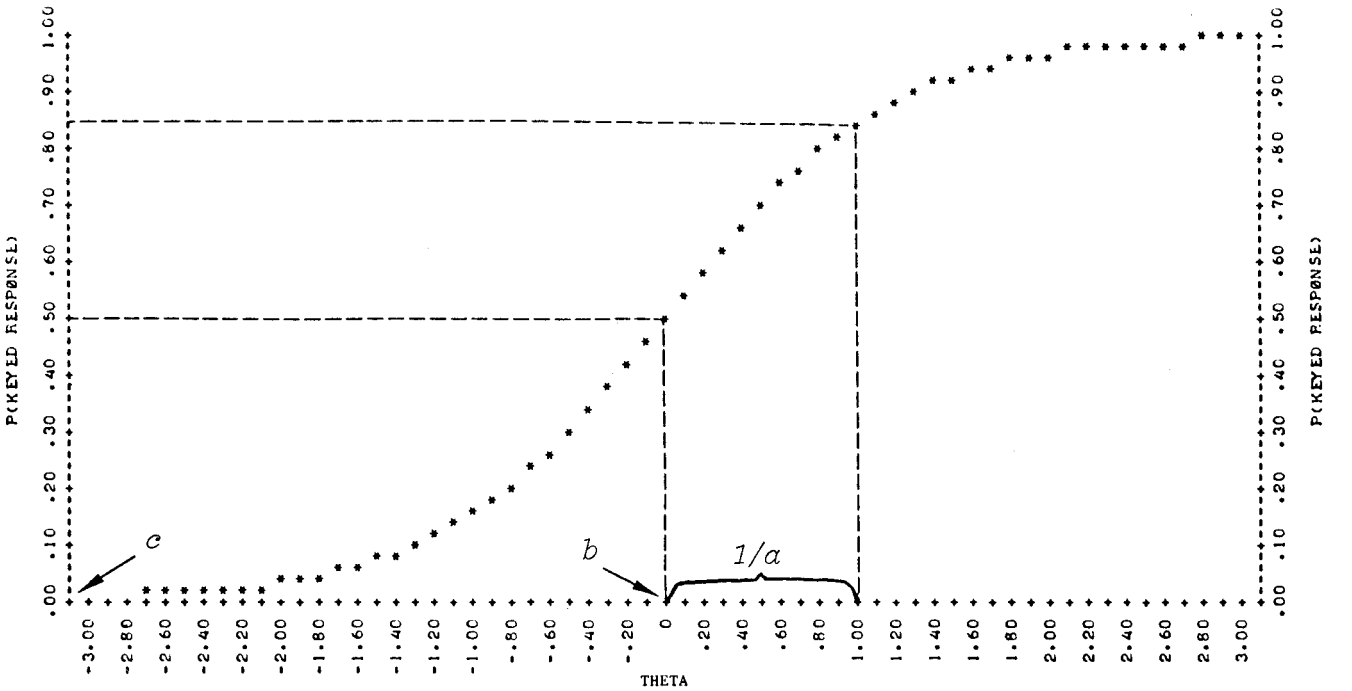


The item parameter c_g is the value of $P_g(\theta)$ when $\theta=-\infty$. It is the lower asymptote of $P_g(\theta)$ and is usually conceived of as the probability of a keyed

response occurring "by chance" when $\theta=-\infty$. The item parameter b_g is known as the *item location parameter*; it indicates the location on the latent trait continuum at which $P_g(\theta)$ is equal to $.5(1+c_g)$. The item parameter a_g is known as the *item discrimination parameter*. It is related to the slope of the response characteristic curve and in this model is equal to the reciprocal of the distance that one must move along the theta continuum in order to increase $P_g(\theta)$ from $.5(1+c_g)$ to approximately $(.8455(1-c_g))+c_g$. Since $a_g=2.0$ and $c_g=.00$ in Figure 1, the distance between the locations on the theta continuum at which $P_g(\theta)=.5$ and $P_g(\theta)\doteq.84$ is equal to $1/a_g=.50$ theta units.

Figure 2 shows a response characteristic curve for an item having $a_g=1.0$, $b_g=0.0$, and $c_g=.00$. The reduced value of a_g, relative to Figure 1, is reflected in the shallower slope of this graph and in the fact that the distance between the locations at which $P_g(\theta)=.50$ and $P_g(\theta)\doteq.84$ is now equal to $1/a_g=1.00$ theta

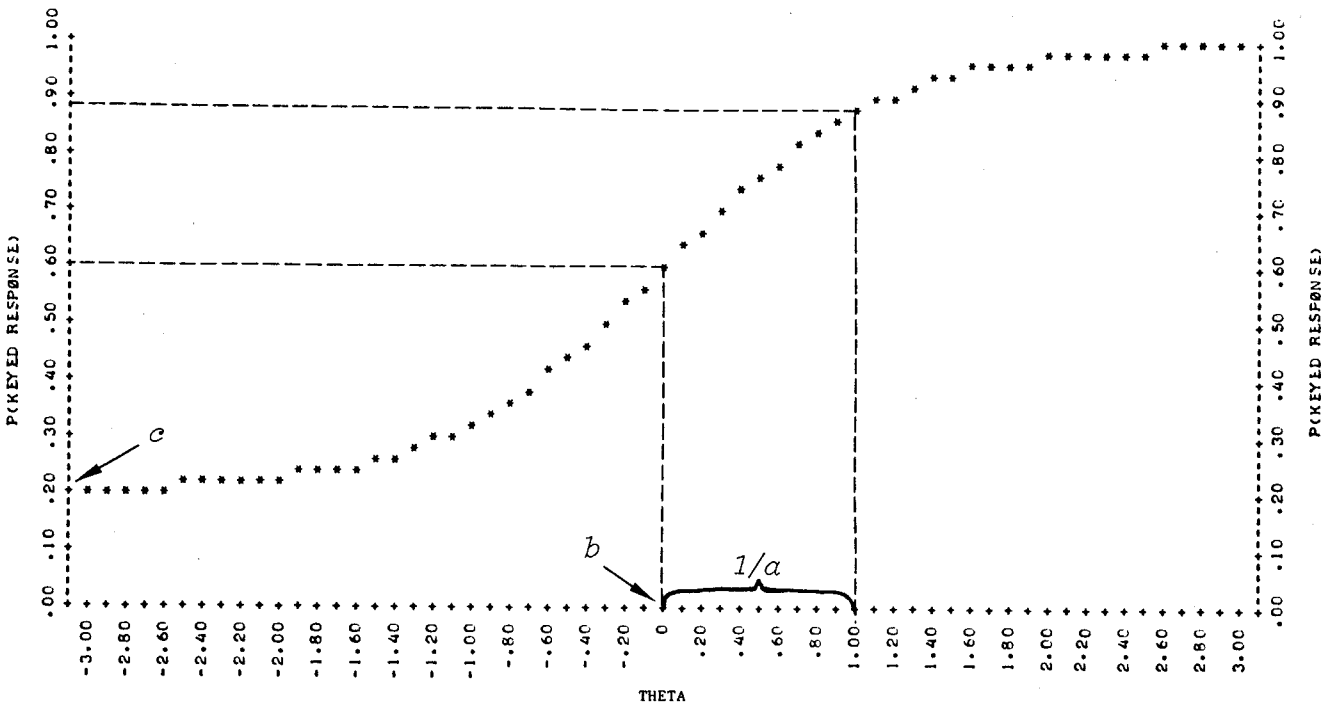
Figure 2
Response Characteristic Curve (a=1.0, b=0.0, c=.00)



units. A value of a_g in the vicinity of 1.0 is typical of many test items. Values of a_g below about .5 are indicative of "poor" items and values of a_g above 2.0, while desirable in many applications, are not common.

Figure 3 shows a response characteristic curve for an item having a_g=1.0, b_g=0.0, and c_g=.20. The value c_g=.20 might be applicable to a multiple-choice test item that has five response alternatives. In accord with the definitions given above, b_g is equal to the location at which $P_g(\theta)$=.5(1+.2)=.60 and a_g is equal to the reciprocal of the distance from the location at which $P_g(\theta)$=.60 to the location at which $P_g(\theta)\doteq$(.8455(1-.2))+.2$\doteq$.88. Note that one of the effects of a non-zero c_g is to reduce the slope of $P_g(\theta)$ at all points along the theta continuum.

Figure 3
Response Characteristic Curve (a=1.0, b=0.0, c=.20)



The Concept of "Information"

Birnbaum (1968) has discussed the concept of "information" available in a test item. Birnbaum's *item information function* is given in Equation 2.

$$I(\theta, u_g) = [P_g'(\theta)]^2/[P_g(\theta)\ Q_g(\theta)] \qquad [2]$$

In this equation, u_g is the *item response variable*. It is equal to 1 when a keyed response is emitted and is equal to 0 otherwise. The quantity $Q_g(\theta)$ is

equal to $1-P_g(\theta)$. The numerator of Equation 2 is the squared first derivative (i.e., the squared slope) of $P_g(\theta)$ at a fixed value of θ. The denominator is the variance of the item response variable, u_g, at a fixed value of θ. The quantity $I(\theta,u_g)$ is an index of the item's ability to discriminate people whose latent trait location equals θ from people at nearby latent trait locations.

In general, a steeper slope for $P_g(\theta)$ implies greater discriminating power. As was noted earlier, high values of a_g and low values of c_g increase the slope of $P_g(\theta)$ and, hence, the information available from an item. The variance of u_g approaches zero at latent trait levels that are deviant from b_g and reaches its maximum value at the latent trait level where $P_g(\theta)=.5$. Figure 4 shows a graph of the function $I(\theta,u_g)$ in the interval from θ=-3.00 to +3.00 for the item shown in Figure 2, which has a_g=1.0, b_g=0.0, and c_g=.00. This graph was generated by evaluating $I(\theta,u_g)$ at 61 points along the theta continuum and rounding the obtained values to the nearest .02.

Figure 4
Information Curve for a Single Item (a=1.0, b=0.0, c=.00)

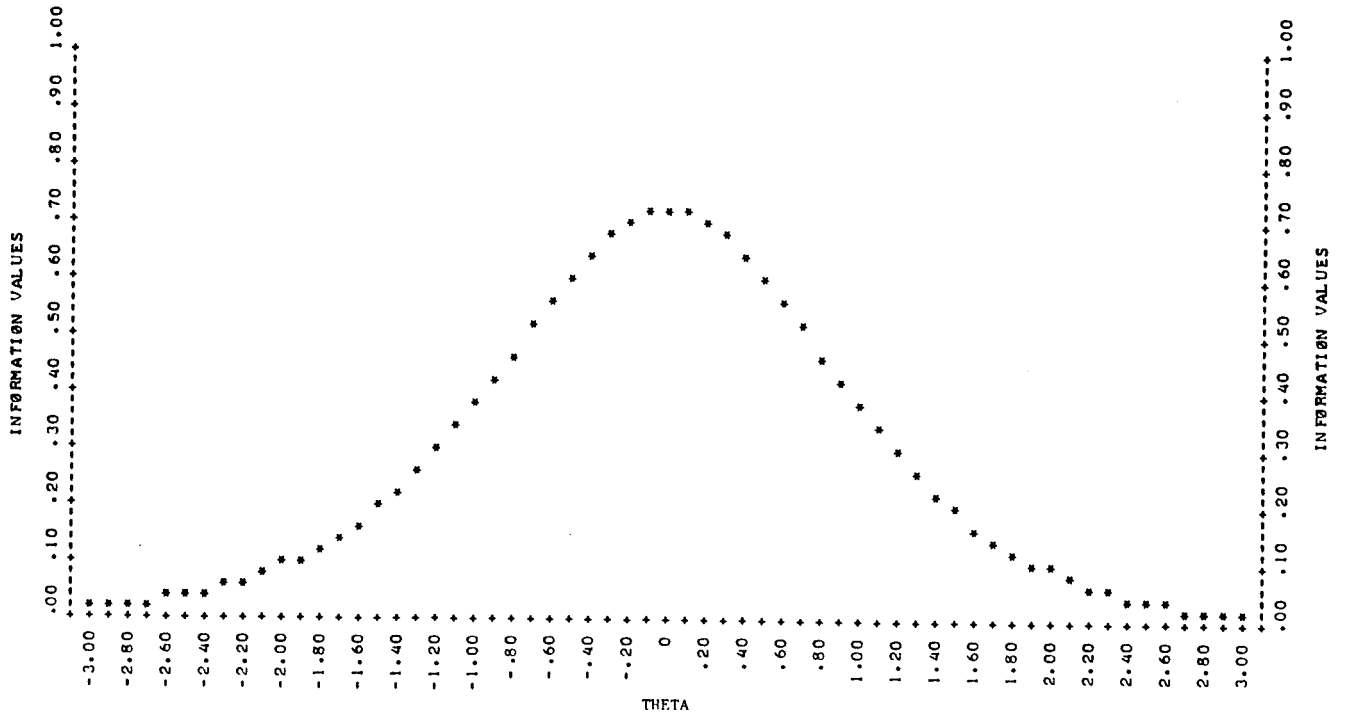


Figure 4 shows that an item provides maximum information in the region of the theta continuum where the item is located (i.e., near b_g) and relatively

little information at levels far below or far above b_g. This result is consistent with intuitive impressions of item discriminating power. If, for example, an ability test item that was suitable for third graders (i.e., $P_g(\theta)$ near .5 among third graders) were administered to college students (in which group $P_g(\theta) \doteq 1.0$), all the college students would probably answer it correctly and no basis for discriminating among college students would exist. Note that in Figure 4 the information curve is symmetric about b_g and attains a maximum value of approximately .72.

Figure 5 shows an information curve for an item having a_g=.85, b_g=0.0, and c_g=.00. This curve, while still symmetric about b_g, attains a lower maximum (approximately .52) and falls off more gradually on either side of b_g than the curve in Figure 4. In fact, the item represented in Figure 5 provides slightly more information than the item represented in Figure 4 in the interval below $\theta \doteq -1.40$ and in the interval above $\theta \doteq 1.40$. However, the gain in these regions is slight compared to the information loss in the interval $-1.40 < \theta < 1.40$.

Figure 6 shows an information curve for an item having a_g=1.0, b_g=0.0, and c_g=.20. This curve is not symmetric about b_g. It attains its maximum value of about .50 near θ=.16. The curve falls off more rapidly on the left of θ=.16 than on the right. This reflects the fact that "chance" keyed responses are more prevalent among people located below b_g than among people located above b_g. Such "lucky" responses contribute error to the estimation of latent trait status and reduce the amount of information available. Note that the information curve in Figure 6 is lower than the curve in Figure 5. Introducing the possibility of "lucky" keyed responses reduces the information available from an item just as if it were an item with lower a_g, but with c_g=.00.

Sequential Estimation in an Adaptive Test

In order to demonstrate the sequential estimation of latent trait status in an adaptive test, a computer program was used to simulate the test responses of a person whose latent trait location is θ=+1.0. Twenty items having a_g=1.0 and c_g=.20 were administered. The items' b_g values changed as a function of responses generated during the simulated test. Table 1 summarizes the results of this 20-item test.

The first column in Table 1 contains item numbers in the 20-item series (g=1,2, . . .,20). The second column contains the b_g values of the items administered. The difficulty of the first item was b_1=0 because this value approximates the mean latent trait score in any population of persons that is sampled to parameterize a set of test items. (An exception to this may be found in Wright and Panchapakesan's (1969) implementation of the Rasch model. They scale the latent trait metric such that the mean of the b_g estimates is

Figure 5
Information Curve for a Single Item (a=.85, b=0.0, c=.00)

