SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

AD-A214 717

1. REPORT NUMBER	2. GOVT ACCESSION NO.
AFOSR-TR-89-1647	

4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED
APPLICATIONS OF LATENT TRAIT THEORY TO THE DEVELOPMENT AND USE OF CRITERION-REFERENCED TESTS	Final Technical Report (Feb. 1, 1978-April 30, 1979)
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Ronald K. Hambleton	8. CONTRACT OR GRANT NUMBER(s) F49620-78-C-0039
9. PERFORMING ORGANIZATION NAME AND ADDRESS Laboratory of Psychometric and Evaluative Research School of Education/University of Massachusetts Amherst, MA 01003	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Department of the Air Force Air Force Office of Scientific Research Bolling Air Force Base, DC 20332	12. REPORT DATE March 1979
	13. NUMBER OF PAGES 41 pages
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)

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited

Approved for public release; distribution unlimited.

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

DTIC ELECTE NOV 29 1989 S B D

18. SUPPLEMENTARY NOTES

A paper presented at an AERA-NCME symposium entitled "Psychometric Approaches to Domain-Referenced Testing," San Francisco, April 1979.

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

89 11 27 158

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

The success of criterion-referenced testing programs depends to a considerable extent upon how effectively tests are constructed, and test scores used to assign examinees to mastery states and/or to estimate examinee domain scores. Methodologies such as decision theory, Bayesian statistics, and generalizability theory have been used successfully to address a variety of technical matters (for example, reliability estimation and domain score estimation). The purposes of this paper were to consider latent trait theory

DD FORM 1 JAN 73 1473

Unclassified

20. Abstract (continued)

as a framework for resolving some of the technical aspects associated with criterion-referenced tests. Specifically, advantages and disadvantages of latent trait theoretic concepts were considered; a discussion of past applications to test development and test score usage were provided; and directions for future research and development were offered.

Features that make the use of latent trait theory attractive include (1) the independence of examinee ability estimates from the particular choice of test items from a "pool" of calibrated test items defining some content domain, (2) the independence of item statistics from the particular examinee sample used to calibrate them, and (3) the availability of a measure of precision for each ability estimate.

Keywords: Latent trait theory, Cognition, Mathematical models, Psychological tests, Aptitude tests. (SDW)

Applications of Latent Trait Theory to the Development and Use of Criterion-Referenced Tests[1,2,3]

Ronald K. Hambleton
University of Massachusetts, Amherst

The success of competency based education will depend to a considerable extent upon how effectively criterion-referenced tests are constructed, and how the test scores are used (1) to assess examinee performance levels and (2) to make mastery/non-mastery decisions. It is common to define a criterion-referenced test as a test which is designed to provide examinee data relative to well-defined objectives being measured by a test (Popham, 1978). "Well-defined" means that each objective is stated in such a way that the relevant pool of possible test items measuring an objective is clear to anyone who makes use of the test scores or who becomes involved in the test development process (for example, item writers and item reviewers).

Up until about five years ago there was a considerable amount of energy being expended in the development of criterion-referenced tests and in the use of criterion-referenced test scores. However, the potential of these criterion-referenced testing programs was not often realized

---

[1]The project was performed pursuant to a contract from the United States Air Force Office of Scientific Research. However, the opinions expressed here do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.

[2]Laboratory of Psychometric and Evaluative Research Report No. 91. Amherst, MA: School of Education, University of Massachusetts, 1979.

[3]A paper presented at an AERA-NCME symposium entitled "Psychometric Approaches to Domain-Referenced Testing," San Francisco, April 1979.

either because of poorly constructed tests or misinterpreted test scores or both. Undoubtedly such a state of affairs existed because of the shortage of technical guidelines to aid both test developers and test score users. Often the test items did not measure the intended objectives, too few test items were used in the tests, performance standards were set without due consideration of the relevant issues and/or using proper methods, and so on.

Fortunately, there is no reason for the problems to exist anymore. There have been a large number of very useful contributions to a criterion-referenced testing technology and you have heard about many of these from the other presenters at this symposium (Brennan, Huynh, Subkoviak). Such contributions are making it possible to develop better criterion-referenced tests and to use the scores in more appropriate ways (Popham, 1978; Hambleton, Swaminathan, Algina, & Coulson, 1978). For example, much is known about steps for developing criterion-referenced tests, assessing content validity, assembling tests, setting performance standards and assessing test reliability.

Before I lull the reader into a state of euphoria, let me be quick to point out that many very important problems remain. For one, what are the best methods for obtaining more accurate estimates of examinees' domain scores (level of performance scores relative to each objective being tested) and for decreasing the frequency of times examinees are misclassified (assigned to "non-mastery" states when they are "masters" and assigned to "mastery" states when they are "non-masters")?

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

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

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

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

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

There is a second solution to the problem sketched out earlier (and other testing problems). This solution involves the use of latent trait theory (Lord & Novick, 1968; Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1979). Considerable research has been done with latent trait models and concepts and many applications to testing have been highly successful but relatively little <u>specific</u> attention to criterion-referenced testing problems has been given. Specific attention is important because norm-referenced tests and criterion-referenced tests are constructed, analyzed, and test scores interpreted in fundamentally different ways (norm-referenced tests are constructed to facilitate comparing one person with another on the ability measured by a test; criterion-referenced tests are constructed to determine examinee level of performance relative to the objectives measured by the test) and therefore latent trait theoretic results which apply to norm-referenced tests will not necessarily apply to criterion-referenced tests. Unfortunately, much of the research and development work has been done with respect to norm-referenced tests (see, for example, work by Hambleton et al., 1979; Lord, in press; Weiss, 1978).

Purposes of the Study

Two important technologies have emerged in the last ten years which have considerable potential for improving the assessment of individuals. The first, criterion-referenced testing technology, is the better known of the two, and is being used throughout the country in a variety of ways (for example screening of students, monitoring student progress in courses, assigning student grades, and licensing and certification). Nevertheless, many technological problems remain and therefore these new criterion-referenced testing programs are not achieving their full potential. The second, latent trait theoretic technology, has developed more slowly, but is now being used in many types of testing programs. A cursory glance at the 1979 AERA and NCME annual meeting program will quickly substantiate the extensive use of latent trait models. There is one notable exception. It is not being used in any extensive way with criterion-referenced tests. This is unfortunate because latent trait models and concepts have lead to many important norm-referenced test developments (see, for example, Hambleton & Cook, 1977), and appear to have the capability of resolving some of the technological problems associated with the construction and uses of criterion-referenced tests.

The goal of this paper is to consider latent trait theory as a framework for resolving some of the technical problems associated with criterion-referenced tests. Specifically, (1) a brief introduction to latent trait models and concepts is offered, (2) features of latent trait models which have special relevance to criterion-referenced testing are considered, (3) several applications of latent trait models are introduced, and (4) conclusions and suggestions for further research are provided. The four sections of the paper correspond to the four specific purposes outlined above.

Brief Introduction to Latent Trait Models and Concepts

A theory of latent traits supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by examinee characteristics, referred to as traits. Scores for examinees on these traits are estimated and used to predict or explain test performance (Lord and Novick, 1968). Since the traits are not directly measurable, they are referred to as latent traits or abilities. A latent trait model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. The relationship between the "observable" and the "unobservable" quantities is described by a mathematical function. The concept of a "latent trait," and a "domain score" in the context of criterion-referenced measurement are the same. The relationship is an algebraic one and is specified by the "test characteristic curve," a term which will be defined later.

When selecting a particular latent trait model to apply to one's test data, it is necessary to consider whether the data satisfy the assumptions of the model. If they do not, different test models should be considered. Alternately, some psychometricians (for example, Wright, 1968) have recommended that test developers design their tests so as to satisfy the assumptions of the particular latent trait model they are interested in using. In this way, the advantages of the particular latent trait model of interest can be utilized.

The three fundamental assumptions underlying the most commonly used

latent trait models are: The unidimensionality of the test items, local independence, and the mathematical form of the item characteristic curves. Each of these assumptions will be discussed briefly. Two other important topics will also be considered: Item and test information curves, and efficiency.

The Assumption of Unidimensionality

The assumption of a unidimensional set of test items is a common one for test constructors, since they usually desire to construct unidimensional tests so as to enhance the interpretability of a set of test scores (Lumsden, 1976). This is certainly the case with criterion-referenced tests since a key characteristic of a good criterion-referenced test is the interpretability of scores derived from the test.

