DOCUMENT RESUME

ED 137 364 TM 006 165

AUTHOR	Hambleton, Ronald K.; And Others
TITLE	Developments in Latent Trait Theory: A Review of Models, Technical Issues, and Applications.
PUB DATE	[Apr 77]
NOTE	126p.; Paper presented at a joint meeting of the National Council on Measurement in Education and the American Educational Research Association (New York, New York, April 1977)
EDRS PRICE	MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS	Ability; Bayesian Statistics; *Cognitive Measurement; Computer Programs; Criterion Referenced Tests; Goodness of Fit; *Mathematical Models; *Measurement; Probability; Response Style (Tests); Test Bias; Test Construction; *Testing; Testing Problems; Test Interpretation; Test Items
IDENTIFIERS	Item Characteristic Curve Theory; *Latent Trait Theory; Maximum Likelihood Estimation; Tailored Testing

ABSTRACT

Latent trait theory supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as traits, estimating scores for examinees on these traits and using the scores to predict or explain test performance (Lord and Novick, 1968). In view of the breakthroughs in several testing problem areas brought about by the use of latent trait theory, it is clear that the field of latent trait theory will become increasingly more important to measurement specialists and test practitioners. This paper comprehensively reviews this field and addresses four matters. First, the nature and characteristics of latent trait theory are introduced. Second, a review of many of the technical developments in the field is provided. Third, several promising applications of latent trait models are described. Finally, some additional areas for research and development are suggested. (RC)

3/29/77

Developments in Latent Trait Theory: A Review of Models, Technical Issues, and Applications[1,2]

Ronald K. Hambleton, Hariharan Swaminathan, Linda L. Cook
Daniel R. Eignor, Janice A. Gifford
University of Massachusetts, Amherst

There are many well-documented shortcomings of standard testing and measurement technology.[3] For one, the values of standard item parameters (item difficulty and item discrimination) are not invariant across groups of examinees that differ in ability. This means that standard item statistics are only useful in test construction for examinee populations very similar to the sample of examinees in which the item statistics were obtained. There are many testing situations where invariant item parameters would be highly desirable. Another shortcoming of standard testing technology is that comparisons of examinees on an ability measured by a set of test items comprising a test are limited to situations where examinees are administered the same (or parallel) test items. While most common standardized achievement and aptitude tests are typically suitable for middle-ability students, these tests do not provide very precise estimates of ability for either high- or low-ability examinees. "Tailored testing" is designed to correct this shortcoming by administering test items to examinees that are carefully selected to "match" their ability levels (Lord, 1970b, 1974b; Weiss, 1976; Wood, 1973). In "tailored testing," it is likely that no two examinees will take the same

[1] A paper presented at a joint meeting of NCME and AERA in New York, April 1977.

[2] Laboratory of Psychometric and Evaluative Research Report No. 47. Amherst, Massachusetts: School of Education, University of Massachusetts, 1977.

[3] "Standard testing and measurement technology" refers to commonly used methods and techniques for test design and analysis.

set of test items (or even the same number of test items). Since some examinees will be administered more difficult sets of test items than other examinees, the usual examinee test scores (or proportion-correct scores) do not provide an adequate basis for ranking examinees on the ability measured by the test items in the "domain of test items" from which test items were drawn. How then can examinees be compared? Certainly standard test models (Lord and Novick, 1968) cannot handle the problem.

Another shortcoming of standard testing technology is that it provides no basis for determining what a particular examinee might do when confronted with a test item. Such information is necessary, for example, if a test designer desires to predict test score characteristics in one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees.