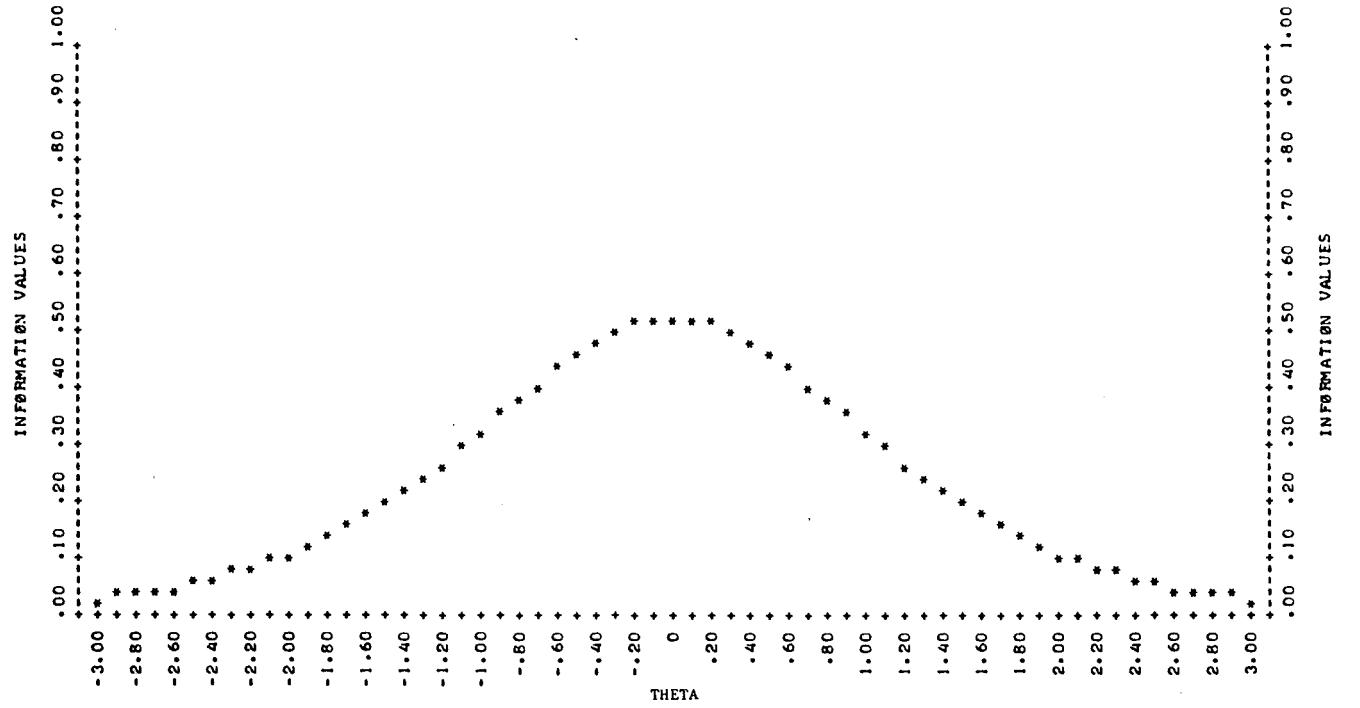
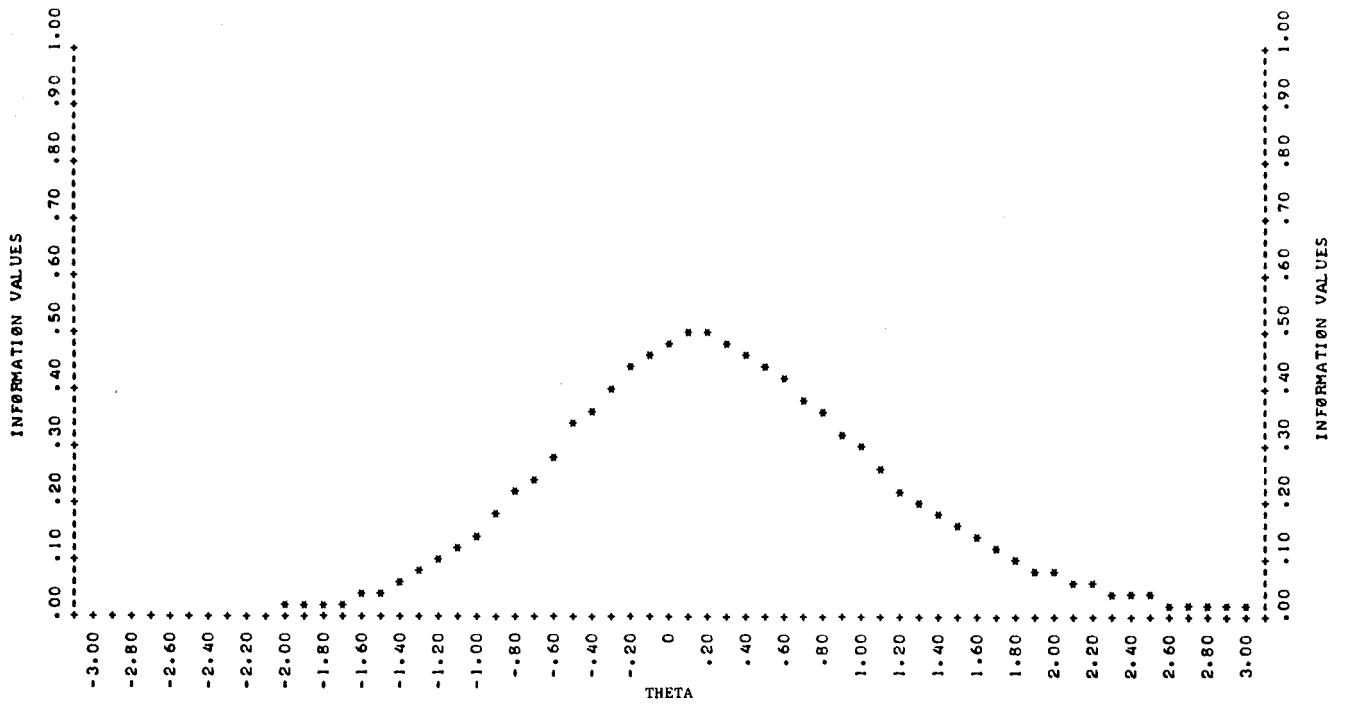


Figure 6
Information Curve for a Single Item (a=1.0, b=0.0, c=.20)



zero and the mean θ estimate among persons is, in general, other than zero.) Following the first item, b_g values either increase or decrease (in accordance with a procedure to be outlined below) depending on whether a keyed or non-keyed response was generated. The item response variable u_g is shown in the third column of Table 1.

Table 1
Sequential Estimation of Latent Trait Status
in a 20-Item Adaptive Test

Item No.	Diff.	Resp.	MAXL Est.	WBL Est.	SBAYES Est.	OBAYES Est.
1	0	1	5.49	1.61	.38	.38
2	1.00	0	.36	-.85	.05	.04
3	0	1	.67	.18	.32	.31
4	.18	1	.89	.82	.53	.54
5	.82	1	1.16	1.25	.75	.78
6	1.25	0	.87	.72	.57	.56
7	.72	1	1.03	1.00	.74	.75
8	1.00	1	1.20	1.21	.89	.93
9	1.21	0	.99	.93	.74	.74
10	.93	1	1.12	1.10	.87	.89
11	1.10	0	.95	.89	.73	.72
12	.89	1	1.05	1.02	.84	.84
13	1.02	0	.91	.85	.72	.70
14	.85	1	.99	.96	.82	.80
15	.96	1	1.07	1.05	.90	.90
16	1.05	0	.96	.92	.80	.78
17	.92	1	1.03	1.00	.88	.87
18	1.00	0	.93	.89	.79	.76
19	.89	1	.99	.96	.86	.84
20	.96	1	1.05	1.03	.92	.92

Likelihood-based estimation. The last four columns of Table 1 contain four different estimates of latent trait status that were calculated after each item was administered. The fourth column of Table 1 contains maximum-likelihood estimates of θ. A maximum-likelihood estimate of θ corresponds to the latent trait location at which the observed pattern of item responses has the maximum probability of occurrence. The probability of a set of item responses, given some fixed value of θ and the item parameters, is obtained using the *likelihood function* given in Equation 3.

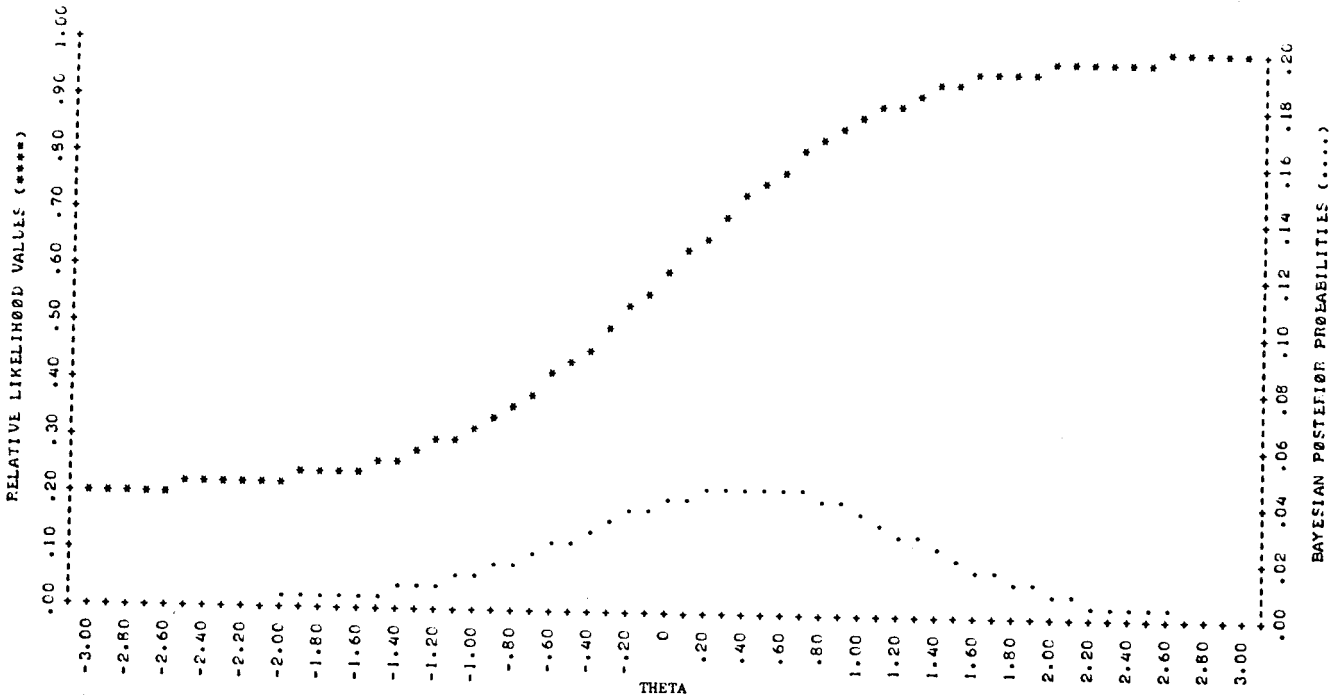
$$L_v(\theta) = \prod_g [P_g(\theta)^{u_g} \; Q_g(\theta)^{1-u_g}] \qquad [3]$$

This equation assumes that the responses of a given person to different test items are independent of one another. The operator Π indicates that a serial product is to be taken over the test items administered up to that point (g=1,2,...k).

After each item was administered, Equation 3 was evaluated at 101 equally spaced θ values in the interval from θ=-5.00 to θ=+5.00 and the largest of the 101 likelihood values was identified. Then, a quadratic function was fitted to this largest likelihood value and the two likelihoods adjacent to it. The value of θ corresponding to the maximum of the quadratic function was used as the "MAXL" estimate. Under most conditions, the estimate of θ obtained in this manner is a good approximation to the estimate that would be obtained if more sophisticated methods of numerical analysis were used to search for a root of the log-likelihood function's first derivative.

The interval between θ=-5.00 and θ=+5.00 will contain at least 96% of the θ estimates in any group that is used to parameterize test items. This is because latent trait item parameterization procedures scale the theta metric such that the mean θ estimate equals zero and the standard deviation among the estimates is 1.0 (again, the Rasch model provides an exception to this general result), and by virtue of Tchebycheff's inequality which states that the proportion of cases which fall more than S standard deviations from the mean cannot exceed $(1/S^2)$ in any distribution (Hays, 1973, p. 253). If the distribution of θ estimates is peaked and unimodal, virtually all of the θ estimates will be between -5.00 and +5.00.

Figure 7
Relative Likelihood and Posterior Probability Curves After 1 Item



Figures 7, 8, and 9 show graphs of the data likelihood function in the interval from θ=-3.00 to θ=+3.00 following the administration of 1, 2, and 3 items, respectively. For plotting purposes, the raw likelihood values were expressed relative to the largest likelihood value in the interval θ=-5.00 to θ=+5.00 and

Figure 8
Relative Likelihood and Posterior Probability Curves After 2 Items

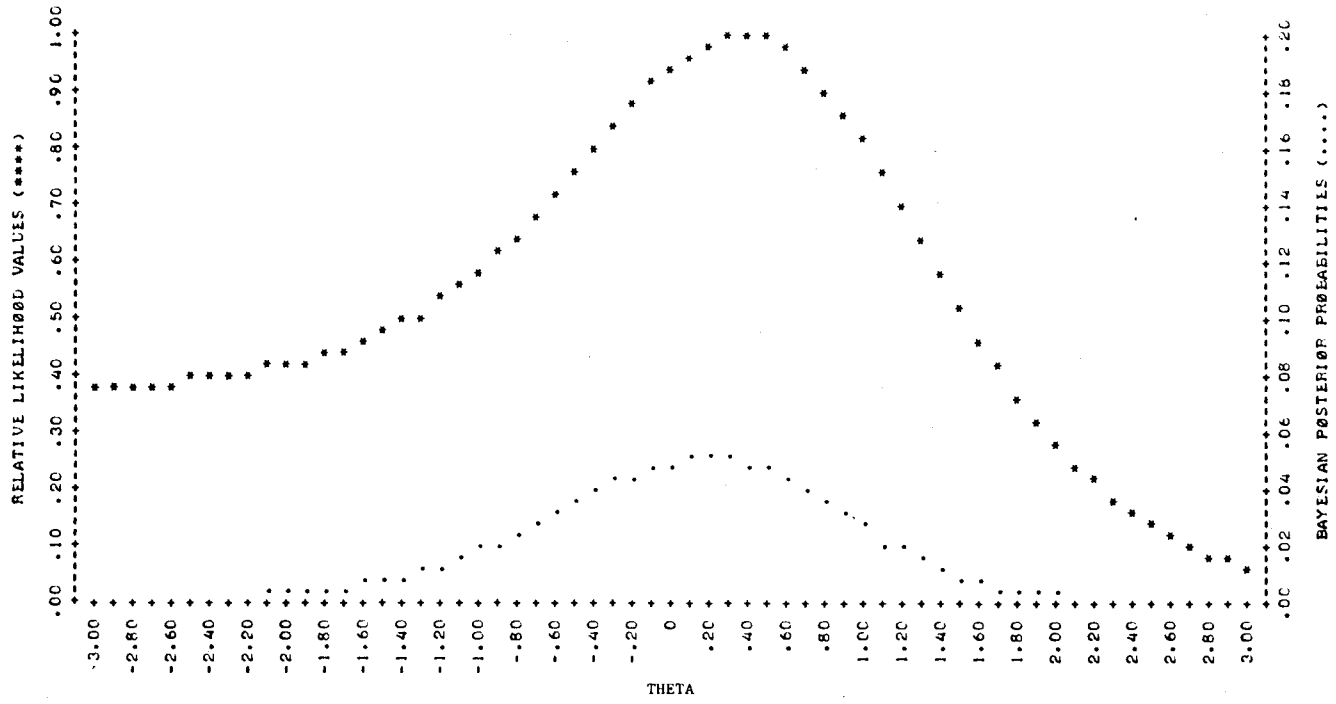
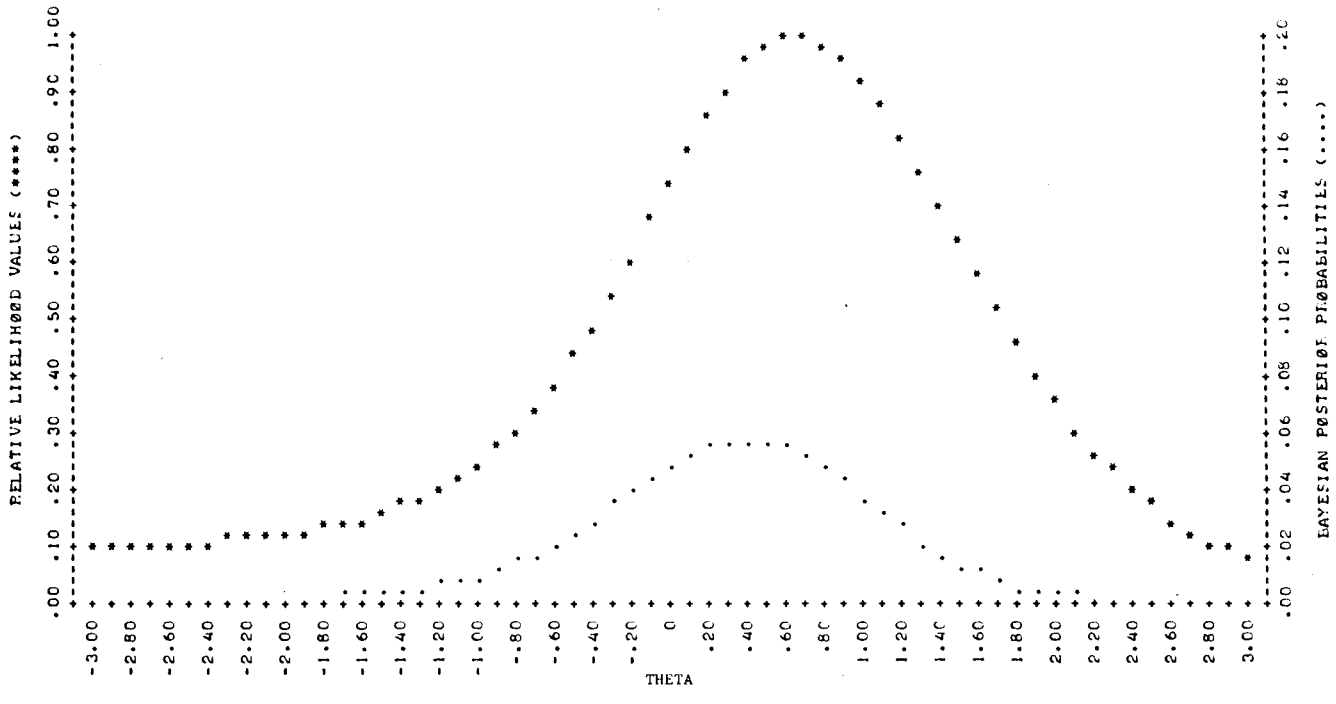


Figure 9
Relative Likelihood and Posterior Probability Curves After 3 Items



then rounded to the nearest .02. As can be seen in Equation 3, after one item is administered the likelihood function corresponds to either $P_1(\theta)$ or $Q_1(\theta)$, depending on whether a keyed or non-keyed response is emitted (compare Figure 7 and Figure 3). The MAXL estimate after a "correct" answer to the first item is +5.49. Actually, since $P_g(\theta)$ is strictly increasing in θ, the estimate should be $\theta=+\infty$, but a finite estimate is certainly more reasonable. After an "incorrect" answer to the second item, with b_2=1.00, the peak of the likelihood curve occurs near θ=+.36 (Figure 8). After the third item, the peak occurs near θ=.67 (Figure 9).

"Weighted-by-likelihoods" (WBL) estimates of latent trait status appear in the fifth column of Table 1. The WBL estimates were obtained by taking a weighted average of 101 equally spaced θ values in the interval from θ=-5.00 to θ=+5.00. The weights used were the data likelihoods at each θ value. That is,

$$\text{WBL Est.} = [\sum_{\theta}(L_v(\theta)\ \theta)]/[\sum_{\theta}(L_v(\theta))] \qquad [4]$$

where θ takes on the values -5.00, -4.90, ..., +5.00. The WBL estimate is influenced by the entire set of 101 likelihood values instead of just the maximum of the likelihood function.