Lumsden (1961) provided an excellent review of methods for constructing unidimensional tests. He concluded that the method of factor analysis held the most promise. Fifteen years later he reaffirmed his conviction (Lumsden, 1976). Essentially, Lumsden recommends that a test constructor generate an initial pool of test items selected on the basis of empirical evidence and a priori grounds. In the jargon of criterion-referenced measurement, items are written to match domain specifications and are discarded when it can be determined that they are invalid for their intended purposes. Such an item selection procedure will increase the likelihood that a unidimensional set of test items within the pool of items can be found. If test items are not preselected, the pool may be too heterogeneous for the unidimensional set of items in the item pool to emerge. In Lumsden's method, a factor analysis is performed and items that are not measuring the dominant factor obtained in the factor solution are removed. The

remaining items are factor analyzed, and again, "deviant" items are removed. The process is repeated until a satisfactory solution is obtained. Convergence is most likely when the initial item pool is carefully selected to include only items that appear to be measuring a common trait. Lumsden proposed that the ratio of first factor variance to second factor variance be used as an "index of unidimensionality." Rejected test items should be studied to determine the possible basis for their misfit. In some instances, it may be necessary to rewrite the domain specifications to reflect the test items which remain.

Local Independence

The second assumption is that of local independence. The assumption states that the test item responses of a given examinee are statistically independent. This means that an examinee's performance on one item does not affect his or her performance on other items in the test. This result would be obtained if the test items measure a single ability.

Item Characteristic Curves

An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the set of items contained in the test. There is no concept comparable to the notion of an item characteristic curve in standard test technology. A primary distinction among different latent trait models is in the mathematical form of the corresponding item characteristic curves. It is up to the user to choose one of the many mathematical forms for the shape of the item characteristic curves. In doing so, an assumption about the items is being made which can be verified later by how well the chosen model "explains" obtained test results.

Each item characteristic curve for a particular latent trait model is a member of a family of curves of the same general form. The number of parameters required to describe an item characteristic curve will depend on the particular latent trait model.

The mathematical expression for the three-parameter logistic curve is:

$$P_g(\theta) = c_g + (1-c_g)\frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}}, \quad g=1, 2, \ldots, n,$$

where:

$P_g(\theta)$ = the probability that an examinee with ability level θ answers item g correctly,

b_g = the item difficulty parameter,

a_g = the item discrimination parameter,

and

D = 1.7 (a scaling factor).

The parameter c_g is the lower asymptote of the item characteristic curve and represents the probability of examinees with low ability correctly answering an item. The parameter c_g is included in the model to account for test response data at the low end of the ability continuum, where among other things, guessing is a factor in test performance. It is now common to refer to the parameter c_g as the pseudo-chance level parameter in the model.

Typically, c_g, assumes values that are smaller than the value that would result if examinees of low ability were to guess randomly to the item. As Lord (1974) has noted, this phenomenon can probably be attributed to the ingenuity of item writers in developing "attractive" but incorrect choices. For this reason, avoidance of the label "guessing parameter" to describe the parameter c_g is desirable.

The popular "Rasch model" or one-parameter logistic test model can be obtained from the three-parameter logistic model by making two assumptions about the test data: (1) the amount of guessing is minimal, and (2) items included in a test are equally discriminating.

Item characteristic curves for several latent trait models are presented in Figure 1.

Item and Test Information Curves

Once a latent-trait model is specified, the precision with which it estimates examinee ability can be determined. Birnbaum (1968) defined the notion of information as a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability. The standard error of estimate of ability is equal to $1/\sqrt{\text{information}}$. When information at an ability level is high, we have narrow confidence bands around our estimates. If information is low, we have wider confidence bands. Because the information function varies with ability level, it has been suggested that test information curves ought to replace the use of classical reliability estimates and standard errors of measurement in test score interpretations.

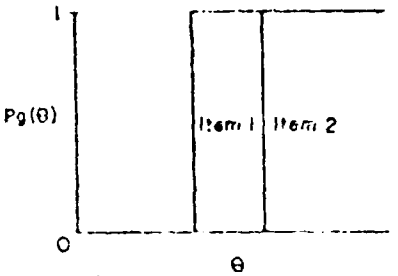
In mathematical terms, Birnbaum (1968) gives the information curve of a given scoring formula by

$$I_y(\theta) = \frac{\left(\sum_{i=1}^{n} w_i P_i'\right)^2}{\sum_{i=1}^{n} w_i^2 P_i Q_i}$$

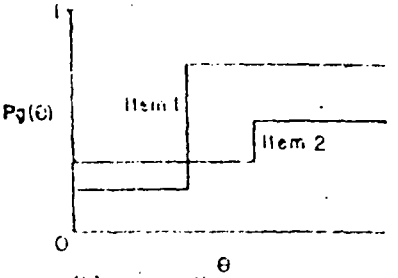
In the expression above, $I_y(\theta)$ is the amount of information at ability level θ provided by the scoring formula y, where

$$y = \sum_{i=1}^{n} w_i X_i.$$

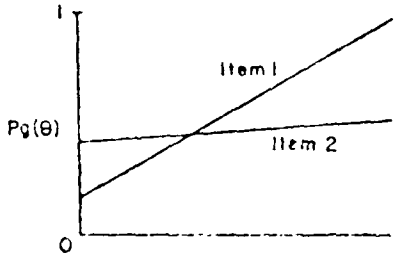
Figure 1. Seven examples of item characteristic curves.



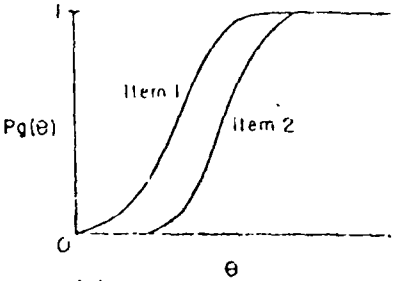
(a) perfect scale curves



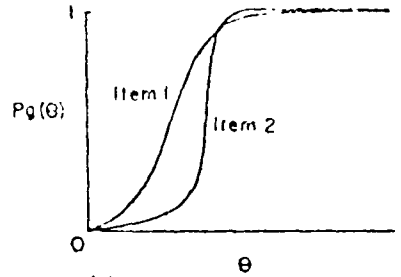
(b) latent distance curves



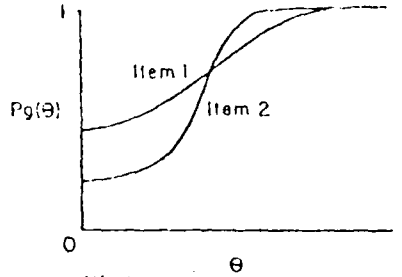
(c) latent linear curves



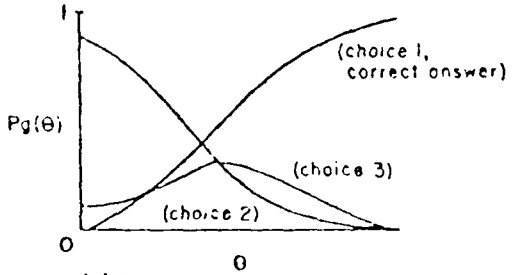
(d) one-parameter logistic curves



(e) two-parameter logistic curves



(f) three-parameter logistic curves



(g) item response curves (single item, 3 choices)

the variable X_g is 0 or 1 depending on whether or not item g is answered correctly; P_g is the probability of a correct answer to item g by an examinee with ability level θ; Q_g is equal to $1 - P_g$; P'_g is the slope of the item characteristic curve at ability level θ; and the item scoring weights are w_g, g = 1, 2, ..., n.

Birnbaum (1968) has shown that the maximum value of $I_y(\theta)$, referred to as the test information curve, is given by

$$I(\theta) = \sum_{g=1}^{n} \left(\frac{P'^2_g}{P_g Q_g}\right) . \qquad [1]$$

The maximum value of the information curve of a given scoring formula is obtained when the scoring weights are chosen, such that

$$w_g = \frac{P'_g}{P_g Q_g} .$$

The quantity $P'^2_g/P_g Q_g$ is the contribution of item g to the information function of the test and is referred to as the item information function. Item information functions have an important role in determining the accuracy with which ability is estimated at different levels of θ. Each item information curve depends on the slope of the particular item characteristic curve and the conditional variance of test scores at each ability level θ. The higher the slope of the item characteristic curve and the smaller the conditional variance, the higher will be the item information curve at that particular ability level. The height of the item information curve at a particular ability level is a direct measure of the usefulness of the item for precisely measuring ability at that level.

The information function for the test composed of the items is obtained by summing the ordinates of the item information curves. From Equation [1] it is clear that items contribute independently to the test information function. Birnbaum (1968) has also shown that with his three-parameter model, an item provides maximum information at an ability level θ, where

$$\theta = b_g + \frac{1}{1.7\, a_g} \log_c 1/2 \left(1 + \sqrt{1 + 8c_g}\right).$$

If guessing is minimal, then $c_g = 0$, and $\theta = b_g$.

Figures 2 to 5 show ten item characteristic curves and corresponding item information curves. The influence of the pseudo-chance level parameter is clear from the figures: When $c_g > 0$, (1) the lower asymptote of the item characteristic curve is different from zero, (2) less information is obtained, and (3) the point of maximum information is shifted to a somewhat higher ability level. Figure 8 shows the calculation of a test information curve from five item information curves. In passing, perhaps it should be noted that when item parameter estimates are used in place of item parameters, test information curves are called "score information curves" by Lord (in press).