Besides the three shortcomings of standard testing technology mentioned above, standard testing technology has failed to provide satisfactory solutions to many testing problems (for example, test design, test score equating, and item bias). For these and other reasons, many psychometricians have been investigating and developing more appropriate theories of mental measurements. Consequently, considerable attention is being currently directed toward the field of latent trait theory, sometimes referred to as item response theory or item characteristic curve theory. Latent trait theory can be traced back to the work of Lawley (1943, 1944). Lazarsfeld (1950) was perhaps the first to introduce the term "latent traits." The work of Lord (1952, 1953a, 1953b), however, is generally regarded as the "birth" of latent trait theory (or modern test theory as it is sometimes called). Progress in the 1950's and 60's was painstakingly slow, in part due to the mathematical complexity of the field, the lack of convenient and efficient computer programs to analyze the data according to latent trait theory, and the general skepticism about the gains that might accrue from this particular line of research. However, important breakthroughs recently in problem areas such as test score equating (Lord, 1975a; Rentz and Bashaw, 1975), tailored testing (Lord, 1974b; Weiss, 1976), test design and test evaluation (Wright, 1968) through applications of latent trait theory, have attracted considerable interest from measurement specialists. Other factors that have contributed to the current interest in latent trait theory include the availability of a number of useful computer programs, publication of a variety of successful applications in measurement journals (Bock, 1972; Lord, 1968, 1974b, 1975d; Samejima, 1969, 1972; Whitely & Dawis, 1974; Wright & Panchapakesan, 1969), and the strong endorsement of the field by authors of the last three reviews of test theory in the Annual Review of Psychology (Keats, 1967; Bock & Wood, 1971; Lumsden, 1976). Another important stimulant of interest in the field was the publication of Lord and Novick's Statistical Theories of Mental Test Scores. They devoted five chapters (four of them written by Allen Birnbaum) to the topic of latent trait theory. A testimony to the current interest and popularity of the topic is the fact that the Journal of Educational Measurement will publish six invited papers on latent trait theory and applications in the summer issue of 1977.

What is latent trait theory? A theory of latent traits supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as traits, estimating scores for examinees on these traits and using the scores 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. For this reason, latent trait models are mathematical models. These mathematical models are based on specific assumptions about the test data. 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.

In view of the breakthroughs in several testing problem areas brought about by the use of latent trait theory, it is clear that the field of latent trait theory will become increasingly more important to measurement specialists and test practitioners. Therefore, given the newness of the field, its rapid growth in recent years, and the diversity of views and contributions, it seems apparent that a comprehensive review of the field is in order.

This document addresses four matters: First, the nature and characteristics of latent trait theory are introduced. Second, a review of many of the technical developments in the field is provided. Third, several promising applications of latent trait models are described. Finally, some additional areas for research and development are suggested.

Latent Trait Theory

Dimensionality of the latent space, local independence, and item characteristic curves are three important notions that arise in connection with latent trait theory. These three notions, along with a discussion of the ability scale, will be provided next.

Dimensionality of the Latent Space

In a general theory of latent traits, it is assumed that a set of k latent traits or abilities underlie examinee performance on a set of test items. The k latent traits can be used to define a k dimensional latent space, with each examinee's location in the latent space determined by the examinee's position on each latent trait. The number of dimensions of the latent space depends on the number of abilities measured by the test in the population of examinees the test is administered to. The latent space is referred to as complete if all latent traits influencing the test scores of a population of examinees have been specified.

It is commonly assumed that only one ability is necessary to "explain," or "account" for examinee test performance. Latent trait models that assume a single latent ability is sufficient to explain or account for examinee performance are referred to as unidimensional. Those models, that assume that more than a single ability is necessary to adequately account for examinee test performance, are referred to as multidimensional. The reader is referred to the work of Mulaik (1972) and Samejima (1974) for discussions of multidimensional latent trait models.

The assumption of a unidimensional latent space 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). What does it mean to say that a test is unidimensional? Suppose a test consisting of n items is intended for use in r subpopulations of examinees (e.g., several ethnic groups). Consider next the conditional distributions of test scores at a particular ability level for the r subpopulations. These conditional distributions for the r subpopulations will be identical if the test is unidimensional. If the conditional distributions vary across the r subpopulations, it can only be because the test is measuring something other than the single ability. Hence, the test cannot be unidimensional.