The MAXL and WBL estimates can differ considerably when only a few items have been administered, as can be seen in Table 1. Inspection of the relative likelihood curve in Figure 8 shows why these two estimators differ after two items have been administered. The WBL estimate is lower due to the fact that the left tail of the likelihood curve is high relative to the right tail. Table 1 also shows that the MAXL and WBL estimators become more similar as the number of items administered increases. Since the WBL estimator has not been proposed previously, future research is planned to study its characteristics.

The procedure by which item b_g values were determined during the simulated test now can be outlined. The general rule followed was: Let the next item have a difficulty level equal to the current value of the WBL estimator, except that in no case shall the new b_g value be more than 1.00 units from the immediately preceding b_g value. Thus, as can be seen in Table 1, item difficulties changed by 1.00 until the third item had been administered and the WBL estimate was .18. After this, each item difficulty corresponded to the value of the WBL estimate following the preceding item. In actual practice, an item is seldom found with b_g exactly equal to the current estimate of latent trait status. In such cases, an item that has b_g close to the desired value is selected for administration.

Bayesian estimation. Columns six and seven of Table 1 contain Bayesian estimates of latent trait status. Given a specified form for the continuous distribution of latent trait scores in a population (i.e., the *prior probability density function* of theta), the item parameters for the items administered, and a vector of item responses (u_g values), it is possible, in principle, to derive the *posterior probability density function* of theta using the inverse probability rule

of Bayes (Hays, 1973, p. 819). In practice, it becomes difficult to obtain analytic expressions for the posterior theta distribution unless the prior distribution and the data likelihood function take on certain restricted forms. To avoid such difficulties, the following approximate procedure can be used.

First, the continuous prior density function of theta is approximated with a discrete probability distribution in which the probabilities are concentrated at 101 equally spaced points along the theta continuum. Thus, for example, the area under the prior density curve between θ=-.05 and θ=+.05 is assigned to the point θ=.00. This is done for θ=-5.00, -4.90, ..., +5.00. Areas beyond θ=-5.05 and θ=+5.05 are assigned to the points θ=-5.00 and θ=+5.00, respectively. (These extreme tail areas should be trivially small. If they are not, the region of the theta continuum in which the procedure is applied can be shifted or extended.) Next, data likelihoods are generated at the same 101 values of θ using Equation 3. The prior probabilities, $f(\theta)$, and the data likelihoods, $L_v(\theta)$, are then entered into into Equation 5 in order to determine the posterior probability of each given θ value.

$$P(\theta|v) \doteq [L_v(\theta)\ f(\theta)]/\sum_{\theta}[L_v(\theta)\ f(\theta)] \qquad [5]$$

The resulting 101 posterior probabilities provide a discrete approximation to the continuous posterior distribution of theta. Finally, the mean of the discrete posterior distribution is obtained with Equation 6 and this value is referred to as the "SBAYES" (simplified Bayesian) estimate at that stage of the testing procedure.

$$\text{SBAYES Est.} = \sum_{\theta}[P(\theta|v)\ \theta] \qquad [6]$$

SBAYES estimates of θ appear in column six of Table 1. Figures 7, 8, and 9 show three of the posterior probability distributions that were generated with the SBAYES procedure when the prior distribution of latent trait scores was specified to be a normal density function with zero mean and unit variance. The first three SBAYES estimates in Table 1 are the means of these discrete distributions.

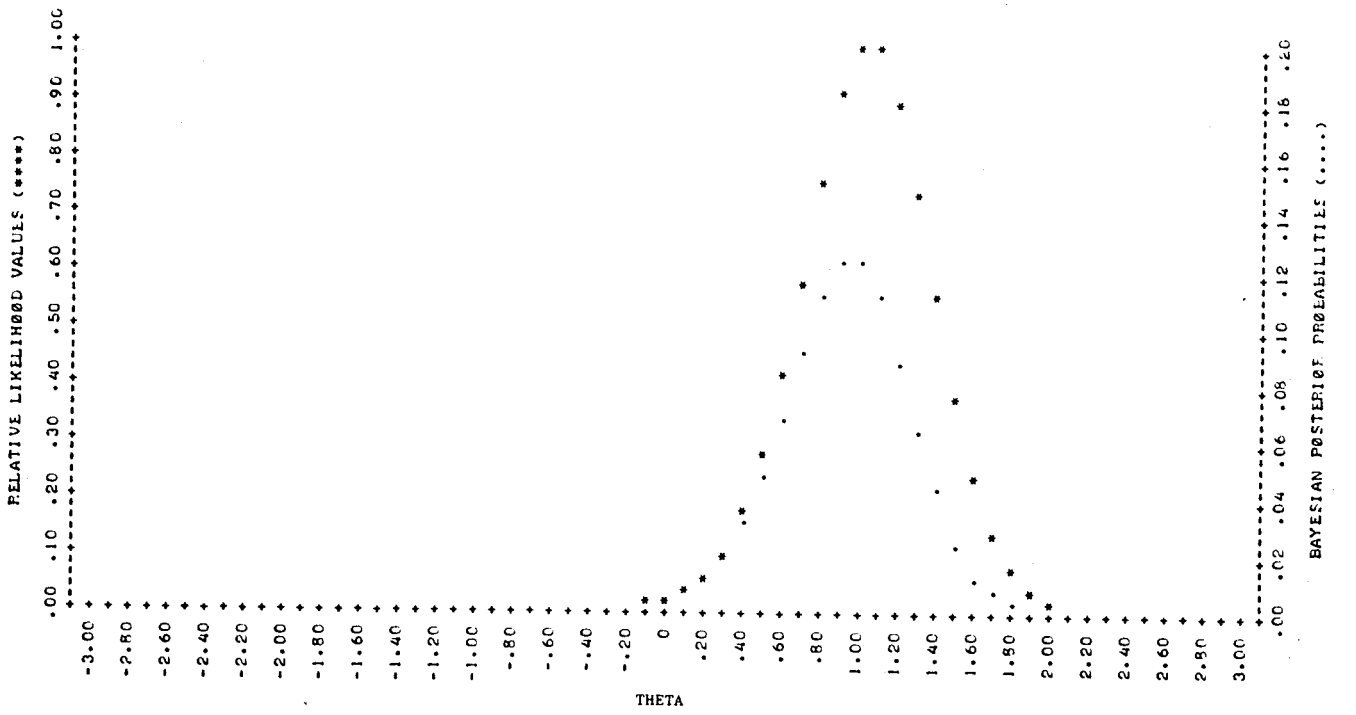
The "OBAYES" (Owen Bayesian) latent trait estimates that appear in column seven of Table 1 were obtained using a procedure described by Owen (1975). While Owen has described both a method for estimating latent trait status and a method for selecting test items, only his estimation procedure was used here. Owen introduced his procedure in the context of a three-parameter normal ogive latent trait model. The close similarity of this model to the logistic model given in Equation 1 allows its application here.

The OBAYES procedure has two drawbacks. First, it is limited to prior distributions that follow a normal density function. The SBAYES procedure described above can accept any type of prior distribution. Second, the OBAYES procedure is order dependent. That is, if a set of items is administered and the item responses are recorded, then the value of the OBAYES estimator will depend partly on the order in which the items are processed by the scoring procedure. The OBAYES procedure implicitly generates an updated prior distribution after each item is scored and then combines this new prior distribution with the likelihood function for the

response to the next item. This in itself would not make the OBAYES procedure order dependent but, in order to simplify the mathematics, Owen proceeded as if each updated prior distribution could be described by a normal density function. This approximation introduces a small amount of inaccuracy into the estimation process and makes the procedure order dependent. The SBAYES procedure does not utilize this type of approximation and is not order dependent.

After administering a single item, SBAYES and OBAYES estimates generally agree to three decimal places when the initial prior distribution of θ is a normal density function. Since the OBAYES estimate is optimal in this particular situation, this level of agreement can be viewed as an indication that very little inaccuracy is introduced by the discrete approximations in the SBAYES procedure. When more than one item has been administered, or when the prior distribution specified for the SBAYES procedure is non-normal, the two estimation methods will not necessarily agree.

Figure 10
Relative Likelihood and Posterior Probability Curves After 20 Items



Comparisons between likelihood-based and Bayesian estimates. Figure 10 shows the relative likelihood and posterior probability curves that resulted after 20 items had been administered. The likelihood curve peaks near θ=1.05 and the posterior probability distribution has a mean of .92 (see Table 1). Both the likelihood curve and the posterior probability curve have shifted to the region of the theta continuum near θ=1.00, and both curves have become more peaked. In fact, as test length (k) approaches infinity, both of these curves approach a vertical

line (i.e., a single-valued distribution) located at the value of θ that is generating the item responses.

Note in Table 1 that the Bayesian estimates of θ tend to stay closer to θ=.00 than the likelihood-based estimates throughout the testing process. This is because Bayesian estimators are "drawn toward" the high density region of the prior distribution. This is appropriate when one's objective is to minimize squared errors of estimation in the population specified by the prior distribution. Unfortunately, for tests of moderate length, a certain amount of bias at the tails of the theta distribution must be accepted in order to achieve this minimization (McBride & Weiss, 1976).

For moderate k, the maximum-likelihood estimator can also be biased. However, for a given value of k and values of θ deviant from the high density region of a peaked prior distribution, the maximum-likelihood estimator will tend to be less biased than the Bayesian estimator. The Bayesian estimator's bias can be reduced by increasing k as the estimate of θ deviates from the high density region of the prior distribution. This can be done readily in an adaptive testing situation.

An interesting relationship exists between the likelihood-based estimators and Bayesian estimators. If one applied the SBAYES estimation procedure and specified that the prior distribution of theta was rectangular in the interval θ=-5.05 to θ=+5.05, then the SBAYES estimate of θ, as determined by Equation 6, would be identical to the WBL estimator. Moreover, the MAXL estimate would closely approximate the mode of the Bayesian posterior probability distribution. Thus, all four types of latent trait estimators that have been presented here can be viewed as Bayesian estimators. The MAXL estimator is a Bayesian modal estimate of θ when the implicit prior is restricted to a rectangular form, the WBL estimator is a least-squares estimate of θ when the implicit prior is restricted to a rectangular form, and the OBAYES estimator is a least-squares estimate of θ when the explicit prior is restricted to a normal form. The SBAYES procedure is the only one of the four methods that does not restrict the form of the prior distribution. By virtue of this flexibility, the SBAYES estimation procedure appears to be the most widely applicable of the four methods.

Total Test Information

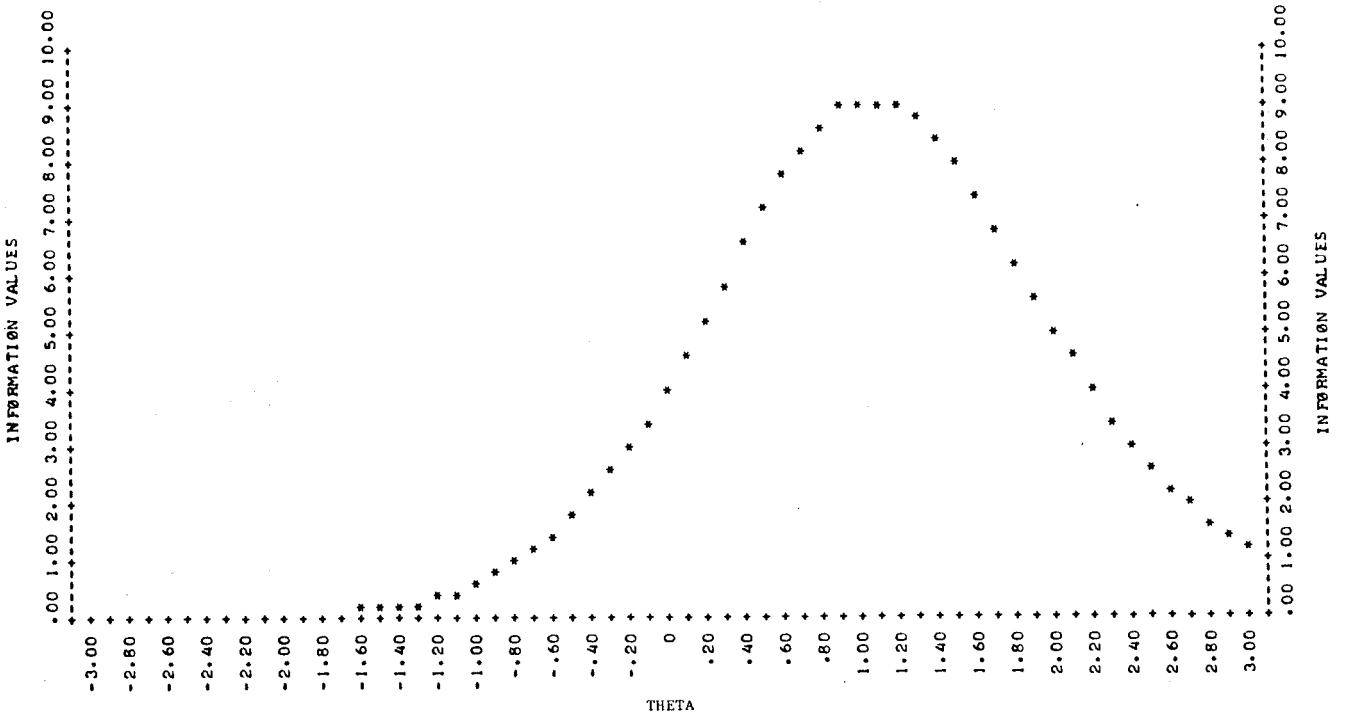
Birnbaum (1968, p. 454) has defined the *information function of a test* as

$$I(\theta) = \sum_g [I(\theta, u_g)]. \qquad [7]$$

This function is the sum of the constituent item information functions and defines the maximum amount of information that can be extracted from a set of items. The amount of information actually extracted depends on how the items are scored.

Information in the adaptive test. Figure 11 shows a graph of the test information function for the 20 items administered in the simulated adaptive test. It was obtained by evaluating Equation 7 at 61 equally spaced points along the theta continuum in the interval from θ=-3.00 to θ=+3.00. This curve shows the maximum amount of information available from these items. The curve peaks near θ=1.00, thus indicating that this set of items provides maximum discrimination among individuals whose latent trait locations fall near θ=1.00. The maximum value of the curve is about 9.00.

Figure 11
Information Curve for 20-Item Adaptive Test



Information in two conventional tests. Figure 12 shows a graph of the test information function for a set of 20 items having a_g=1.0, c_g=.20, and b_g values equally spaced in the interval from -3.00 to +3.00 (i.e., $b_g \doteq$-3.00, -2.68, -2.37, . . ., +3.00). This would commonly be referred to as a "rectangular" distribution of item difficulties. This test provides a fairly uniform level of information across a broad range of the theta continuum. Unfortunately, the level of information is relatively low. The curve attains its maximum value of about 3.20 in the interval $-1.00 \lessapprox \theta \lessapprox 1.90$.

Figure 13 shows a graph of the test information function for a set of 20 items having a_g=1.0, c_g=.20, and b_g=0.0 for all items. This is a "perfectly peaked" test. The shape of this information curve is rather similar to the curve in Figure 11, but it is shifted to the left. The curve in Figure 13 attains its maximum value of 9.80 near θ=.16. At θ=1.00, the value of this information curve is about 5.80.

Figures 12 and 13 represent two rather idealized non-adaptive tests. Both of these tests deliver less information at θ=1.00 than the items selected by the adaptive testing procedure. What is the implication of this result? If, for some practical purpose, it were necessary to order a testee with $\theta \doteq$1.00 relative to other individuals falling at nearby θ values, fewer errors would be made if θ estimates derived from the adaptive test's items were used than if estimates derived from either conventional test were used.