Efficiency

A concept closely related to test information is the concept of efficiency. An efficiency curve is formed by calculating the ratio of two information curves at different points on an ability continuum. The efficiency curve provides a measure of the relative effectiveness of two

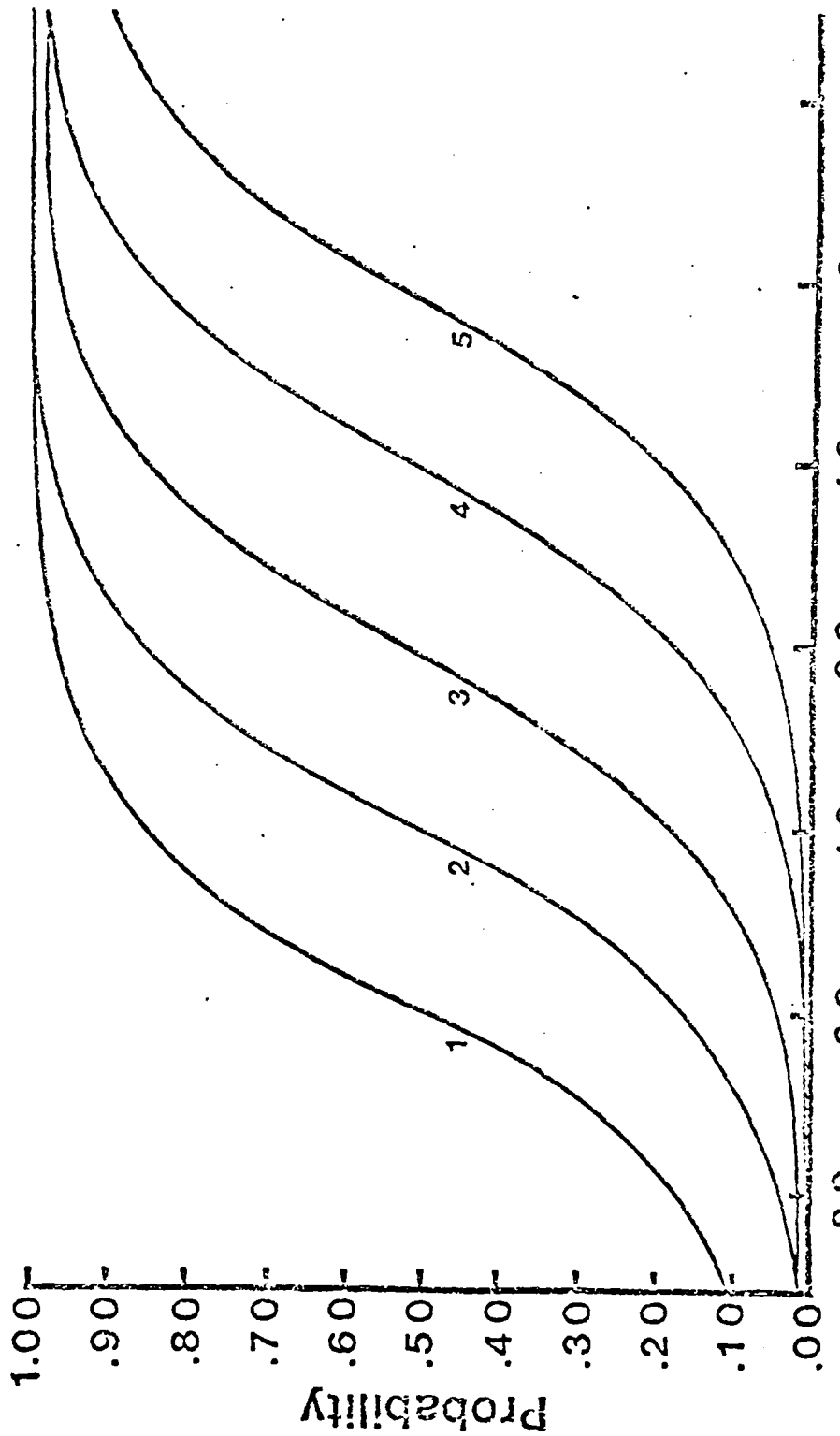


Figure 2. Graphical representation of five item characteristic curves [b=-2.0, -1.0, 0.0, 1.0, 2.0; a=.99; c=.00].

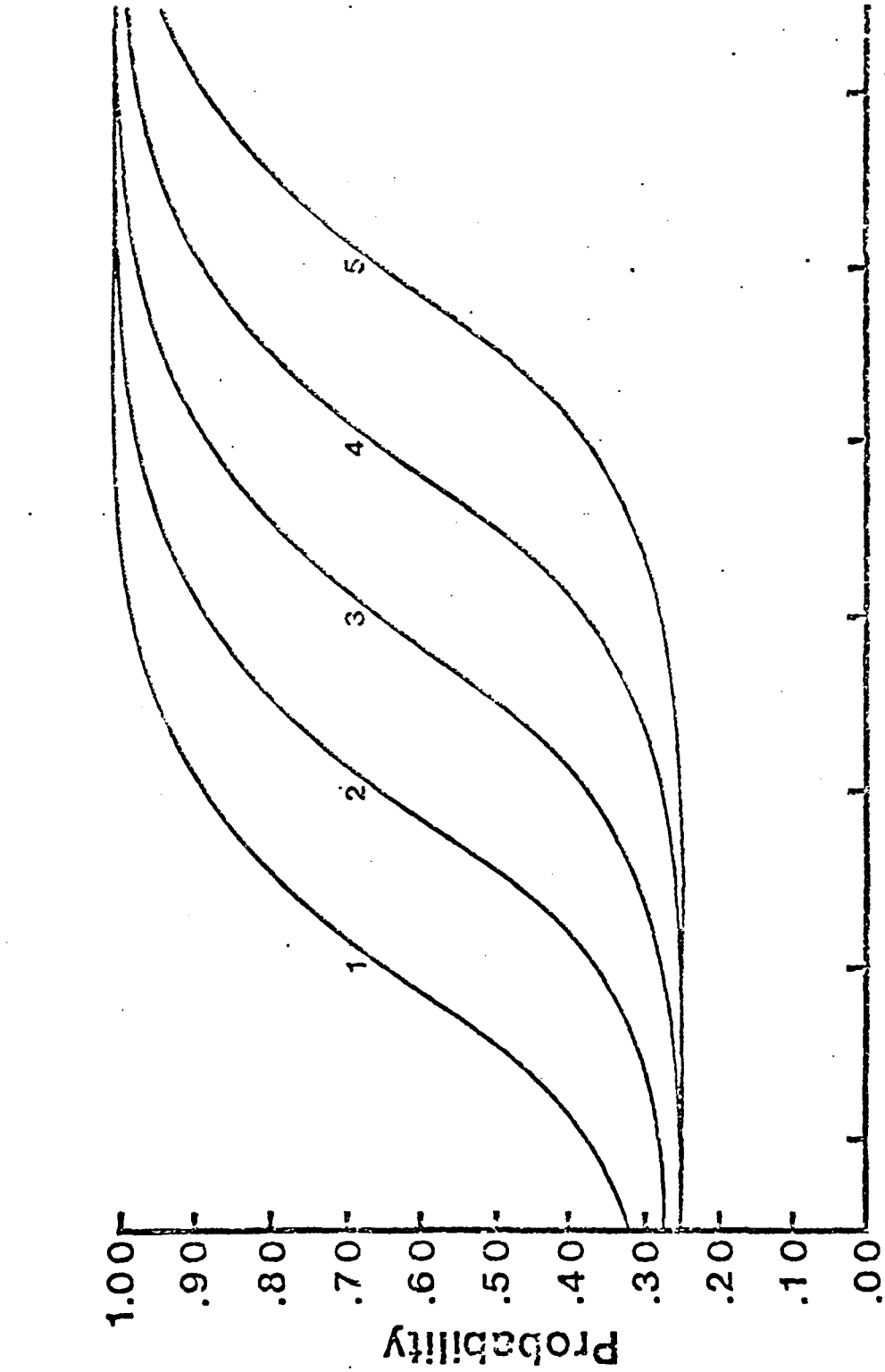


Figure3. Graphical representation of five item characteristic curves [b = -2.0, -1.0, 0.0, 1.0, 2.0; a= .99; c= .25].