Figure 12
Information Curve for 20-Item Rectangular Test ($-3.0 \leq b \leq +3.0$)

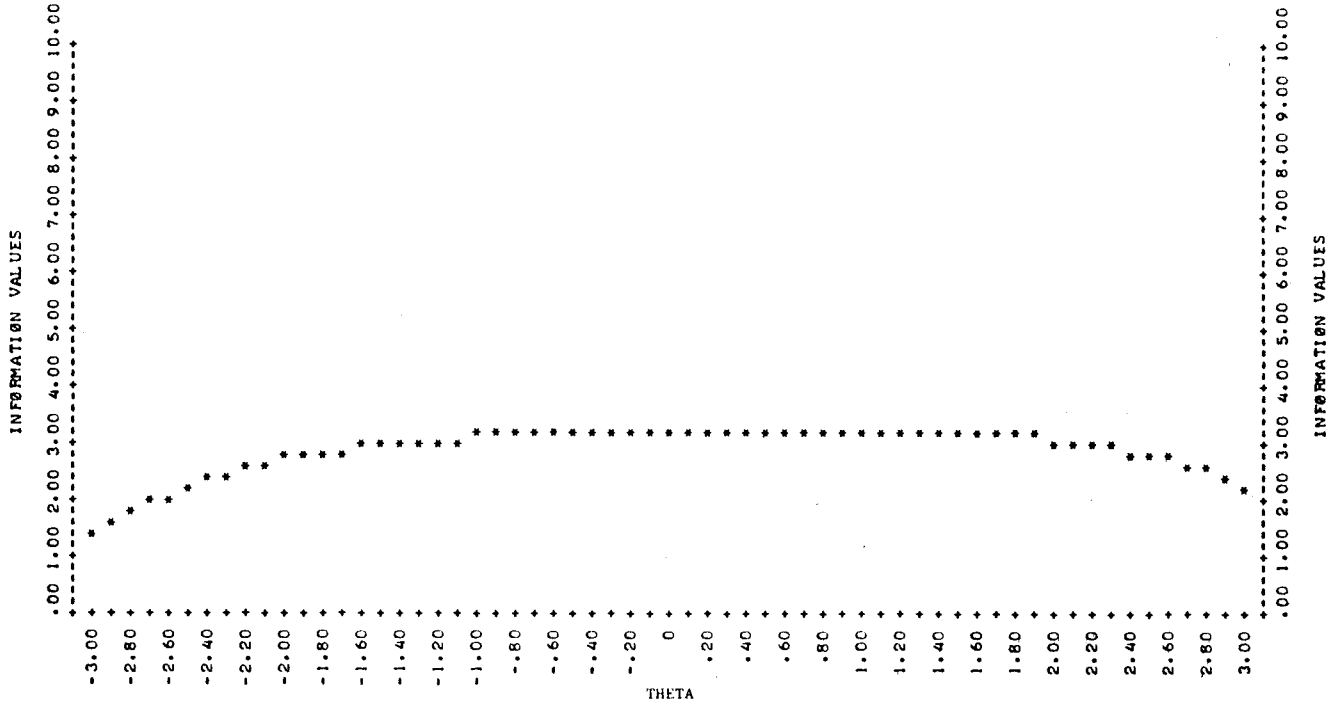
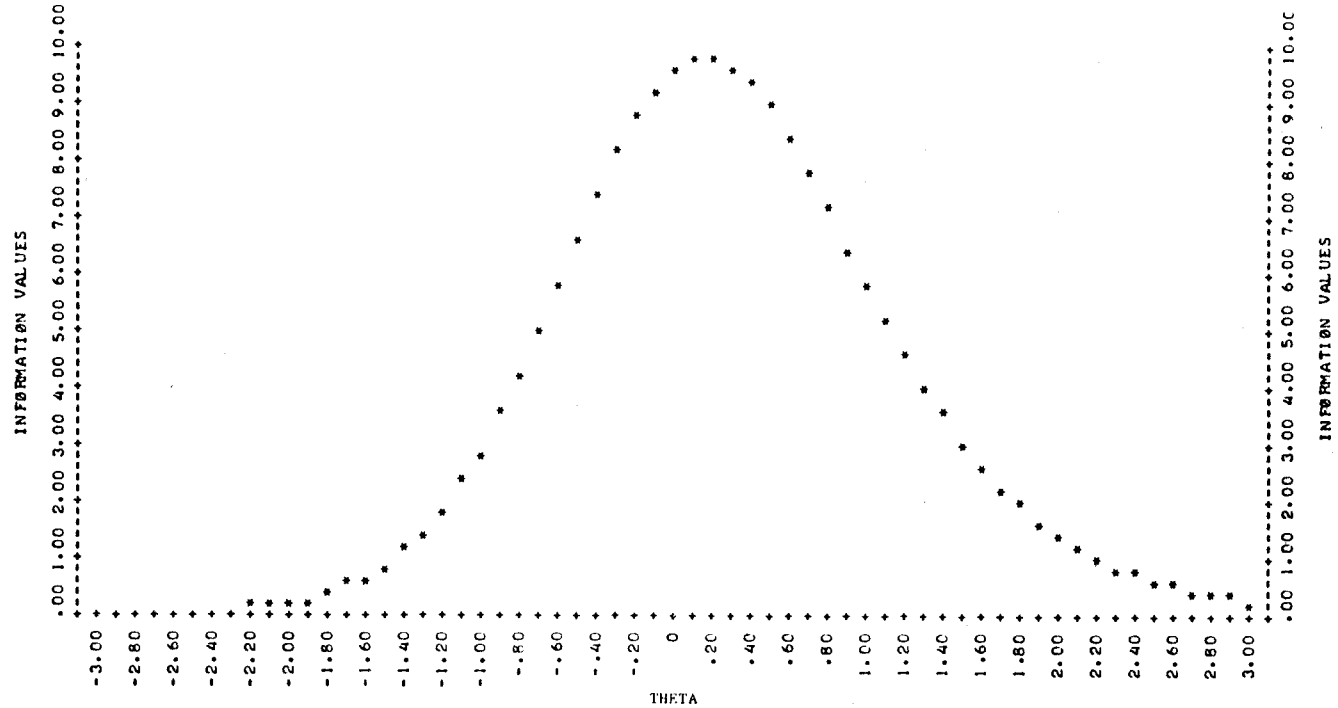


Figure 13
Information Curve for 20-Item Peaked Test (b=0.0 for all items)



Summary

Several procedures for estimating latent trait status have been presented. It has also been suggested that adaptive testing procedures often can provide more accurate estimates of latent trait status than conventional tests. Though there is no necessary connection between latent trait theory and adaptive testing, there is a strong natural impetus toward their joint application. Latent trait theory provides adaptive testing with a coherent theoretical foundation. It is a guide to procedures for designing and scoring adaptive tests. On the other hand, adaptive testing offers the opportunity to take maximum advantage of the potentialities of latent trait theory. At this point in time, both a new type of test theory and a new type of testing technology are available. Their joint effect might possibly exceed the sum of the two parts.

ADAPTIVE TESTING AND THE PROBLEM OF CLASSIFICATION

C. DAVID VALE
University of Minnesota

Two basic goals in the use of ability tests are measurement and classification. When a test is used for measurement, the objective is to accurately determine where a testee's ability lies on the latent ability continuum. When a test is used for classification, the objective is to determine on which side of a cutting score (or between which cutting scores) a testee's ability lies. Such classification decisions should be made so as to minimize the errors of misclassification. Once a classification is made, there is no necessity for a more precise determination of an individual's ability level.

This paper is concerned with the classification of abilities into discrete categories. The general goals of classification will be explicated and alternative means that may practically be used to achieve these goals will be presented and compared using monte carlo computer simulations.

The Classification Problem

Classification Errors and Utility Functions

The goal of this classification is to determine, with a minimal probability of being in error, on which side of a cutting score or between which of several cutting scores, a testee's ability falls. There are two kinds of error probabilities that can be examined in making these classifications. One is the conditional probability of being in error (i.e., for a single testee or at a specific ability level); the other is the expected or unconditional probability of being in error across a group of testees. The conditional probability is a function of the test, the testee's ability level and the placement of the cutting score (for the moment, limiting the discussion to one cutting score). For a given test of fixed length, the probability of making an error of classification for a testee is usually high if the testee's ability level (θ) is near a cutting score (θ_c), and lower if the ability level is distant from the cutting score. This conditional probability of misclassification [$P(M|\theta)$] is described by a function like that shown in Figure 14.

The unconditional probability of misclassification for a group of testees [$P(M)$], is a function of the conditional reliability function and the distribution of abilities within the group under consideration. For a large group with abilities distributed N(0,1), this probability is given by Equation 8.

$$P(M) = \int_{-\infty}^{\infty} P(M|\Theta)\ \phi\ (\Theta)\ d\Theta \qquad [8]$$

where $\phi(\Theta) = [2\pi \cdot \exp(\Theta^2)]^{-\frac{1}{2}}$

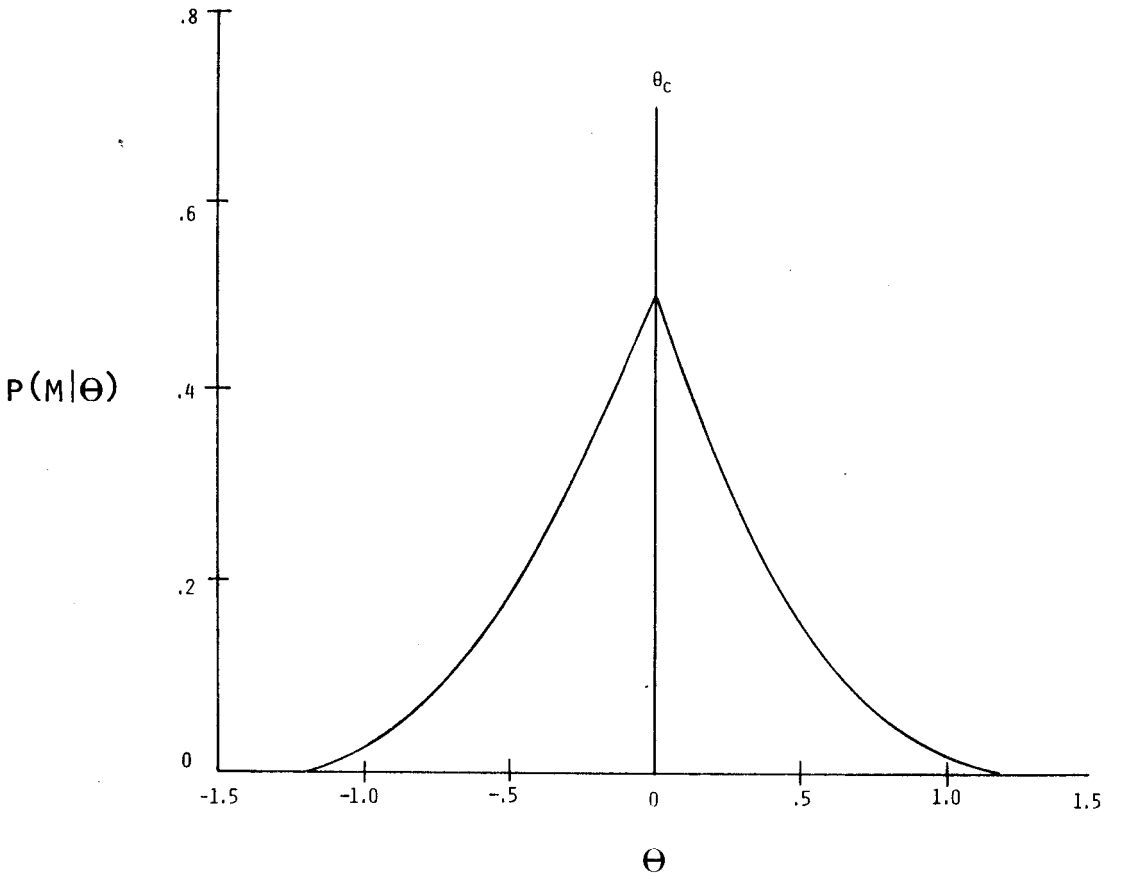
In practical situations, it may be desirable to minimize the quantity in Equation 8. This unconditional probability is a scalar quantity and as such can be

---

This research is supported by contract N00014-76-C-0243, NR150-382, with the Personnel and Training Research Programs, Office of Naval Research.

minimized. A function such as the conditional probability function can only be minimized at a single point and this is typically of little practical value because theoretically, assuming a continuous distribution of ability, the probability of anyone having an ability at that point is zero.

Figure 14
A Conditional Probability of Misclassification Curve



A more viable approach to making classification decisions is one that will, over a group of individuals, maximize some form of utility such as the quality of performance extracted from the work force. The unconditional probability of misclassification reflects errors of classification into categories along a latent continuum and it may be errors of classification along an observable success-failure continuum that are of interest. This possibility is important because two individuals, one with an ability level slightly above a cutting score on the latent continuum and the other with ability slightly below the cutting point, probably have a trivial difference between their probabilities of success on a job. If both are classified above the cutting score, however, one will be considered a "hit" and the other a "miss" when classification occurs on the latent continuum. In order to assess the practical value (i.e., cost effectiveness) to an organization of an adaptive testing strategy, utility functions of θ for each decision must be specified. As an example of such utility functions, consider the following:

For three classifications--low, middle, and high--three utility functions might be:

$$U_{low} = .5 \qquad [9]$$

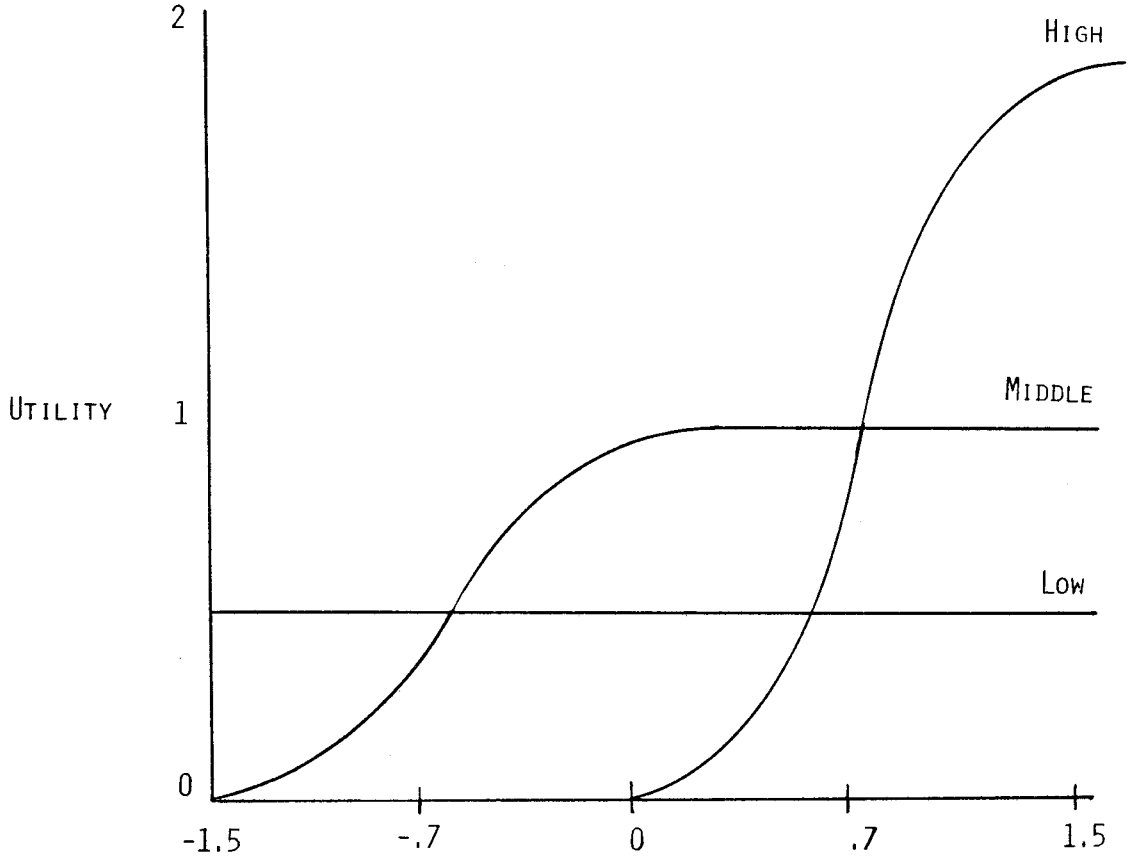
$$U_{medium} = \Phi(3.0(\theta+0.7)) \qquad [10]$$

$$U_{high} = 2.0(\Phi(3.0(\theta-0.7))) \qquad [11]$$

where $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$

A practical situation in which these utility functions might arise is as follows: There are three jobs requiring an ability, θ. One is so easy that almost anyone can do it but when performed satisfactorily, it is only .5 utility units of

Figure 15
Conditional Utilities for each of Three Decisions



value to the organization. A second job is fairly easy and 50% of people with θ above -.7 can perform it satisfactorily. Differences in ability near -.7 make greater changes in the probability of success than do differences around, say,

θ=0.0. Ninety-eight percent of people with θ above 0.0 will be successful on the job and additional increments in θ are of little importance in predicting job success. Success in this job is worth one unit of value. A third job requires higher θ to be successful, but is worth two units of value when performed satisfactorily. The utility functions defined by Equations 8, 9, and 10 result in the three utility curves presented in Figure 15. As can be seen, there is a clear reason for assigning high θ people to the third job and lower θ people to the second and first jobs.

Test Design for Classification Problems

Although it may be possible to determine that quantity (e.g., probability of misclassification or expected utility) which is to be minimized or maximized, it is difficult to design a test explicitly for that purpose. The goal of optimal test design can be approached practically via one of several approximation strategies. Two general types of testing strategies that have been researched in the ability measurement domain are the conventional testing strategy and the adaptive testing strategy. In the former, test items are selected to best measure the abilities of members of a group, and the same test is given to everyone. In the latter, a test is tailored, during the testing process, to each individual's level of ability, and a different test may be given to each person. This permits higher measurement precision over most of the ability continuum than that attained with a conventional test.

In the remainder of this paper, two forms of a conventional test and one form of an adaptive test will be compared. The conventional tests will be a unimodally peaked test with all item difficulties of one value and a bimodally peaked test (i.e., the simplest form of a multimodally peaked test) with difficulties of two values. As will be discussed later, these are, respectively, attempts to put items at a level where they best measure most people or at a level where people need to be measured best. The adaptive test to be compared will be Owen's (1975) Bayesian strategy. This strategy starts with some estimate of an individual's ability, chooses an appropriate item, administers the item, and forms a new estimate of the individual's ability. Using this estimate, it chooses the next item and continues this procedure until the end of the test.

These strategies will be compared along the criteria previously discussed. Since utility functions are peculiar to an organization, the majority of the comparisons will be in terms of misclassification probabilities. The utility functions presented above will, however, be discussed as examples in some later comparisons.

Simulation Procedures

The comparisons presented in this paper assume that classification decisions are made in the following way:

1) A testing strategy selects a subset of items from a large pool of items;

2) These items are then administered to a testee, and from his responses to those items an estimate of ability level is obtained;

3) The testee is then classified into that category which:

   a) in the case where probability of misclassification is of interest, is the one in which his estimated ability falls, or

   b) in the case where utility maximization is of interest, is the one which for his estimated ability predicts the highest utility.

To simplify the analyses and interpretations, availability of an infinitely large item pool was assumed. This pool contained items of all difficulties with their discriminating powers fixed at a constant level. It was further assumed that these items could not be correctly answered by guessing. These assumptions reduced the problem of item selection to determining the difficulty of the next item to be administered in the adaptive test. Finally, to make a determination of the unconditional probability of misclassification possible, ability was assumed distributed N(0,1).

Owen's (1975) Bayesian testing procedure requires a prior estimate of a testee's ability to administer and score a test. For all data presented in this paper, a fixed prior ability distribution which was N(0,1) was used for all testees. Owen's scoring procedure was used to score the conventional tests and again a N(0,1) prior was used.

Generation of Misclassification Probabilities and Expected Utilities

Conditional probability of misclassification was calculated for each of 30 values of θ equally spaced between θ=-1.45 and θ=1.45. The simulation procedure followed that described by McBride and Weiss (1976) or Vale and Weiss (1975). Ten-item "tests" were administered to 200 "testees" at each of 30 points. The means and standard deviations of the ability estimates were calculated at each point, a normal distribution with these parameters was determined, and the proportion of that distribution falling outside the correct cutting score interval was taken as the probability of misclassification at that level of ability. These probabilities were then visually fitted into the smooth curves shown in the figures.