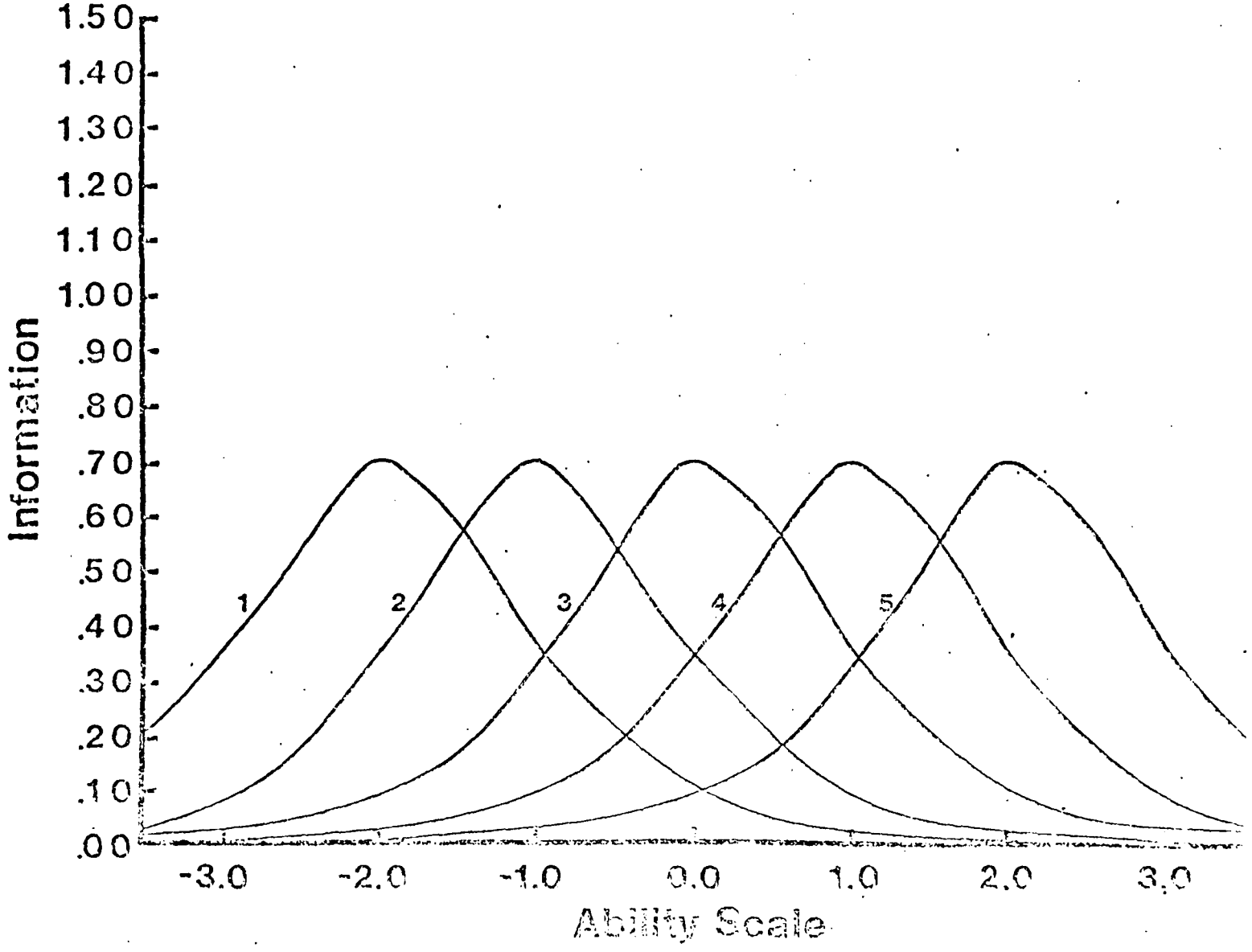


Figure 4. Graphical representation of five item information curves [b=-2.0, -1.0, 0.0, 1.0, 2.0; a= .99; c= .00].

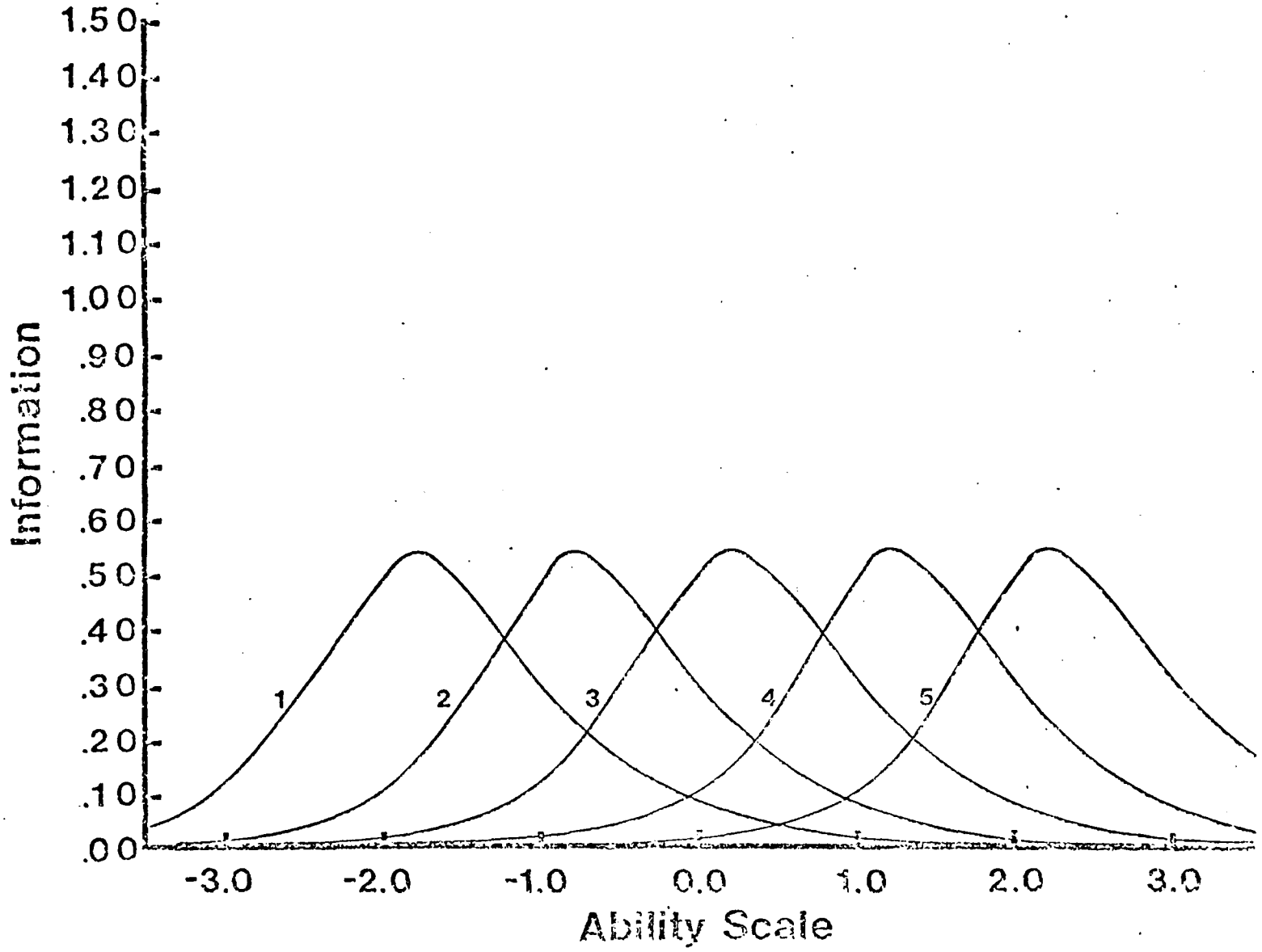


Figure 5. Graphical representation of five item information curves [b = -2.0, -1.0, 0.0, 1.0, 2.0; a = .99; c = .25].