To determine the unconditional probability of misclassification, ten-item "tests" were administered to 2,000 "testees" with ability levels randomly sampled from a N(0,1) population of ability levels (the same sample of 2000 ability levels was used for all comparisons). The predicted category for individuals was the score interval in which their ability estimate fell. The true category was the interval in which their true ability fell. An individual was considered misclassified if the predicted category was not the same as the true category. The number of misclassified individuals divided by 2000 was taken as the unconditional probability of misclassification.

Expected utility was determined by generating 2000 ability estimates following the same procedures used in the calculation of expected probability of misclassification. The optimal decision to make for an individual was taken as the decision corresponding to the utility function with the highest value at the estimated level of ability. The actual utility was the value of the utility function corresponding to the decision made, evaluated at the "testee's" true level of ability. The

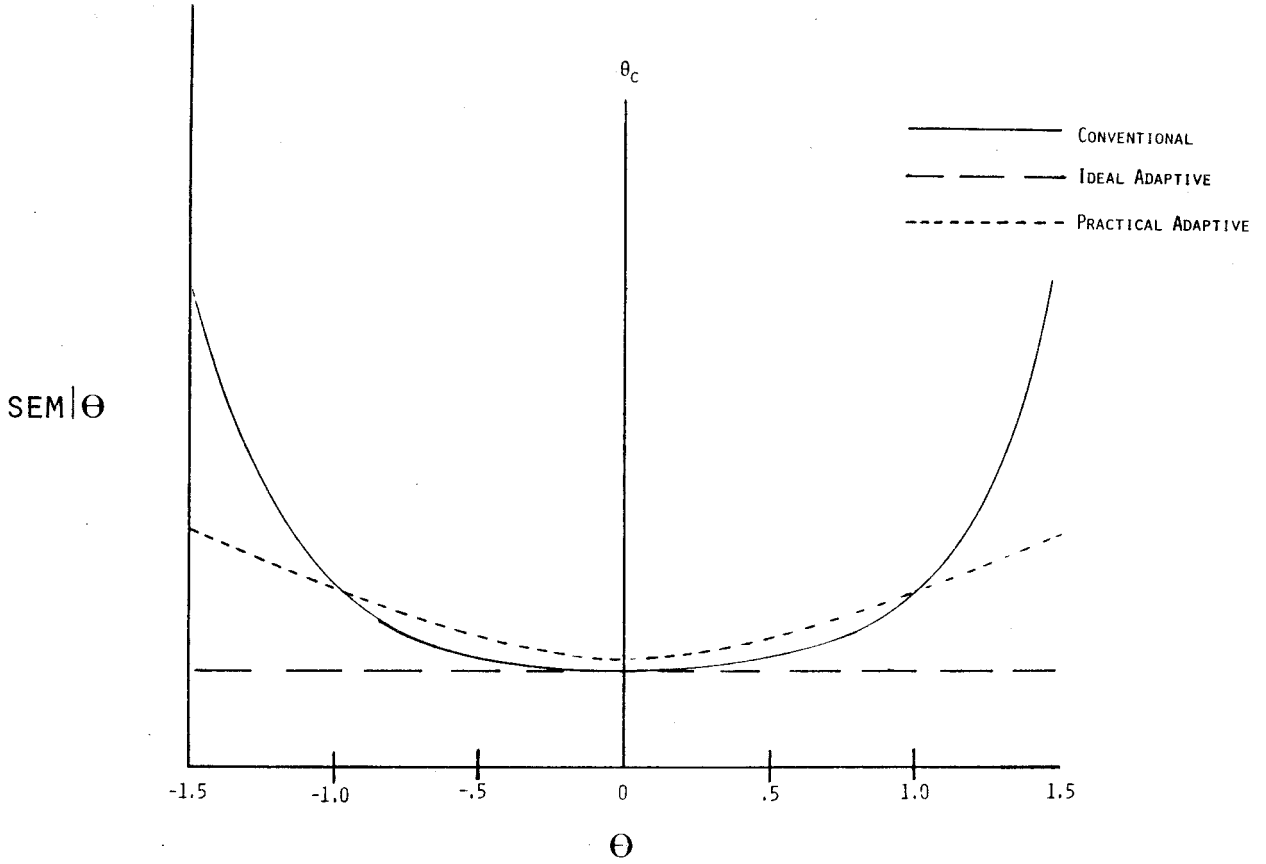
expected utility was simply the mean of these 2000 actual utility values. These values are reported only in comparisons of tests in decisions involving more than one cutting score.

Results

A Single Cutting Score

The simplest categorization situation to investigate is where there is one cutting score placed in the middle of the ability distribution at θ_c=0.0. The best conventional test for making this decision is one with all of its items peaked at b=0.0. Figure 16 shows curves representing standard error of measurement functions

Figure 16
Standard Error of Measurement Curves for Three Tests



(the reciprocal square root of the information functions) for three ten-item tests with a=2.0; a peaked conventional test with all items having b=0.0, an ideal adaptive test with all items having b=θ, and a practical adaptive test with items having difficulties at the estimated ability level at each stage. The conventional test provides a low error level at θ=0.0, but higher error levels distant from that point. The ideal adaptive test provides the same low level of error at all ability levels but is unrealistic because in order to implement it, it is necessary to know

a testee's ability level before the test is administered. A practical adaptive test provides a standard error function lower than that of the conventional test at ability levels distant from θ=0.0, but relatively higher near θ=0.0.

Assuming errors of measurement at a level of θ are distributed $N(\theta, SEM^2)$, the probability of misclassifying an individual is given by Equation 12.

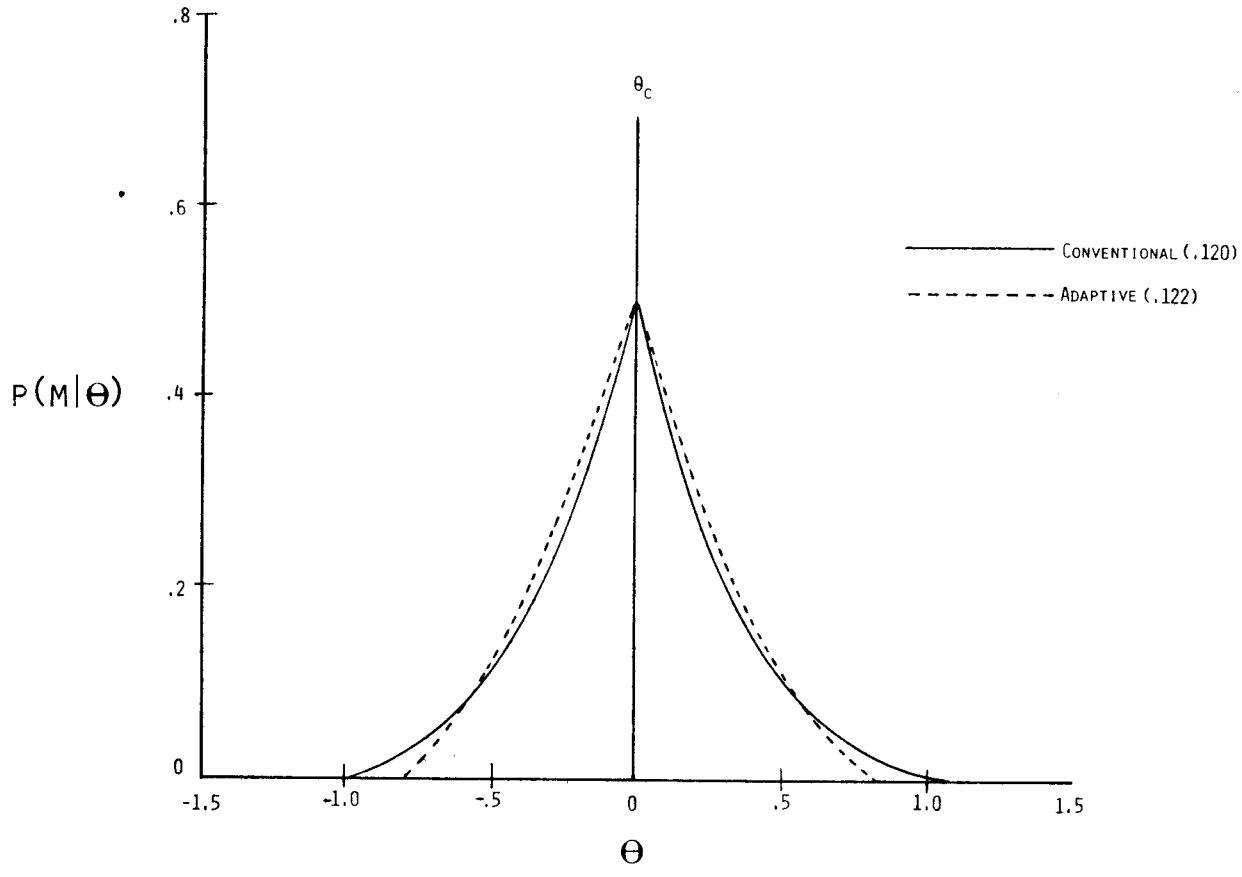
$$P(M|\theta) = 1 - \Phi \frac{|\theta_c - \theta|}{SEM}$$

$$= 1 - \Phi[\sqrt{I(\theta)\ (\theta_c - \theta)^2}] \qquad [12]$$

where θ_c is the cutting score, and $I(\theta)$ is the test information function evaluated at θ.

It can be shown from Equation 12 that when θ_c is fixed, $P(M|\theta)$ is a monotonic increasing function of the standard error of measurement. Thus, the ordering of the

Figure 17
Conditional Probability of Misclassification, a=1.0



three testing strategies on $P(M|\theta)$ is the same as their ordering on conditional standard errors of measurement at any level of θ. It can then be seen from these curves that a practical adaptive test can provide a lower expected probability of misclassification if it approximates the ideal adaptive test. How well a given adaptive testing strategy approximates the ideal is, or course, an empirical question.

Figure 17 presents the $P(M|\theta)$ curves for a ten-item conventional test, with difficulties peaked at b=0.0, and a ten-item Bayesian adaptive test, both with item discrimination fixed at a=1.0 and both scored by Owen's method. The curves appear very similar, being high near the cutting point (indicating a high probability of making an error) and low distant from the cutting point. The conventional test allows somewhat better decisions for values of θ nearer to the cutting score. The differences in the conditional probability of misclassification function yield a very small difference between unconditional probability of misclassification values for the two strategies, which were .120 for the conventional test and .122 for the Bayesian test. (Unconditional probabilities are shown in parentheses beside the legend in Figure 17 and successive figures.)

Figure 18
Conditional Probability of Misclassification, a=2.0

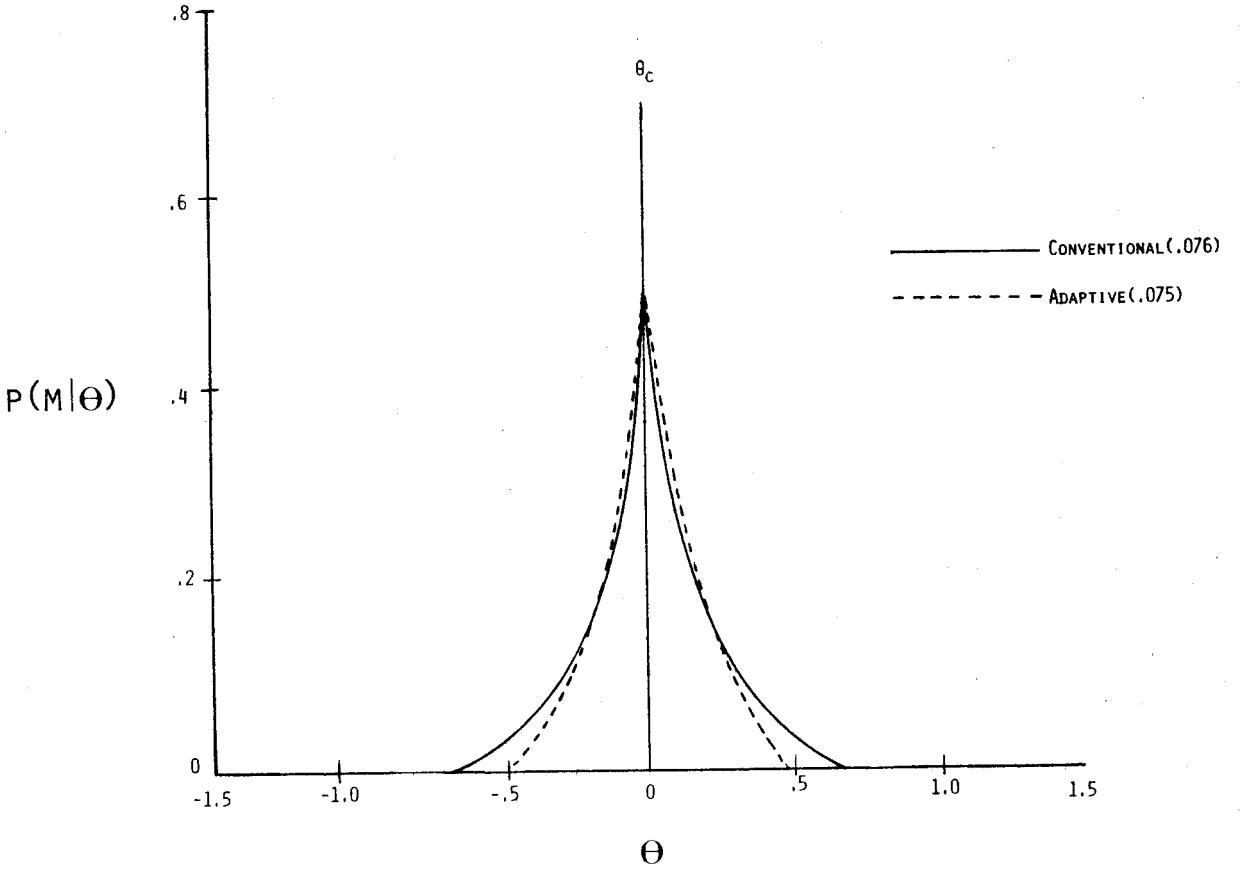


Figure 18 shows $P(M|\theta)$ curves for the same strategies with item discriminations of a=2.0. The same general results were obtained, except that the differences

at values of θ distant from the cutting score were more pronounced, and the range of superiority of the conventional test was smaller. Due to the N(0,1) shape of the ability distribution, however, small differences near the cutting point are as important in the determination of the expected probability of misclassification as large differences distant from the cutting point. Difference in expected probability was still very low (.076 versus .075).

Figure 19
Conditional Probability of Misclassification, a=3.0

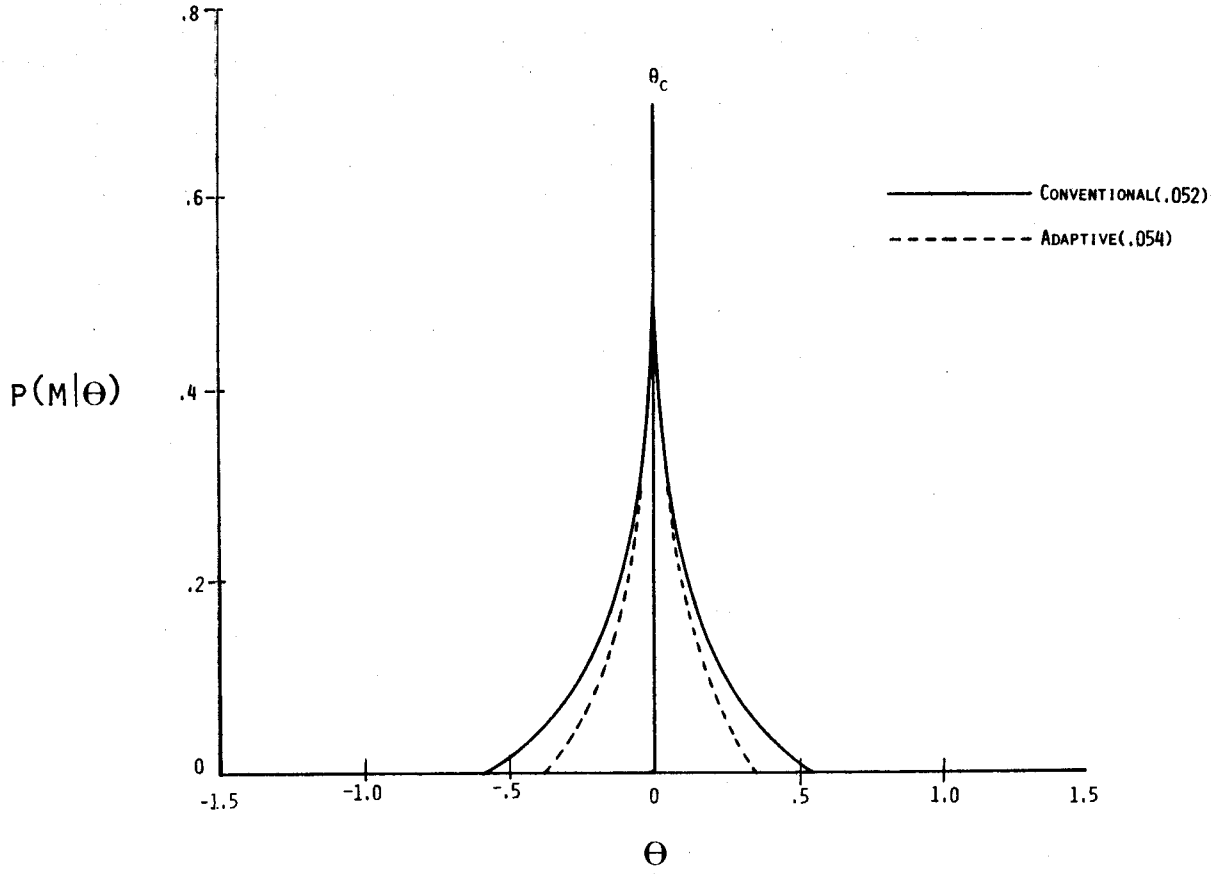


Figure 19 shows curves for tests with high item discrimination (a=3.0). Again, similar results were obtained and the difference in expected probability of misclassification was still small (.052 versus .054).