Item	b_i	a_i	c_i
10	1.1	2.0	.05
11	−1.5	.9	.20
13	−0.1	1.6	.16
30	2.4	1.1	.09
47	−0.4	.4	.20

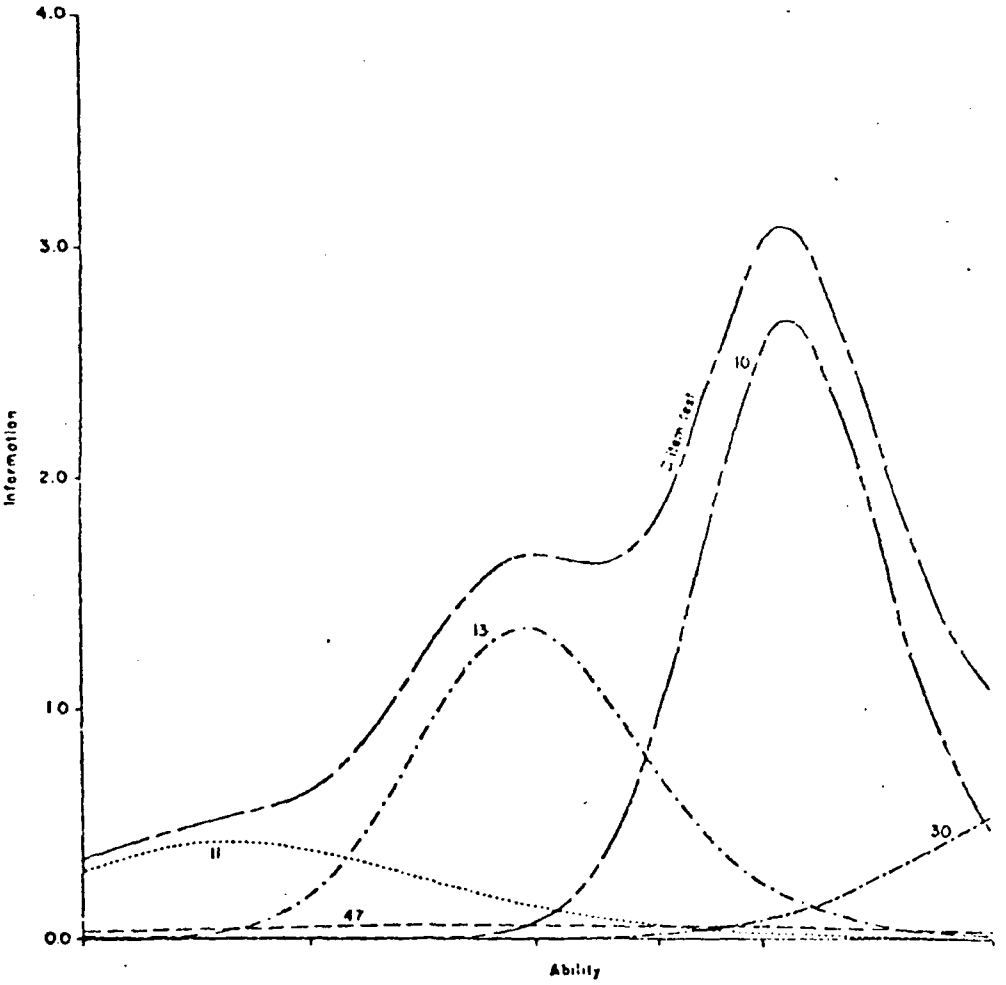


Figure 8. Information curves estimated for five items and a five-item test. The items are from the verbal section of the SAT. This figure is reproduced by permission from Lord (1968).

tests (each characterized by a different information curve) for measuring ability. In test development work it is common to compare the efficiency of different test designs (i.e., tests composed of different items) for measuring ability at different locations on the ability continuum. Whereas the shapes of test information curves depend on the metric chosen for measuring ability, efficiency curves do not, and therefore they are particularly useful in test development work.

The process of determining the relative efficiency of two tests is employed more often as part of the analysis of existing tests than as a part of the test development process.
The distinction between test analysis and test development has been made by Rentz and Bashaw (1977). Basically they define the test development process as one that allows items that do not fit the model to be discarded, whereas in test analysis applications, "the model becomes fixed and data are in effect 'fitted' to it." The distinction made by Rentz and Bashaw between test development and test analysis is a useful one.

The Ability Scale and Test Characteristic Curves

If we were to administer two criterion-referenced tests, that measured the same objective (or objectives), to the same group of examinees, and the tests were not strictly parallel two different test score distributions would result. The extent of the differences between the two distributions would depend, among other things, on the difference between the difficulties of the two tests. Unfortunately, there is no basis for preferring one distribution over the other. What this example reveals is that, in general, the test score distribution provides no information about the distribution of ability scores.

The problem occurs because the raw-score units from each test are unequal and different. On the other hand, the scale on which ability scores are measured is one on which examinees will have the same ability score across non-parallel tests measuring a common ability. Thus, even though an examinee's test scores will vary across non-parallel forms of a test measuring an ability, the expected ability for an examinee will be the same on each form.

Most measurement specialists are familiar with the concept of domain score, the expected test score (on a sample of test items) for an examinee. What is the relationship between domain scores and ability scores? The test characteristic curve, which is obtained by summing the ordinates of the ICC's, provides the relationship. This is easily seen from the following argument. Consider the proportion-correct score, $\hat{z} = \frac{x}{n}$. Then

$$E(\hat{z}|\theta) = \frac{1}{n} \sum_{g=1}^{n} P_g(\theta), \qquad [2]$$

$$Var(\hat{z}|\theta) = \frac{1}{n}2 \sum_{g=1}^{n} P_g(\theta)\, Q_g(\theta). \qquad [3]$$

$E(\hat{z}|\theta)$ is the test characteristic curve (scaled by 1/n) introduced earlier. It is the sum of item characteristic curves for items included in the test. Suppose next we lengthen the test by adding an infinite number of parallel-forms. By definition, $E(\hat{z}|\theta) = \pi$, the domain score. Also Var $(\hat{z}|\theta) \to 0$, as $n \to \infty$, and so π and θ will be related by a monotonic increasing transformation which is the test characteristic curve. Clearly then, the two concepts, π and θ, are the same, except for the scale of measurement used to describe each. One important difference is that domain score is defined on the interval [0, n] or [0, 1] whereas ability scores are usually defined on the interval $[-\infty, +\infty]$.

There are other differences between domain scores and ability scores. A domain score is defined for <u>each</u> sample of test items. It is the expected test score for an examinee. An examinee's domain score will vary across non-parallel measures of the same ability. On the other hand, ability score is defined for a "pool" or "universe" of items measuring a single ability. An examinee's domain score in different samples of items would (in general) vary. However, ability score is defined in terms of the "pool" of items from which the sample was drawn. Latent trait models specify relationships between examinee item performance and ability, and so it is always possible to "transform" examinee performance on a particular sample of items (defining a test) onto an ability scale defined for the large "pool" of test items. Thus, while an examinee would have (in general) a different domain score for each sample of items drawn from the pool and would obtain different test scores in each sample of items, the expected estimate of examinee ability from each sample of test items would be the same. More will be said about this important relationship later.

Ability scores can be used with item characteristic curve parameters for items included in a test to estimate examinee test performance. Recall,

$$E(X|\theta) = \sum_{g=1}^{n} P_g(\theta). \qquad [4]$$

Thus, ability scores provide a basis for <u>content-referenced interpretations</u> of examinee test scores. When the quantities in Equation [4] are scaled by $1/n$, $E(X/n|\theta)$ represents the expected proportion of items in a test that an examinee will answer correctly and this interpretation will have meaning regardless of the test performance of other examinees. Of course, ability scores provide a basis for <u>norm-referenced interpretations</u> as well.

Special Features of Latent Trait Models

When latent trait models fit particular data sets, three advantages are obtained. Perhaps the most important advantage of latent trait models is that, given a set of test items that have been fitted to a latent trait model (that is, item parameters are known), it is possible to estimate an examinee's ability on the same ability scale from any subset of items in the domain of items that have been fitted to the model. (Of course, the domain of items needs to be homogeneous in the sense of measuring a single ability. If the domain of items is too heterogeneous, the ability estimates will have little meaning.) In fact, regardless of the number of items administered (as long as the number is not too small) or the statistical characteristics of the items, the ability estimate for each examinee will be an asymptotically unbiased estimate of true ability, provided the latent trait model holds. Ability estimation independent of the particular choice (and number) of items represents one of the major advantages of latent trait models. Hence, latent trait models provide a way of comparing examinees even though they may have taken quite different subsets of test items. In latent trait models, the difficulty of items is accounted for by the model and reflected in the ability estimates. Thus, two students, who receive identical scores on an easy and difficult subset of the test items, respectively, will differ in their ability estimates (the second student will receive a higher estimated score than the first).

Another advantage of latent trait models is that item parameters are invariant across sub-groups of examinees chosen from an examinee population. In principle, item parameters should remain the same, regardless of the subgroup tested. Invariant item parameters have been sought by measurement specialists for a long time; their advantages are obvious for test development work. Certainly classical item statistics, such as item difficulty

will vary from group to group, depending upon the average ability of the group being tested. This invariance property is shown graphically in Figure 6.

Yet another advantage is that they provide a measure of the precision of ability estimation at each ability level. Thus, instead of providing a single standard error of measurement that applies to all examinees, regardless of their test scores, latent trait models make it possible to provide separate estimates of error for each examinee or for each ability level.

Examinee-free item statistics are especially useful in "item banking" and criterion-referenced test development. Item-free ability estimates permit the "tailoring" of tests to individuals and situations. The concepts of *information* and *efficiency* are useful in both test development work and determination of precision of ability score estimates. Some of the applications will be considered in the next section of the paper.

But also, it is now time to consider the price which must be paid for the "goods" which are "delivered" via latent trait models. First, the special features will only be obtained when there is a reasonably close match between the researcher's latent trait model and his/her data. How close? That question is currently under study by many researchers. Second, it is unlikely that the features will be obtained with "short" tests. Hard figures are difficult to come by but it would appear that tests of 15 or more items are required. Also, sample sizes of 200 or more examinees will be required to produce stable item statistics with the one-parameter model and somewhat larger samples are required with the two- and three-parameter logistic test models (Swaminathan & Gifford, 1979).

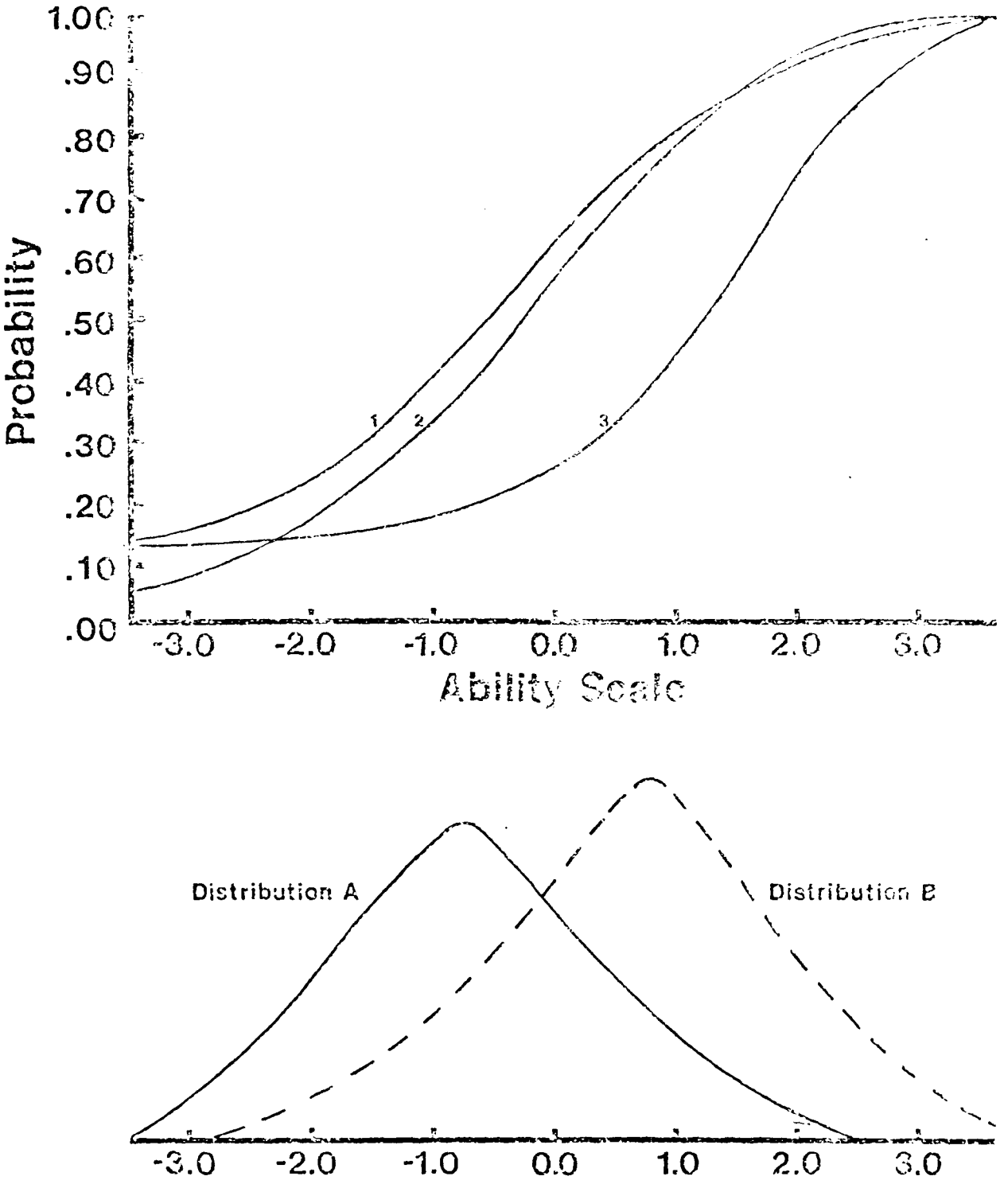


Figure 6. A diagram showing the independence of the shape of item characteristic curves from the underlying ability distribution.

Other practical problems include (1) the training of practitioners to use these models, and (2) the handling of examinees who get "rejected" because their test scores are too high or too low.

Applications

Item Banking

The development of criterion-referenced testing technology has resulted in the increasing importance of item banking (Choppin, 1976). An item bank is a collection of test items, "stored" with known item characteristics and made available to test constructors. Depending on the intended purpose of the test, items with described characteristics can be drawn from the bank and used to construct a test with known properties. Although classical item statistics (item difficulty and discrimination) have been employed for this purpose, they are of limited value for describing the items in a bank because these statistics are dependent on the particular group used in the item calibration process. Latent trait item parameters, however, do not have this limitation, and consequently are of much greater use in describing test items in an item bank (Choppin, 1976; Wright, 1977). The invariance property of the latent trait item parameters makes it possible to obtain item statistics that are comparable across dissimilar groups. Let us assume that we are interested in describing items using the two-parameter logistic test model. The single drawback is that because the mean and standard deviation of the ability scores are arbitrarily established, the ability score metric is different for each group. Since the item parameters depend on the ability scale, it is not possible to directly compare latent trait item parameters derived from

different groups of examinees until the ability scales are equated in some way. Fortunately, the problem is not too hard to resolve since Lord and Novick (1968) have shown that the item parameters in the two groups are linearly related. Thus, if a subset of calibrated items is administered to both groups, the linear relationship between the estimates of the item parameters can be obtained by forming two separate bivariate plots, one establishing the relationship between the estimates of the item discrimination parameters for the two groups, and the second, the relationship between the estimates of the item difficulty parameters. Having established the linear relationship between item parameters common to the two groups, a prediction equation can then be used to predict item parameters for those items not administered to the first group. In this way, all item parameters can be equated to a common group of examinees and corresponding ability scale.

Test Development

The important differences between developing tests using standard methods and methods based on latent trait theory occur during the following steps: (1) Item analysis, (b) selection of test items, and (c) reliability assessment.

Item analysis techniques involve (1) the characterization of test items and (2) the use of statistical information for revising and/or deleting test items. The major problem with item statistics (item difficulty and discrimination) derived from standard item analyses is that they are sample dependent. This problem is overcome by characterizing items in terms of latent trait parameters.

Latent trait theory not only provides the test developer with sample invariant item parameters but also with a far more powerful method of item selection (Birnbaum, 1968). This method involves the use of information curves, i.e., items are selected depending upon the amount of information they contribute to the total amount of information supplied by the test. One of the useful features of item information curves is that the contribution of each item to the test information function can be determined without knowledge of the other items in the test. When standard testing technology is applied the situation is very different. The contribution of any item to such statistics as test reliability cannot be determined independently of the characteristics of all the other items in the test.

Lord (1977) outlined a procedure, originally presented by Birnbaum (1968), for the use of item information curves building a test to meet any desired set of specifications. The procedure employs a pool of calibrated items, with accompanying information curves, such as might be obtained from the item banking methods previously described. The procedure outlined by Lord consists of the following steps:

1. Decide on the shape of the desired test information curve. Lord (1977) calls this the target information curve.
2. Select items with item information curves that will fill up the hard-to-fill areas under the target information curve.
3. After each item is added to the test, calculate the test information curve for the selected test items.
4. Continue selecting test items until the test information curve approximates the target information curve to a satisfactory degree.

Examples of the application of this technique to the development of tests for differing ranges of ability (based on simulated data) are given by Cook and Hambleton (1979). Some results from their study are reported in Figure 7.

An excellent discussion of item selection, as it pertains to tests developed according to Rasch model procedures, is presented by Wright and Douglas (1975) and Wright (1977). The item selection procedure basically consists of specifying the ability distribution of the group for whom the test is intended and then choosing items such that the distribution of item difficulties matches the distribution of abilities. This procedure is equivalent to that originally introduced by Birnbaum (1968), since in this case, the item information curves depend only on the difficulty parameters.

In latent trait theory test information curves replace the familiar concepts, reliability and standard error of measurement. The use of the test information curve as a measure of accuracy of estimation is appealing for at least two reasons: (1) Its shape depends only on the items included in the test, and (2) it provides an estimate of the error of measurement at each ability level.

Test Score Interpretations

One primary use of a criterion-referenced test is to obtain an estimate of an examinee's level of mastery (or "ability") on an objective. Thus, a straightforward application of one of the latent trait models (the assumption of unidimensionality would not likely be a problem) would produce examinee ability scores. Among the advantages of this application would be that items could be sampled (for example, at random) from an item pool for each examinee, and all examinee ability estimates would be on a common scale.