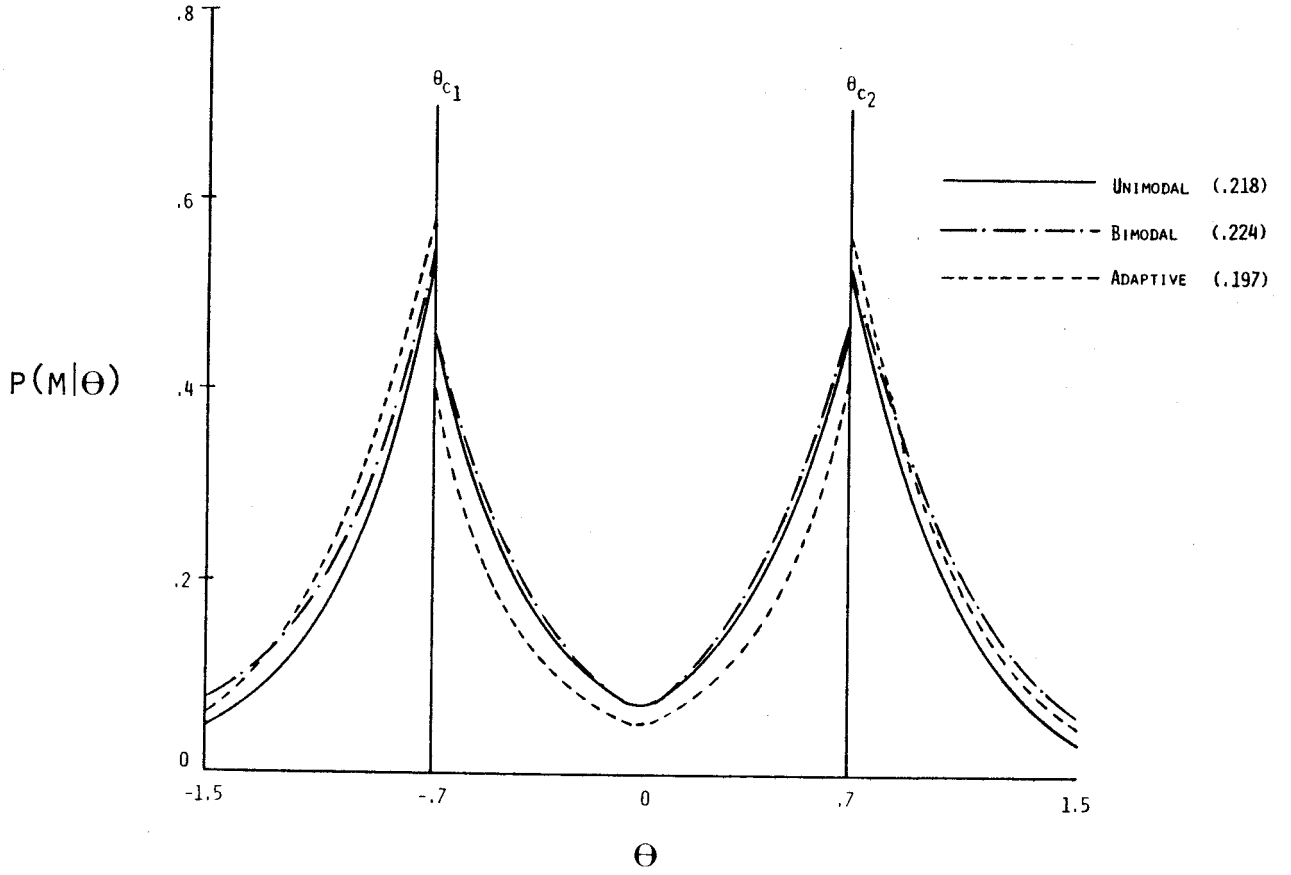
These results suggest that an adaptive test makes classification decisions about as well as a conventional test in this simple case where a conventional test should perform better in comparison to an adaptive test. However, it should be noted that the conventional test was superior to the adaptive test in an increasingly narrower range of θ with increasing item discriminations.

More than One Cutting Score

Design of conventional tests is more complicated, however, when the cutting scores deviate from the center of the ability distribution. A given increase in information, which corresponds to a given decrease in standard error, has its

greatest effect on the conditional probability of misclassification at ability levels near a cutting score. This suggests that items should be peaked at the cutting scores. But a given reduction in conditional probability of misclassification has its greatest effect on the expected probability of misclassification at levels of ability where most of the people are located. This, assuming $\theta \sim N(0,1)$, suggests peaking the item difficulties at b=0.0. As a result, when the cutting score is at some value of θ other than 0.0, the two suggestions are in conflict. The optimal point(s) to peak the difficulties will be some function of the location of the cutting scores, the discriminating powers of the items, and the underlying ability distribution. Determination of such an optimal design of a conventional test is beyond the scope of this paper. However, comparisons of some standard conventional test designs with an adaptive test will be informative.

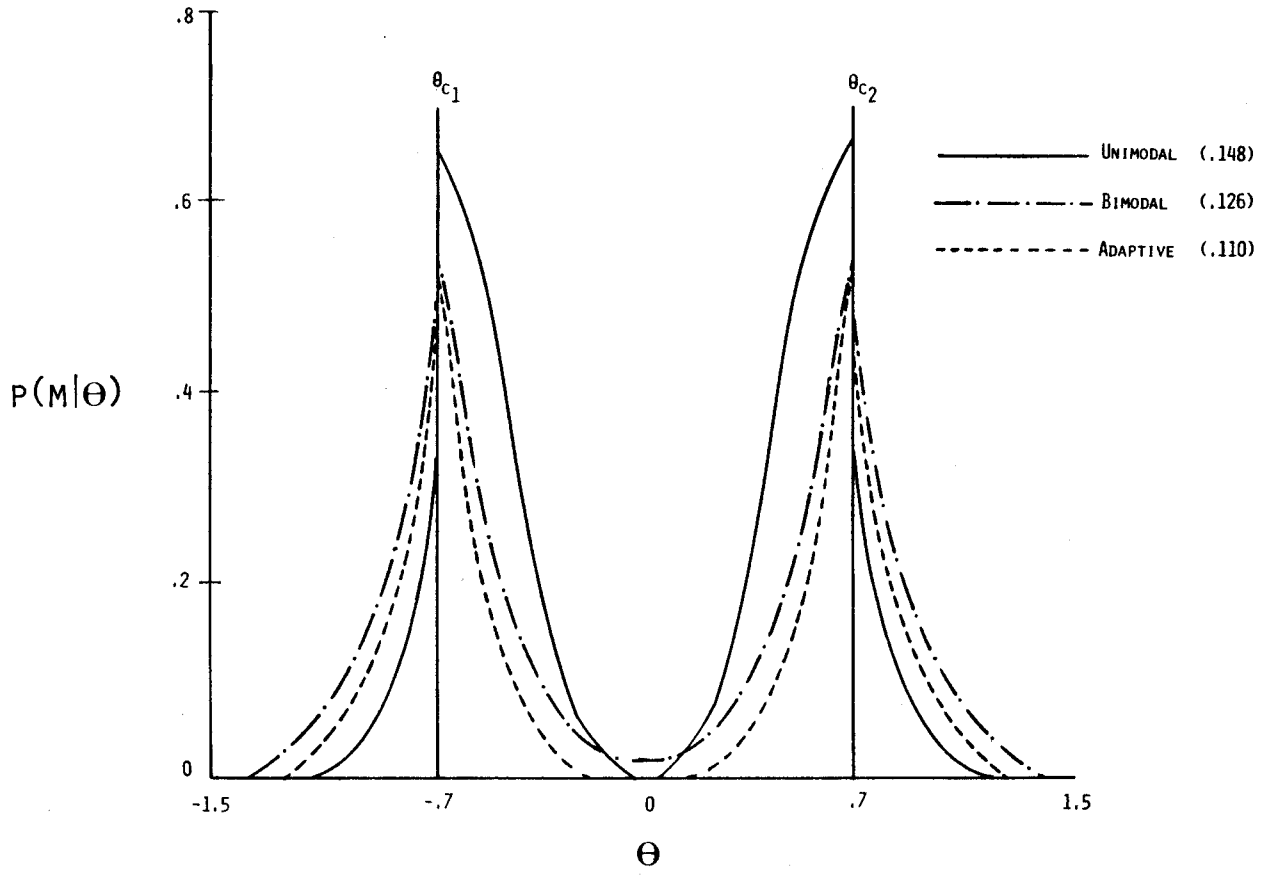
Figure 20
Conditional Probability of Misclassification, a=1.0



Assume that there are two cutting scores, one at θ_c=-.7 and the other at θ_c=.7, and that all errors of misclassification are equivalent in terms of importance. One classical approach to designing a conventional test involves peaking half of the items at each of the two cutting scores, where the fine distinctions need to be made; such a test can be referred to as a bimodal conventional test. Another approach is to peak all the items at b=0.0; this test can be called a unimodal conventional test.

Figures 20 through 22 present the conditional probabilities of misclassification for each of the unimodal and bimodal conventional tests, and the Bayesian adaptive test, at three levels of item discrimination. Figure 20 shows the curves for the case when a=1.0. There is little suggestion in Figure 20 as to which strategy is better. But an interesting discontinuity is observed for estimates from all testing strategies at the cut points. This characteristic is due to the fact that, for finite-length tests (which include 10-item tests like those used here), the Owen's Bayesian score is biased (i.e., the expected value of the score at a given level of θ is not θ). Specifically, in this case, the Bayesian score is biased in the vicinity of the cutting scores toward the center of the population ability distribution at θ=0.0. This causes more testees to be classified into the middle interval than would be by an unbiased score. The effect is that fewer errors of classification are made for ability levels in the middle interval and more are made for individuals in the two extreme intervals. Comparing expected probabilities of misclassification, the adaptive test yields the lowest probability (.197) and the bimodal conventional, the highest (.224).

Figure 21
Conditional Probability of Misclassification, a=2.0

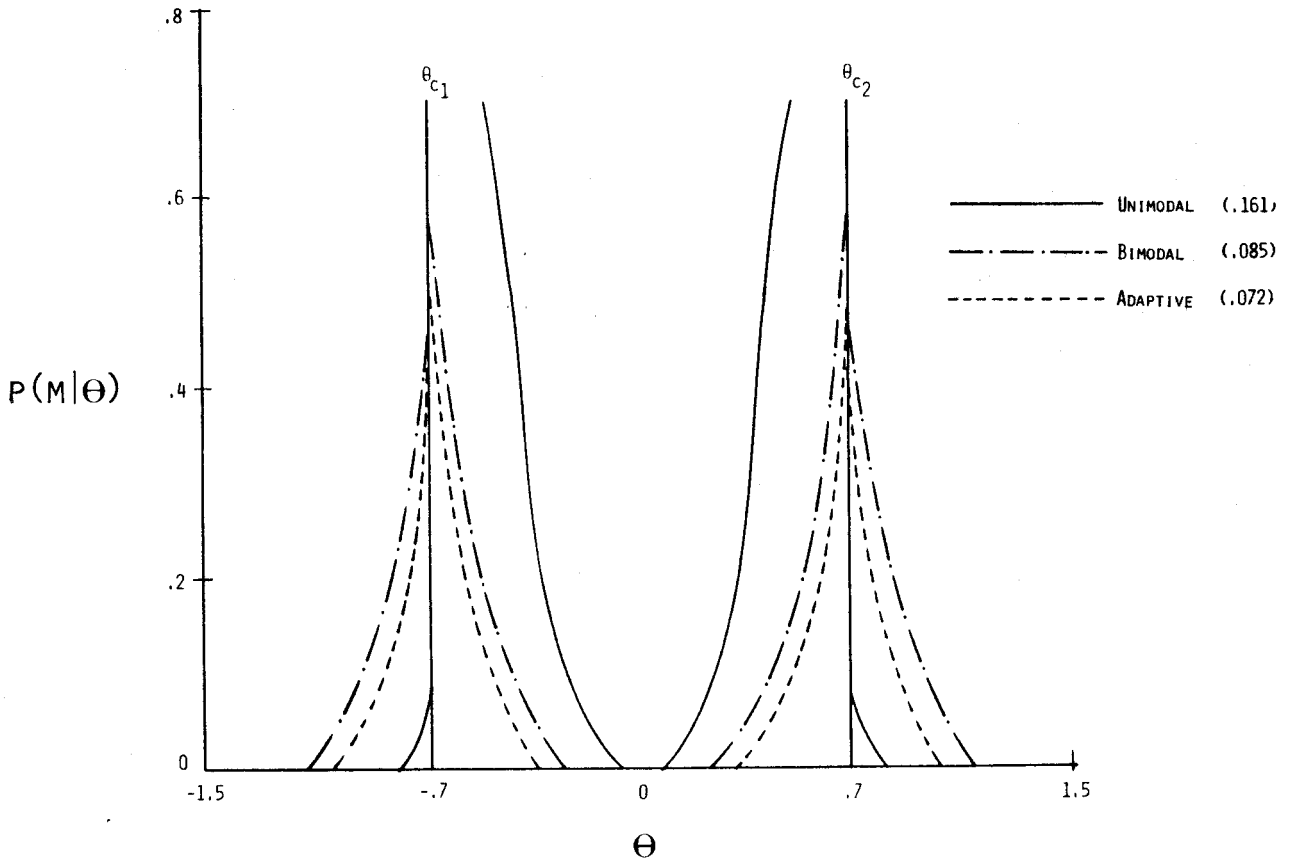


It is difficult to say in this case, however, whether the adaptive test provides a lower expected probability of misclassification because it makes better decisions or because it is conservative. The conservatism results in more classifi-

cations errors in the extreme categories, and fewer errors at central ability levels where more individuals' ability levels lie.

When α=2.0 (Figure 21), the unimodal conventional test shows pronounced discontinuity suggesting that scores are too extreme near the cutting points. The adaptive test provides the smallest conditional probabilities of misclassification over most of the ability range. It makes a few more errors in the extreme intervals than does the unimodal conventional test, but the unimodal test's superiority is offset by extreme error rates in the middle interval. In terms of expected probabilities of misclassification, the adaptive test is again superior [P(M)=.110]. With an expected probability of misclassification of .126, the bimodal conventional test, its nearest competitor, is expected to make 1.15 times as many errors of classification.

Figure 22
Conditional Probability of Misclassification, α=3.0



When α=3.0, as shown in Figure 22, the same general results were obtained. The expected probability of misclassification for the bimodal conventional test (.085) was 1.18 times as large as that of the Bayesian adaptive test (.072). It should be noted, however, that items this discriminating are rare in practice.

Utility Comparisons

It is tempting to take these values at this point and say that adaptive testing can greatly reduce overall errors of classification by up to 15 percent in a realistic classification situation. But, as was discussed earlier, the errors of classification presented thus far are based on a latent ability continuum rather than an observable success-failure continuum. Using the utility functions presented earlier and choosing the decision yielding the highest expected utility for the estimate of ability, average utilities for the bimodal conventional test (the best conventional test in previous comparisons) and the Bayesian test were .808 and .820, respectively, using the items of a=1.0. For the same sample of abilities and a=2.0, the utilities were .831 and .849. With a=3.0, the values were .855 and .858. Whether these differences are practically significant depends on what these units of utility mean in a particular context. But such utilities (of which these are only an example) must ultimately be considered in determining the comparative values of conventional versus adaptive testing for classification decisions.

Conclusions

These results suggest that adaptive testing may offer important advantages to an organization involved in making classification (e.g., selection and placement) decisions. Specifically, the data show that while a conventional test classifies as well as an adaptive test when there is one cutting score at the middle of the ability distribution, an adaptive test will provide better categorization when there is more than one. The determination of the cost effectiveness of adaptive testing in an organization, however, will depend on the utility functions specified by the organization.

APPLICATIONS OF ITEM CHARACTERISTIC CURVE THEORY TO THE PROBLEM OF TEST BIAS

STEVEN M. PINE
University of Minnesota

One of the most challenging and important issues facing test developers and users today is whether or not ability tests are biased against minority groups, and if so, how test bias can be reduced. In recent years, there has been considerable research activity concerned with the identification and reduction of bias and unfairness in various settings. For the most part, these efforts have been unsuccessful. One possible reason for this lack of progress is the fact that almost all the research on test bias and fairness has been based on classical test theory.

In his recent review of test theory, Lumsden (1976) refers to the true score model of classical test theory as the "Model-T Theory" and suggests that classical test theory reflects a very restricted range of test behavior. For example, classical test theory emphasizes group-oriented measurement; but group-oriented measurement is likely to be unproductive if tests are to be relevant to individuals of varied backgrounds. Consequently, it is unlikely that this approach will be useful in resolving problems as complex as those involved in test bias.

Bias in testing is caused by the failure of tests to take into account a number of important variables in their construction, administration, and scoring (Angoff, 1975; Green, 1976; Pine & Weiss, 1976; Sattler, 1974). These variables include individual differences in motivation, ethnic background and related variables.

Tests based on classical test theory may ignore certain types of individual differences because they are constructed using item statistics which can be expected to vary between population subgroups, and because they require all testees to take identical test items. If progress is to be made in this critical research area, a test theory that permits the testing process to be adapted or tailored to individuals is needed. This capability now exists in the form of item characteristic curve theory, coupled with the technology of adaptive test administration.

An Item Response Model of Bias

Item characteristic curve theory. Recently, a new test theory called "item characteristic curve (or latent trait) theory," specifically designed for the measurement of individuals, has emerged. Item characteristic curve theory (Lord & Novick, 1968) is based on the idea that the responses which individuals make to a given ability test item are determined by their ability on one or more underlying dimensions (latent traits), and the parameters of the test items, i.e., their difficulty, discriminating power, and probability of being guessed correctly by chance. This idea is expressed mathematically by the Item Characteristic Curve (ICC) which gives the probability that a testee with a given ability level on the underlying dimension will correctly answer a given test item.