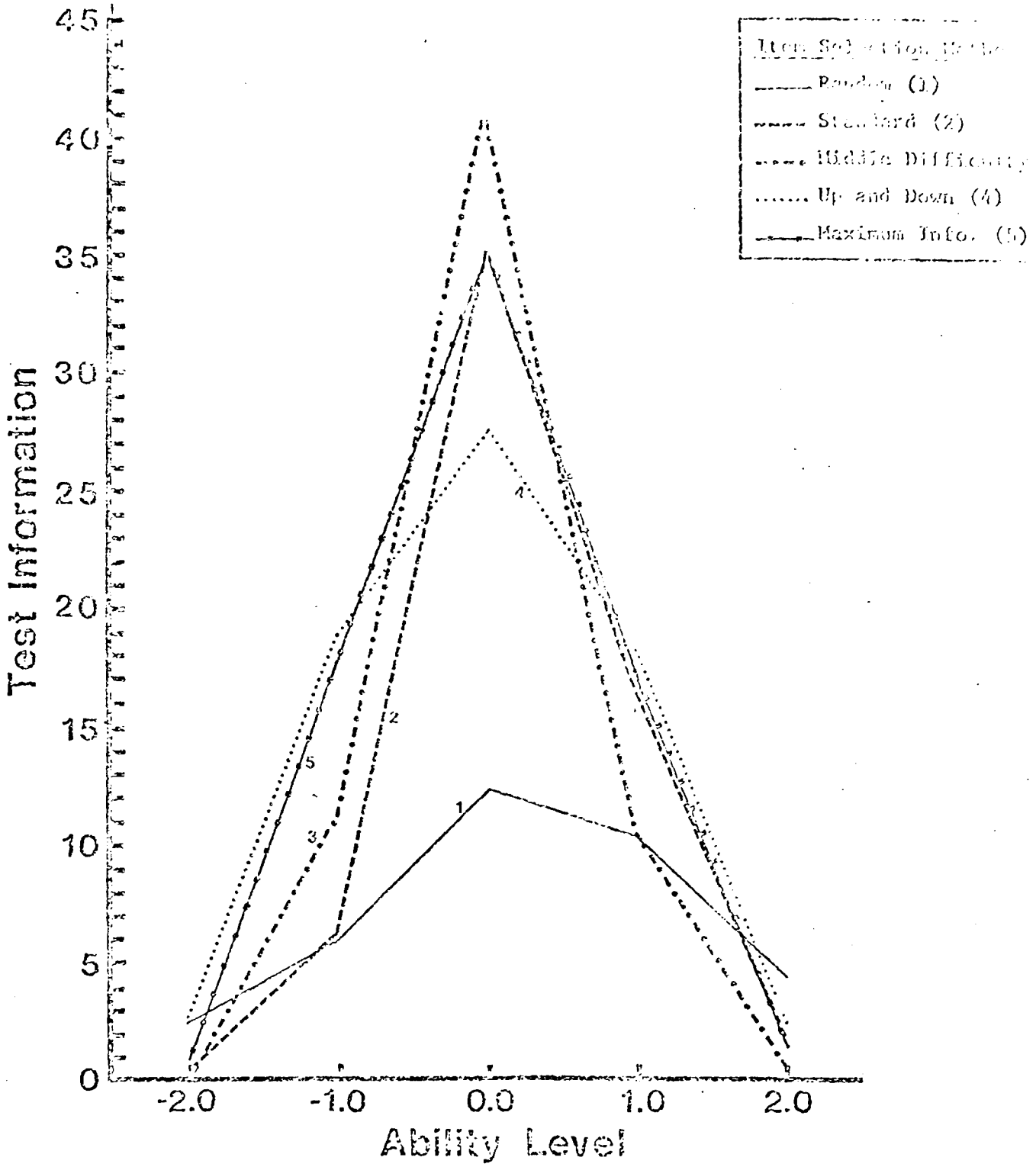


Figure 7. Test Information Curves Produced With Five Item Selection Methods [30 Test Items]

Since item parameters are invariant across groups of examinees, it would be possible to construct criterion-referenced tests to "discriminate" at different levels of the ability continuum. Thus, a test developer might select an "easier" set of test items for a pretest than a posttest, and still be able to measure "examinee growth" by estimating examinee ability at each test occasion on the same ability scale. This cannot be done with classical approaches to test development and test score interpretation. If we had a good idea of the likely range of ability scores for the examinees, test items could be selected so as to maximize the test information in the region of ability for the examinees being tested. The optimum selection of test items would contribute substantially to the precision with which ability scores were estimated. In the case of criterion-referenced tests, it is common to observe lower test performance on a pretest than on a posttest; therefore, the test constructor could select the easier test items from the domain of items measuring an objective for the pretest and more difficult items could be selected for the posttest. This would enable the test constructor to maximize the precision of measurement of each test in the region of ability where the examinees would most likely be located. Of course, if the assumption about the location of ability scores was not accurate, gains in precision of measurement would not be obtained.

The results reported in Tables 1 to 4 show clearly the advantages of "tailoring" a test to the ability level of a group. Of course, the potential improvements depend on the validity of a test developer's assumption about the examinee ability distribution. If he or she uses an incorrect prior distribution as a basis for designing a test, the resulting test will certainly not have the desired characteristics.

Table 1

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items of Equal Discrimination and Pseudo-chance Levels Equal to Zero

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	1.04	1.45	.59	1.40	.57	40%	-43%
-2.0	2.27	3.03	1.45	1.34	.64	34%	-36%
-1.0	3.85	4.66	3.03	1.21	.79	21%	-21%
0.0	4.62	4.63	4.66	1.00	1.01	0%	1%
1.0	3.78	2.98	4.63	.79	1.22	-21%	22%
2.0	2.19	1.42	2.98	.65	1.36	-35%	36%
3.0	1.00	.58	1.42	.58	1.43	-42%	43%

Table 2

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items with Varying Discrimination Indices and Pseudo-chance Levels Equal to Zero

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	1.04	1.42	.61	1.36	.58	36%	-42%
-2.0	2.21	2.91	1.42	1.32	.64	32%	-36%
-1.0	3.72	4.51	2.91	1.21	.78	21%	-22%
0.0	4.47	4.48	4.51	1.00	1.01	0%	1%
1.0	3.62	2.83	4.48	.78	1.24	-22%	24%
2.0	2.06	1.36	2.83	.66	1.37	-34%	37%
3.0	.96	.58	1.36	.61	1.42	-39%	42%

Table 3

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items of Equal Discrimination and Pseudo-chance Levels Equal to .20

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	.22	.36	.07	1.63	.31	63%	-69%
-2.0	.86	1.31	.36	1.53	.42	53%	-58%
-1.0	2.08	2.81	1.31	1.35	.63	35%	-37%
0.0	3.04	3.29	2.81	1.08	.92	8%	-8%
1.0	2.76	2.28	3.29	.82	1.19	-18%	19%
2.0	1.69	1.12	2.28	.66	1.35	-34%	35%
3.0	.79	.46	1.12	.59	1.42	-41%	42%

Table 4

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items with Varying Discrimination Indices and Pseudo-chance Levels Equal to .20