This research is supported by contract N00014-76-C-0244, NR No. 150-383, with the Personnel and Training Research Programs, Office of Naval Research.

The ICC curves and their associated item parameters are the building blocks of this new test theory. Once item parameters are determined for each test item, they can be used to describe how individuals at a given ability level are likely to perform on each item. ICC theory allows probabilistic statements to be made about the ability level of testees regardless of their subgroup membership or which subset of items they have been administered. This property provides a means for creating tests which can be adapted to individual testees since it is no longer necessary that identical items be administered to every testee, thus making ICC theory potentially valuable for developing less biased tests. Furthermore, the bias-reducing potential of ICC theory is not tied to its use with any particular testing strategy, although the greatest benefits can be expected when it is used in conjunction with adaptive testing (Pine & Weiss, 1977; Weiss, 1974).

Definition of item bias. *A test item can be considered to be unbiased if all individuals having the same underlying ability level have an equal probability of correctly answering the item, regardless of their subgroup membership.*

As indicated, the ICC gives the probability of correctly answering an item at a given ability level. Therefore, the above definition of an unbiased item is equivalent to requiring that a test item have the same ICC for all subgroups. Since an ICC is described by its difficulty, discrimination and guessing parameters, this is also equivalent to requiring that the values of these parameters be invariant within a linear transformation from subgroup to subgroup. The linear transformation assumption is necessary to account for the fact that subgroups in which the parameters are calculated may have ability distributions with different means and variances.

Applying the Model to Detect Test Bias

The following discussion is restricted to tests that consist entirely of homogeneous items. Homogeneity implies that the items measure essentially one ability dimension. This definition allows for the possibility that a homogeneous set of items may measure one or more extraneous dimensions in addition to the single primary dimension which the test is purported to measure. For instance, test items intended to measure vocabulary ability may inadvertently also measure several cultural variables. Although the present discussion is restricted to homogeneous items, the concepts developed here could in principle be extended to the multidimensional case.

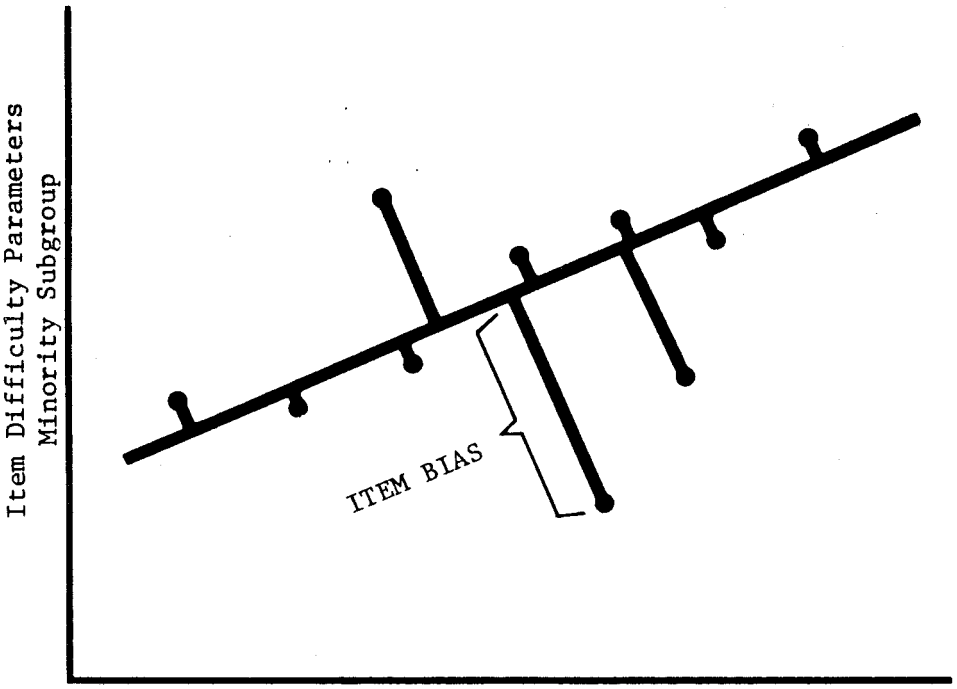
It is also assumed here that test items fit an underlying response model for all subgroups. This model is the function which specifies the shape of the ICC curve and indicates, at each ability level, the probability that an individual at that level will correctly answer the administered item. This constraint is not as limiting as it may appear to be, since one can empirically test the fit of the item data to the assumed response model and eliminate those items that do not fit prior to carrying out any of the analyses described here.

Given the above restrictions, the first step in investigating whether a set of items is biased is to screen out those items which do not fit the underlying response model. Most of the existing computer programs for estimating item response

parameters (e.g., Urry, 1974a; Wingersky & Lord, 1973) reject items that do not fit the assumed model as a matter of course. Therefore, with these programs, it can be assumed that all items for which parameter values are available fit the response model.

The next step is to demonstrate that these items are homogeneous, i.e., the same trait accounts for the major portion of underlying variance in each subgroup's inter-item correlation matrix. If they are homogeneous, Lord and Novick (1968, pp. 359-360) have shown that their item response parameters will be invariant (within a linear transformation) across subgroups. According to the definitions given earlier, invariant test items are unbiased. Therefore, a *sufficient* method for demonstrating that a set of test items is unbiased is first to factor analyze the matrix of inter-item correlation coefficients within each of two or more subgroups and demonstrate that the same single factor accounts for the major portion of variance in each subgroup's matrix, and then show that this is the factor that the test was intended to measure.

Figure 23
Item Bias Shown as a Perpendicular Distance
in a Scatter Plot of Subgroup Item Difficulties



A second approach for determining whether a set of test items is biased is also implicit in the work of Lord and Novick. If the same dimension underlies a set of test items for a population of testees (which would, therefore, make the items unbiased), the item parameters for any two subgroups in the population should have

a linear relationship (Lord & Novick, 1968, p. 380). This condition can be tested directly by plotting the discrimination (a), difficulty (b), or guessing (c) parameters of a set of items derived from one subgroup against those from another and and testing for linearity. A plot of this type, based on the item response difficulty parameters for a 10-item test, is shown in Figure 23. If factor analysis indicates that a single dimension underlies a set of items, the presence of a linear relation between subgroups for ICC parameters is both a necessary and sufficient demonstration that these items are unbiased.

In Figure 23, the perpendicular distance between each item and the best fitting line through all the points can be interpreted as the degree of item bias; the greater the distance, the more item bias is implied. By comparing the relative item parameter values between subgroups, it is possible to identify the specific test items which contribute the most to a non-linear relationship between subgroup parameters. In the language of analysis of variance, this non-linear relationship would be an item-by-group interaction. Plots similar to Figure 23 and related interpretations could also be made for item discrimination and guessing parameters.

The degree-of-item-bias index illustrated in Figure 23 has several applications. It could be used to screen out the most biased items during the construction of a conventional test. Or, it could be used within an adaptive testing framework as an additional criterion for item selection.

The assessment of item bias by plotting a scatter diagram of item parameters for one subgroup against another is not in itself new. A very similar method has been used at Educational Testing Service (ETS) for several years. The essential difference between the present method and the ETS method is that ETS uses item parameters based on classical test theory. It can be shown (Lord & Novick, 1968, p. 301) that classical item parameters will generally not be linearly related across subgroups of a population. This means that the test for bias using classical parameters can lead to an artifactual detection of bias. Furthermore, the difficulty parameter of classical test theory is confounded by level of discrimination and guessing effects (Urry, 1974b). Thus, if an item is relatively more difficult for one subgroup than another, it is not clear whether this is because the item varies only on difficulty, or whether this result is caused by differences in discrimination and/or guessing. The item parameters from ICC theory, on the other hand, provide relatively unconfounded measures of difficulty, discrimination, and guessing. Therefore, by plotting these parameters on separate graphs, it is possible to determine exactly why an item is biased. For instance, it may be that a given item is biased not because it is relatively more difficult for a minority subgroup, but because that subgroup is less effective at guessing. This kind of detailed analysis is impossible using classical item parameters.

Another interesting consideration in the use of ICC versus classical item parameters is the fact that if classical item parameters are linearly related among subgroups, thereby implying an unbiased set of items, ICC parameters will of necessity *not* be linearly related and will, therefore, imply the presence of bias in these same items. This fact would seem to have particular relevance for the work of researchers such as Jensen (1975) who have concluded that tests are generally *not* biased against Blacks based on the presence of a linear relationship between classical item parameters correlated across Black and White subgroups.

An example with real data. To demonstrate how these analyses might be used and interpreted, they have been applied to the difficulty parameter from 75 multiple-choice vocabulary items administered in a racially mixed high school in Minneapolis. The sample sizes in this study were not optimal (58 Blacks, 168 Whites), but the data provide a good example of the technique.

First the homogeneity assumption was tested by factor analyzing the inter-item correlation matrices. A subset of 45 items was chosen and two tetrachoric intercorrelation matrices were calculated, one for the Black and one for the White subsamples. The matrices were then factor analyzed using the principal axis method; communalities were estimated using the highest off-diagonal entry for each item, and the factor solution was iterated until the estimated communalities stabilized. Eight factors were extracted from each matrix, in each case accounting for all of the estimated common variance. The eigenvalues from the two factor analyses are shown in Table 2.

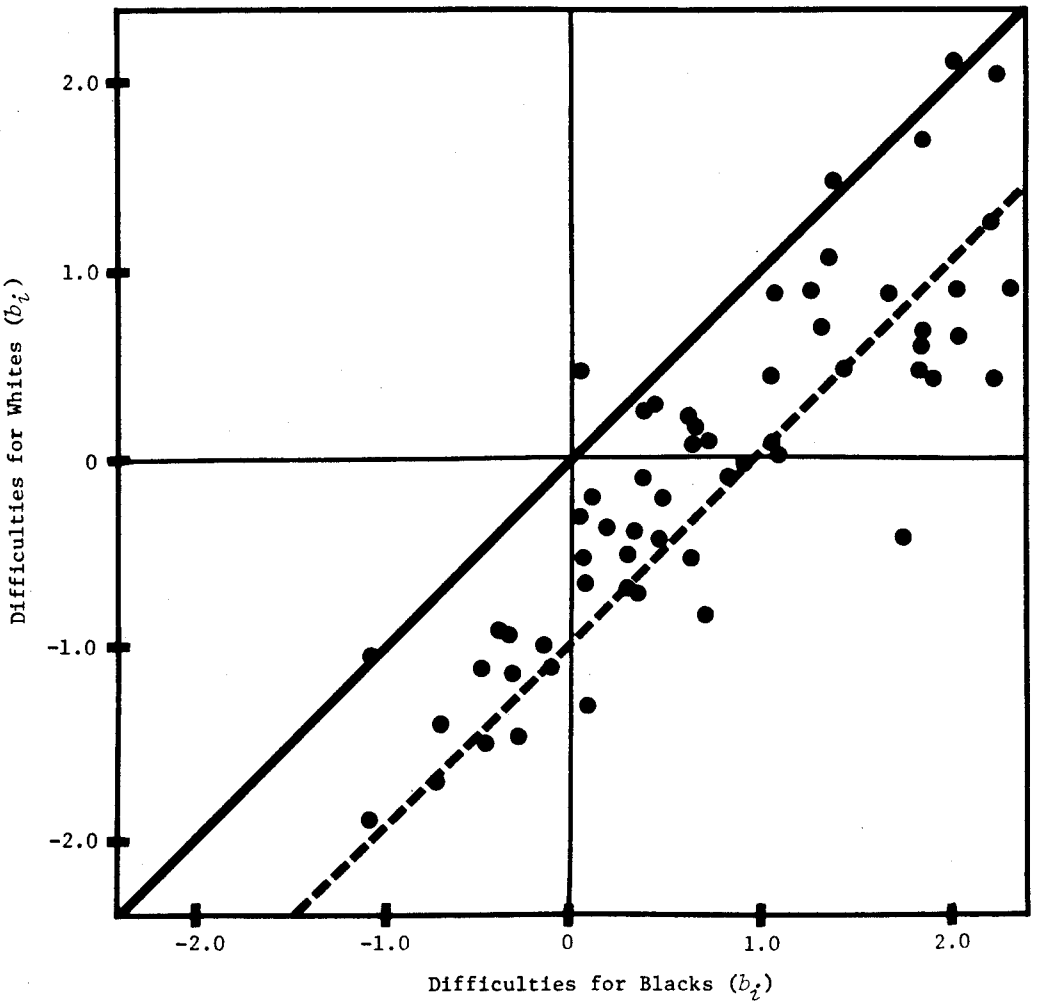
Table 2
Eigenvalues from Factor Analyses of Black and White Subgroup Item-Intercorrelation Matrices

Subgroup	Factor	Eigenvalue	Percent of Common Variance	Cumulative Percent
Whites				
	1	19.26	64.8	64.8
	2	2.32	7.8	72.7
	3	1.67	5.6	78.3
	4	1.58	5.3	83.7
	5	1.37	4.6	88.3
	6	1.20	4.1	92.4
	7	1.18	4.0	96.4
	8	1.08	3.6	100.0
Blacks				
	1	16.33	47.9	47.9
	2	3.70	10.9	58.7
	3	3.01	8.8	67.5
	4	2.64	7.7	75.3
	5	2.35	6.9	82.2
	6	2.26	6.6	88.8
	7	2.06	6.0	94.9
	8	1.75	5.1	100.0

For both the Black and the White data, the first eigenvalue was very large in comparison to the remaining eigenvalues, providing evidence supportive of the unidimensionality assumption. Furthermore, the items appear to be measuring the same dimension in both subgroups, since the coefficient of congruence (Rummel, 1970, p. 461) calculated between the 45 corresponding loadings for Factor 1 in the two subgroups was .97. It also seems reasonable to conclude, based on the pattern of loadings, that Factor 1 is measuring vocabulary ability.

The results of a further analysis of bias for these 75 items are shown in Figure 24. The scatter plot in Figure 24 is based on the estimated ICC difficulty parameter values calculated separately for the White and Black subsamples.

Figure 24
Graphical Analysis of the Bias in 75 Multiple-Choice Vocabulary Items



The data plotted in Figure 24 show that almost all of the items are relatively more difficult for Blacks than for Whites. This is indicated by the fact that the dots representing the items tend to fall below the diagonal line. If the items were equally difficult for Blacks and Whites, the data points would fall on this line.

However, the mere fact that the items are relatively more difficult for Blacks cannot necessarily be taken as an indication of bias, since bias in the test items is assessed by evaluating the degree of linearity in the plot. The Pearson product-

moment correlation coefficient between the item parameter values for Blacks and Whites is r=.86, indicating a high degree of linear relationship. This is consistent with the results of the factor analysis and suggests that these vocabulary items, when taken as a group, are essentially unbiased. It is possible, however, that even though the items taken as a group are unbiased, one or more of the items taken individually might be biased. For instance, in these data, several items appear to have larger departures from the dotted line fitted through the item points in Figure 24. Of course, it is possible that these large departures may be due only to sampling error. To eliminate possible misinterpretations that would occur if this were the case, a technique is under development to establish confidence limits for the best fitting line. This technique will permit the identification, with some known degree of confidence, of biased items.

Related Developments

The material presented here is only one example of how item characteristic curve theory can potentially be applied to the problem of test bias. It is only a small part of the research related to test bias and unfairness currently underway at the University of Minnesota.

Additional developments involve a method of correcting for bias in the ICC item parameters. Very briefly, this method consists of determining item parameter estimates that will depend only on the extent to which an item loads on the factor it is supposed to be measuring. In essence, this approach is based on the notion that to obtain unbiased test items, all that is necessary is to know how each test item behaves (i.e., what its parameters are) in the various subgroups which comprise our test population. Using the method now under development, bias in an item can be eliminated by correcting its parameter values to account for the degree of bias. Then, if the resulting ability estimates are based not on the total number of correct answers, but on some function of the corrected item parameter values, the resulting ability estimates will be unbiased.

This method for correcting item bias is now being studied by computer simulation techniques. In this way, the bias-corrected item parameter values can be directly compared to the known, true item parameter values. If the results of these studies are favorable, the technique will permit the reduction or elimination of the effects of item bias on ability test scores.

Does this mean that we can now write the final chapter on test unfairness? Not at all! First, some may disagree that bias has been eliminated as long as differences exist in the mean test scores of various subgroups. Secondly, bias in the estimation of item parameters is only one source of possible unfairness in the testing process. A test can be unfair for a myriad of other reasons, including those attributable to elements in the testing environment, and to the psychometric properties of the procedure used to select and administer test items (Pine & Weiss, 1977; Weiss, 1975). To explore the possible psychometric influences on test unfairness, a series of computer simulations designed to investigate how item characteristics interact with the choice of a testing strategy is currently in progress. Also in progress is a live computerized testing study designed to investigate how well some of the bias-reducing procedures described in this paper operate in a real test administration. This study will also investigate a computerized adaptive test designed explicitly to reduce bias in test scores. In addition, the study is designed to replicate a previous finding that computerized tests increase the test-taking motivation of minority testees (Betz & Weiss, 1976b; Weiss, 1976).

APPLICATIONS OF ADAPTIVE TESTING IN MEASURING ACHIEVEMENT AND PERFORMANCE

ISAAC I. BEJAR
University of Minnesota

The purpose of achievement testing is to locate individuals on an achievement scale. Usually, to interpret achievement test scores, a transformation is applied to the scores which allows an interpretation in terms of the relative standing of an individual with respect to the norming group. In many instructional settings, this interpretation is not adequate and, as a result, instructional personnel have requested more concrete kinds of interpretation. Criterion-referenced testing, mastery testing and similar approaches have been developed to meet these needs.

What is unique about criterion-referenced and mastery testing is that the items that constitute the test are sampled from a population of items which is isomorphic with the objectives of the instructional program in which achievement is to be measured (Shoemaker, 1975). Because of this, it is possible to interpret scores in terms of the specific areas of achievement that a student has mastered in relation to the objectives of the instructional program.