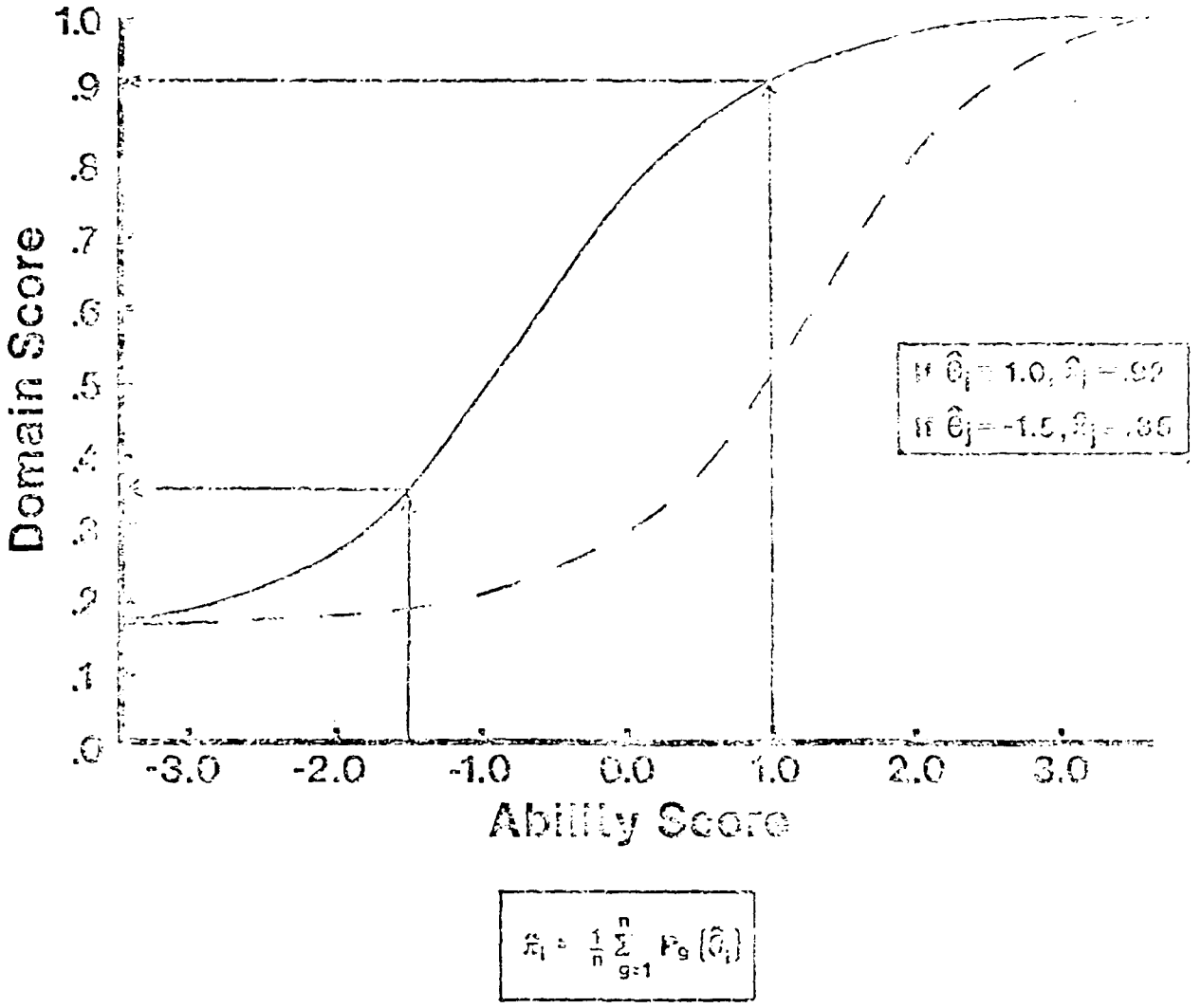
Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	.24	.37	.08	1.58	.35	58%	-65%
-2.0	.86	1.27	.37	1.48	.44	48%	-56%
-1.0	2.02	2.71	1.27	1.35	.63	35%	-37%
0.0	2.94	3.18	2.71	1.08	.92	8%	-8%
1.0	2.65	2.16	3.18	.81	1.20	-19%	20%
2.0	1.59	1.06	2.16	.67	1.36	-33%	36%
3.0	.75	.46	1.06	.61	1.41	-39%	41%

A second important use of criterion-referenced tests is to produce examinee test scores that can be used to obtain "domain score estimates." Much has already been made of the "item-free" ability estimates which are derivable from latent trait models. However, while ability estimates have the definite advantage of being "item-free," ability scores are measured on a scale which appears to be far less useful to practitioners than the domain score scale. After all, what does it mean to say, $\theta = 1.5$? Domain scores can be defined on the interval [0, 1] and provide information about examinee levels of performance (proportions of content mastered) relative to the objectives (described by domain specifications) measured on the test. As long as the test items are a representative sample of test items from the domain of items from the domain of items measuring an objective, the associated "test characteristic curve" (or more correctly, the "score characteristic curve") can be used to obtain domain score estimates from ability score estimates. When a non-representative set of test items is included in a test, examinee performance on the set of test items is used to estimate examinee ability and the score information curve for the <u>total</u> pool of calibrated test items measuring an objective is used to estimate domain scores from ability scores.

Figure 9 provides a graphical representation of the procedure outlined.

..... Test characteristic Curve for the Sample of Items Included in a Criterion-Referenced Test

——— Test characteristic Curve for the Pool of Items Measuring an Objective



$$\hat{\pi}_j = \frac{1}{n} \sum_{g=1}^{n} P_g(\hat{\theta}_j)$$

Figure 9. A comparison between the test characteristic curves for [1] the domain of item measuring an objective, and [2] the particular items included in a test which measure the objective.

I will briefly introduce one additional criterion-referenced testing problem which probably can be resolved by using latent trait models and concepts. It is common for instructors to change their tests from one group of examinees to the next. This is often done to improve the tests, to insure test security, to reflect minor adjustments in courses and so on. The problem is to insure that the standards of performance required of students across the different versions of a test are the same. The fact that a candidate must achieve a test score of (say) 90% on either test to receive a passing score does not guarantee the equivalence of the two tests. For example, it may turn out that one test is somewhat easier than the other. Required is a method for "equating" scores from one test to another. Equating of test scores will improve the usability of the derived scores for individual interpretations and course evaluations. Equating of test scores on norm-referenced tests has occupied a great deal of attention and much useful work has been done. Currently, most test score equating is being done via the use of latent trait models (the one- and three-parameter logistic test models are the most popular). In fact, there is evidence to suggest that latent trait model approaches to equating are often far superior to classical methods. However, with criterion-referenced tests we often have relatively short tests and modest numbers of examinees and therefore latent trait model equating methods need to be developed for use in this special testing situation. To date, equating studies have often been done with rather large numbers of examinees and test items.

Conclusions

The exploration of latent trait models and their application to educational testing and measurement problems has been under study for about ten years now. Certainly there are many problems requiring resolution but enough is known about latent trait models to use them successfully in solving many testing problems. With respect to the field of criterion-referenced testing, the task as I see it is one of identifying those problems which can be handled by latent trait model technology rather than whether or not the technology should be used.

On the positive side,

1. Latent trait models appear to provide an excellent basis for equating non-parallel forms of competency tests at the district and state level.

2. Several useful computer programs exist to carry out required analyses.

3. Several new textbooks and articles are now available to the interested practitioner (Hambleton, Lord, Wright & Stone, and Warm, to name four).

4. Other promising applications of latent trait models are in the areas of adaptive testing, item bias, test development, and test score interpretations. For example, Weiss and his colleagues at the University of Minnesota have some impressive results on the effects of adaptive testing in the area of criterion-referenced testing. Bob Rentz and his colleagues at Georgia State University are doing some excellent work on the study of test score reporting systems.

On the other hand,

1. I see little reason to recommend the use of latent trait models in daily classroom management of students. Latent trait models will offer little more than a headache to classroom teachers. Because (1) criterion-referenced tests are typically short, (2) sample sizes are small (although item banks may reduce the importance of this factor), (3) the required time for training of teachers in a new system of measurement would be extensive, and (4) any gains in measurement precision that might accrue would be marginal, I cannot recommend applications in this particular area.

2. No data set will ever be fit perfectly by a model. What is not known is how much misfit can be tolerated by a model and still have any advantages of the model hold in practice. Latent trait models are strong, i.e., based on restrictive assumptions, and therefore this general area requires considerably more research.

The viability of latent trait models for test development work is clear but more effective implementation could be achieved if several questions were satisfactorily answered:

1. The choice of a model is one question. At the test development stage, the practitioner has the option of developing items to fit a specific latent trait model. It would greatly facilitate the test development process, if practical guidelines existed that provided a logical basis for making this choice.

2. A second question concerns the reason for item misfit. At the present level of technical sophistication, the test developer, faced with a misfitting item, can do little more than subjectively examine the item and hope that the reason for misfit will be apparent.

3. The problem of determining whether or not a pool of items can be considered unidimensional in an important one. Factor analytical techniques are often used for this purpose but there are problems (Hambleton et al., 1979; Lord & Novick, 1968).

4. One area of current interest involves the equating of a criterion-referenced test to a norm-referenced test so that CRT scores can be reported in terms of a norm-referenced framework without actually carrying out a national norming study. Such an equating study is often discussed within the context of Title I evaluations. Legal issues aside, how best to do the equating is not clear (for example, how large and representative a sample of examinees is needed?) nor is the minimum size of the correlation between the two sets of scores which is needed to insure a stable equating known.

Numerous test developers are now considering the use of latent trait models in their work. Hopefully this paper will provide some newcomers to the area with a suitable introduction to the topic.

References

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.

Choppin, B. H. Recent developments in item banking: A review. In D. DeGruijter & L. J. Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement. New York: Wiley, 1976.

Cook, L. L., & Hambleton, R. K. A comparative study of item selection methods utilizing latent trait theoretic models and concepts. Laboratory of Psychometric and Evaluative Research No. 88. Amherst, MA: School of Education, University of Massachusetts, 1979.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

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

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

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

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. E., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1979, in press.

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

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, in press.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

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

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

Rentz, R. R., & Bashaw, W. L. The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Swaminathan, H., & Gifford, J. A. Estimation of parameters in the three-parameter latent trait model. Laboratory of Psychometric and Evaluative Research Report No. 93. Amherst, MA: School of Education, University of Massachusetts, 1979.

Weiss, D. J. (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota, 1978.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968.

Wright B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.