Undoubtedly, this attention to content is bound to increase the quality of achievement test scores. However, the degree of improvement possible in achievement test scores using any approach to achievement test construction is limited by the nature of the test item. When typical multiple-choice test items are used, a very limited range of student performance is measured. The cognitive skills involved appear to be the processes of recall of information coupled with recognition of the correct answer, and the result is usually expressed as either "correct" or "incorrect". However, achievement or knowledge is seldom all or none, and proceeding as if it were, as in the typical "correct-incorrect" multiple-choice achievement test, does not extract all the potential information about an individual's achievement level. This paper describes research concerned with the integration of testing procedures which take partial information into account with methods of computerized adaptive achievement test administration, and discusses some implications of this research for performance testing.

Partial Knowledge

Background. Intuitively it seems clear that extracting partial knowledge from test responses should lead to better assessment of achievement. However, the research literature (e.g., Wang & Stanley, 1970) does not show consistent

---

This research is supported by contract N00014-76-C-0627, NR 150-389, with the Personnel and Training Research Programs, Office of Naval Research.

increases in both reliability and validity when partial knowledge is taken into account. The results of the typical investigation (e.g., Hakstian & Kansup, 1975) show that, while reliability is usually increased by taking partial knowledge into account, the validity of the scores remains the same or even diminishes. Such findings are usually interpreted as evidence against the usefulness of the assessment of partial knowledge. However, a careful consideration of the problem suggests that something is amiss. One possible explanation is that the test and the criterion are not unidimensional.

To illustrate, consider two tests, A and B, measuring a single construct. Test B can be referred to as the "criterion test"; the correlation between A and B will be referred to as the validity of Test A. Both Test A and Test B correlate .60 with the construct. This can be summarized as follows:

$$\Lambda = \begin{matrix}\text{Test} \\ \text{A} \\ \text{B}\end{matrix}\begin{bmatrix}.60 \\ .60\end{bmatrix} \qquad [13]$$

Then the intertest correlation matrix can be expressed (Jöreskog, 1971; Maxwell, 1971) as Equation 14.

$$\Sigma = \Lambda\Lambda' + \Psi^2, \qquad [14]$$

where Ψ^2 is a diagonal matrix of error variances. For the Λ in Equation 13, Equation 14 becomes,

$$\Sigma = \overset{\Lambda}{\begin{bmatrix}.60 \\ .60\end{bmatrix}} \overset{\Lambda'}{\begin{bmatrix}.60 & .60\end{bmatrix}} + \overset{\Psi^2}{\begin{bmatrix}.64 & .00 \\ .00 & .64\end{bmatrix}}$$

$$= \overset{\Lambda\Lambda'}{\begin{bmatrix}.36 & .36 \\ .36 & .36\end{bmatrix}} + \overset{\Psi^2}{\begin{bmatrix}.64 & .00 \\ .00 & .64\end{bmatrix}}$$

$$= \overset{\Lambda\Lambda' + \Psi^2}{\begin{bmatrix}1.00 & .36 \\ .36 & 1.00\end{bmatrix}} \qquad [15]$$

The off-diagonal element of $\Lambda\Lambda'$ is equal to the validity of A and the diagonal elements are reliabilities. In this case both A and B have reliabilities of .36 and the validity of Test A is .36.

Now, suppose Test A is administered under conditions that allow for partial knowledge and that, as a result, its correlation with the construct goes from .60 to .70. Following the same procedure shown in Equation 15, the reliability of Test A becomes .49 while that of Test B remains at .36. At the same time, the validity of Test A increases from .36 to .42. In short, when there is a single common factor underlying the responses to a criterion and a predictor, an increase in the reliability of the predictor will lead to an increase in its validity. This is not so when more than one factor is common.

To illustrate this, assume that Tests A and B, both administered conventionally, have in common a method factor (θ_m), in addition to the construct, and that both correlate .40 with it. That is,

$$\Lambda = \begin{matrix} \theta_c & \theta_m \\ \end{matrix} \begin{bmatrix} .60 & .40 \\ .60 & .40 \end{bmatrix} \begin{matrix} \text{Test} \\ \text{A} \\ \text{B} \end{matrix} \qquad [16]$$

Assuming that the construct and the method factor are uncorrelated, the correlation matrix for Tests A and B, according to the model in Equation 14, is given by:

$$\Sigma = \overset{\Lambda}{\begin{bmatrix} .60 & .40 \\ .60 & .40 \end{bmatrix}} \overset{\Lambda'}{\begin{bmatrix} .60 & .60 \\ .40 & .40 \end{bmatrix}} + \overset{\Psi^2}{\begin{bmatrix} .48 & .00 \\ .00 & .48 \end{bmatrix}}$$

$$= \overset{\Lambda\Lambda'}{\begin{bmatrix} .52 & .52 \\ .52 & .52 \end{bmatrix}} + \overset{\Psi^2}{\begin{bmatrix} .48 & .00 \\ .00 & .40 \end{bmatrix}}$$

$$= \begin{bmatrix} 1.00 & .52 \\ .52 & 1.00 \end{bmatrix} \qquad [17]$$

In this case, the validity of Test A is .52.

Now, suppose that the same Test A is again administered under conditions that allow for the scoring of partial information and that, as a result of this, its correlation with the construct becomes .70. At the same time the correlation of Test A with the method factor drops from .40 to .20; i.e., Λ becomes:

$$\Lambda = \begin{matrix} \theta_c & \theta_m \\ \end{matrix} \begin{bmatrix} .70 & .20 \\ .60 & .40 \end{bmatrix} \begin{matrix} \text{Test A (with partial knowledge)} \\ \text{Test B} \end{matrix} \qquad [18]$$

and

$$\Lambda\Lambda' = \begin{bmatrix} .53 & .50 \\ .50 & .52 \end{bmatrix} \qquad [19]$$

Thus, as a result of introducing partial knowledge, the validity was reduced from .52 to .50. However, it is clear that this seemingly disappointing result is not inconsistent with the true improvement that occurred, namely an increase in the correlation of Test A with the construct.

Although this example contains many assumptions, it seems that something similar occurs with real data. Hakstian and Kansup (1975) compared the validity of a verbal ability test administered under conventional and elimination scoring (Coombs, Millholland, & Womer, 1956) instructions. Validity was defined as the

correlation with school grades in language arts. This correlation was .49 under conventional administration and .39 under elimination scoring. However, the correlation with another verbal ability test was .59 under conventional scoring and .67 under elimination scoring. Thus, when validity is defined as the correlation with school grades, elimination scoring appears to be less valid; but when validity is defined as the correlation with another verbal ability score, elimination scoring is more valid. These results are not contradictory but simply provide evidence of the fact that performance on verbal ability tests measured either with multiple-choice or elimination items is explained by the same ability, whereas school grades on language arts do not depend exclusively on verbal ability.

Advantages of using partial information. If methods for the assessment of partial knowledge are to yield improved test scores, the tests must be such that there will be an opportunity for partial knowledge to emerge. With few exceptions, most notably Coombs et al. (1956), the presence of partial knowledge is never tested. Some theoretical results suggest that when partial knowledge is allowed to emerge and is scored, dramatic improvements in test scores follow.

To illustrate this, consider the information functions of two latent trait models. Information at a given point on the underlying trait is the reciprocal of the variance of the maximum likelihood estimator at that point. Therefore, the larger the information value, the more precise is the estimate of the location of an individual on the trait. One latent trait model studied was the two-parameter normal ogive (Lord & Novick, 1968, Chap. 16) which is appropriate for dichotomous scoring. The other model was Samejima's (1969) graded response model, which is an extension of the two-parameter normal ogive model to polychotomous scoring. Information levels of the graded model can be considered to be the case when partial knowledge is taken into account, whereas the information provided by the dichotomous model is that provided when partial information is ignored.

To simplify the comparison, the mean information for each model was computed, assuming that the underlying trait was normally distributed. In addition, it was assumed that each test consisted of 60 items, each having the same item-trait correlation (r). The distribution of item difficulty in the dichotomous case can be described as a truncated normal distribution with a mean of 0.0 and maximum and minimum equal to $1/r$ and $-1/r$, respectively. The distribution of difficulty of the highest category in the graded model was also a truncated normal distribution but with a mean of $.40/r$ and maximum and minimum $1/r$ and $-.20/r$. Within each graded item, the difficulty of each of the lower categories was set in such a way that the categories would be chosen by the same proportion of testees. The comparison assumes that there are five graded-response categories. This choice of difficulties approaches the optimal conditions for the two models.

The ratio of the mean information for the graded model over that of the dichotomous model for several levels of test homogeneity is seen in Table 3. For example, at an item-trait correlation of r =.55 the ratio was 1.42. This

means that, on the average, the use of partial knowledge will be 42% more informative than if it is ignored. Note that this improvement, due to incorporating partial information into the scores, increased as the discrimination of the test increased. In other words, the better the test, the more it will benefit from adding partial knowledge. This is also true when reliability rather than information is used as the evaluative criterion (Bejar & Weiss, in press).

Table 3
Ratio of Mean Information of Graded to
Dichotomous Model, as a Function of Item-Trait Correlation

	Item-Trait correlation					
	.55	.63	.71	.77	.84	.95
Ratio of mean information	1.42	1.43	1.48	1.52	1.58	1.90

The advantages derived from taking partial knowledge into account can only materialize under the proper conditions. In the typical multiple-choice test item, even though partial knowledge influences which alternative is chosen, the response is scored as correct or incorrect. One way of allowing credit to be given for partial knowledge is to instruct testees to segregate alternatives into different categories. Coombs' (1956) procedure is an instance of the approach where the categories are "correct" and "incorrect". Other categories are possible, though; e.g., verbal items may be classified as "synonyms", "antonyms", or "neither".

Computerized Testing

Recording and scoring responses to non-dichotomous test items is not, however, convenient with paper-and-pencil test administration. One obvious use of interactive computers, therefore, is to handle the recording and scoring of responses to non-dichotomous achievement test items. But, as previous presentations in this report suggest, the computer can also be used to adapt or tailor the test to each individual.

These presentations (and indeed most of the research in computerized adaptive testing) have been oriented toward ability measurement. In achievement testing, it is possible to distinguish between two kinds of adaptive test administration: One involves adapting the length of the test; in the other, the difficulty of the test is adapted.

Adapting the length of the test to the individual is appropriate in instructional settings where each individual is allowed as much time as is necessary to complete a given unit of instruction. Under those conditions, individual differences with respect to knowledge are minimized and it becomes profitable to adapt the length of the test rather than its difficulty. The research of Ferguson (1970) is an example of this type of adaptive testing. In his system, an individual is tested until he is classified into a non-mastery or mastery category. The statistical basis of this system is Wald's sequential likelihood ratio test. Ferguson's model assumes that the difficulty and discrimination of all items are the same. It is not known how

sensitive the procedure is with respect to violation of these assumptions. Thus, research addressed to this question is needed. It would also be desirable to study the possibility of relaxing the model to allow for unequal item difficulties and discriminations as well as allowing for polychotomous responses.

Although self-paced instruction has many advantages, limited resources often do not permit its full implementation. As a result, the sample under instruction will likely be heterogeneous with respect to achievement. Similarly, if a test is intended to measure retention of achievement or levels of achievement acquired prior to instruction, there will be wide variation in levels of performance. Under these conditions, adapting the test to an individual's level of achievement will be more efficient than the conventional non-adaptive procedure.

Most of the research on adaptive testing has been done in the context of dichotomous response models. The exceptions are to be found in the work of Bayroff, Thomas, and Anderson (1960), Wood (1971), and Samejima (1976). One of the major aims of the achievement/performance testing research at the University of Minnesota is to combine the advantages of partial knowledge scoring and adaptive testing. Bayroff et al. (1960) seem to be the only researchers who have actually implemented an adaptive testing strategy using non-dichotomous items. Essentially what they did was to branch an individual according to the correctness of the alternative chosen. Although they used a polychotomous item for the first item only, this can be readily extended to include all items. Other branching rules are possible. Wood (1971) suggested that the optimal branching rule will administer as the next item the most discriminating of those items with a midpoint of adjacent categories closest to the individual's current estimated achievement. Samejima (1976) implemented a simulation on live data of a similar procedure, which she referred to as tailoring the dichotomization of the item to the individual. She noted substantial improvements by comparing the plot of scores based on a uniform dichotomization and tailored dichotomization against the scores based on the polychotomous responses.

Summary and Conclusions

Two recent developments in test theory hold promise for the improvement of achievement test scores. In combination, adapting the test to the individual and simultaneously extracting more information from each response by recording partial knowledge should result in greater improvements in achievement test scores than either taken alone. The use of non-dichotomous item formats, now made possible by the administration of achievement test items on interactive computers, should result in achievement tests which more accurately measure what a student has learned as a result of instruction.

Although the use of polychotomous models in the measurement of partial knowledge has been emphasized here, it is clear that these models have much to offer in performance testing as well. Fitzpatrick and Morrison (1970) define a performance test as "one in which some criterion situation is simulated to a much greater degree than represented by the usual paper-and-

pencil test." Unlike paper-and-pencil tests, performance tests are relatively expensive and it is this cost consideration that highlights the necessity for extracting as much information as possible from a testee's set of responses. Polychotomous response models make this feasible. The use of interactive computers also has much to offer in the area of performance testing, for computerized test administration can make it possible to represent simulated situations conveniently and economically. Additional savings are likely by testing individuals only on those skills which match the individual's level of training.

In short, it seems that coupling polychotomous response model theory with interactive computer administration of tests is likely to result in more accurate and, in the long run, more economical assessments of achievement and performance.

REFERENCES

Angoff, W. H. The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias, December 1975.

Bayroff, A. G., Ross R., & Fischl, M. A. Development of a programed testing system. (Technical Paper 259). Arlington, VA: Army Research Institute for the Behavioral and Social Sciences, December 1974.

Bayroff, A. G., Thomas, J. J., & Anderson, A. A. Construction of an experimental sequential item test. (Research Memorandum 60-1). Personnel Research Branch, Department of the Army, January 1960.

Bejar, I. I., & Weiss, D. J. A comparison of empirical differential option weighting procedures. Educational and Psychological Measurement. (In press.)

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance. (Research Rep. 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (AD A027147) (a)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing. (Research Rep. 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (AD A027170) (b)

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, MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37(1), 29-51.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40(1), 5-32.

Coombs, C. H., Millholland, J. E., & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 17-37.

Cronbach, L. J. Essentials of psychological testing. New York: Harper and Row, 1961.

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

Ferguson, R. L. A model for computer-assisted criterion-referenced measurement. Education, 1970, 91, 25-31.

Fitzpatrick, R. S., & Morrison, E. J. Performance and production evaluation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington DC: American Council on Education 1971.

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

Green, D. R. Reducing bias in achievement tests. Paper presented at the NIE Conference on Test Bias, San Francisco, April 1976.

Hakstian, A. R., & Kansup, W. A comparison of several methods of assessing partial knowledge in multiple choice tests: II. Testing procedures. Journal of Educational Measurement, 1975, 12, 231-240.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(3), 159-170.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart, & Winston, 1973.

Jensen, A. R. Test bias and construct validity. Revised address to the American Psychological Association Annual Meeting, Chicago, December 1975.

Jöreskog, K. G. Statistical analysis of sets of cogeneric tests. Psychometrika, 1971, 36, 109-133.

Lord, F. M. Applications of item response theory to practical testing problems. Paper presented at the convention of the American Psychological Association, Washington, DC, September 1976.

Lord, F. M. A theory of test scores. Psychometric Monograph. 1952, No. 7.

Lord, F. M. Individualized testing and item characteristic curve theory. In Krantz, Atkinson, Luce, & Suppes (Eds.), Contemporary developments in mathematical psychology. San Francisco: W. H. Freeman, 1974.

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

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

Lumsden, J. Test theory. In M. R. Rosenzweig, & L. W. Porter (Eds.), Annual review of psychology. Palo Alto, CA: Annual Reviews, Inc., 1976.

Maxwell, A. E. Estimating true scores and their reliabilities in the case of composite psychological tests. British Journal of Mathematical and Statistical Psychology, 1971, 24, 195-204.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy. (Research Rep. 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (AD A022964)

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

Pine, S. M., & Weiss, D. J. A comparison of the fairness of adaptive and conventional testing strategies. (Research Rep. 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. (In press.)

Pine, S. M.,& Weiss, D. J. Effects of item characteristics on test fairness. (Research Rep. 76-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1976. (AD A035393)

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Neilson & Lydiche (for Danmarks Paedagogiske Institut), 1960.

Rummel, R. J. Applied factor analysis. Evanston, IL: Northwestern University Press, 1970.

Samejima, F. Estimating latent ability using a response pattern of graded responses. Psychometrika, 1969, Monograph Supplement No. 17.

Samejima, F. Graded response model of the latent trait theory and tailored testing. Proceedings of the First Conference on Computerized Adaptive Testing. United States Civil Service Commission, Bureau of Policies and Standards, 1976.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39(1), 111-121.

Sattler, J. J. Assessment of children's intelligence. Philadelphia: W. B. Saunders Company, 1974.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

Urry, V. W. Ancillary estimators for item parameters of mental test models. Unpublished paper. Personnel Research and Development Center, U. S. Civil Service Commission, 1974. (a)

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269. (b)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing. (Research Rep. 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (AD A020961)

Wang, M. W., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

Waters, B. J. Development of a microcomputer-controlled, interactive testing terminal. Lowry AFB, CO: Air Force Human Resources Laboratory, in preparation, 1977.

Weiss, D. J. Adaptive testing research at Minnesota--Overview, recent results and future directions. (Professional Series 75-6). Washington, DC: Proceedings of the First Conference on Computerized Adaptive Testing, June 1975.

Weiss, D. J. Computerized ability testing, 1972-1975. (Final Report). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1976. (AD A024516)

Weiss, D. J. Strategies of adaptive ability measurement. (Research Rep. 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (AD A004270)

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? Minneapolis, University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (AD 757788)

Wingersky, M. S., & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. (Research Memorandum 73-2). Princeton, NJ: Educational Testing Service, 1973.

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.