It is possible for a test to be unidimensional within one population of examinees and not unidimensional in another. Consider a test with a heavy cultural loading. This test could appear to be unidimensional for all populations with the same cultural background. However, when administered to populations with varied cultural backgrounds, it may in fact have more than a single dimension underlying the test score. Examples of this situation are seen when the factor structure of a particular set of test items varies from one cultural group to another.

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. 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 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."

Factor analysis can also be used to check the reasonableness of the assumption of unidimensionality with a set of test items (Hambleton & Traub, 1973). However, the approach is not without problems. For example, much has been written about the merits of using tetrachoric correlations or phi correlations (McDonald & Ahlawat, 1974). The common belief is that using phi correlations will lead to a factor solution with too many factors, some of them "difficulty factors" found because of the range of item difficulties among the items in the pool. McDonald and Ahlawat (1974) concluded that "difficulty factors" are unlikely if the range of item difficulties is not extreme and the items are not too highly discriminating.

Tetrachoric correlations have one attractive feature. A sufficient condition for the unidimensionality of a set of items is that the matrix of tetrachoric item intercorrelations has only one common factor (Lord & Novick, 1968). On the negative side, the condition is not necessary. Tetrachoric correlations are awkward to calculate (the formula is complex and requires some numerical integration), and, in addition, do not necessarily yield a correlation matrix that is positive definite, a problem when factor analysis is attempted.

Local Independence

The assumption of local independence states that the probability of an examinee answering a test item correctly is not affected by his or her performance on any other item in the test.

If we let U_g, g = 1, 2, ..., n, represent the binary responses (1, if correct; 0, if incorrect) of an examinee to a set of n test items, P_g = the probability of a correct answer by the examinee to item g, and $Q_g = 1 - P_g$, then the assumption of local independence leads to the following statement:

$$\text{Prob }\{U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n\}$$
$$= \prod_{g=1}^{n} P_g^{u_g} Q_g^{1-u_g}. \qquad [1]$$

That is, the probability of an examinee response pattern is given by the product of probabilities of the item responses.

One result of the assumption of local independence is that the frequency of test scores across examinees for fixed ability, denoted θ, is given by

$$f(x|\theta) = \sum_{\Sigma u_g = x} \prod_{g=1}^{n} P_g^{u_g} \; Q_g^{1-u_g}, \qquad [2]$$

where x is an examinee's test score which can take on values from 0 to n.

The assumption of local independence for the case when θ is unidimensional, and the assumption of a unidimensional latent space are equivalent. First, suppose a set of test items measure a common ability. Then, for examinees at a fixed ability level θ, item responses are statistically independent. For fixed ability level θ, if items were not statistically independent, it would imply that some examinees have higher expected test scores than other examinees of the same ability level. Consequently, more than one ability would be necessary to account for examinee test performance. This is a clear violation of the original assumption that the items were unidimensional. Second, the assumption of local independence implies that item responses are statistically independent for examinees at a fixed ability level. Therefore, only one ability is necessary to account for the relationship among a set of test items.

It is important to note that the assumption of local independence does not imply that test items are uncorrelated over the total group of examinees (Lord & Novick, 1968, p. 361). Positive correlations between pairs of items will result whenever there is variation among the examinees on the ability measured by the test items.

Because of the equivalence between the assumptionsof local independence and of the unidimensionality of the latent space, the extent to which a set of test items satisfy the assumption of local independence can also be studied using factor analytic techniques. Also, a rough check on the statistical independence of item responses for examinees at the same ability level was offered by Lord (1953a). His suggestion was to consider examinee item responses for examinees within a narrow range of ability. For each pair of items, a χ^2 statistic can be calculated to provide a measure of the independence of item responses. If the proportion of examinees obtaining each response pattern (00, 01, 10, 11) can be "predicted" from the marginals for the group of examinees, the item responses on the two items are statistically independent. The value of the χ^2 statistic can be computed for each pair of items, summed, and tested for significance. The process would be repeated for examinees located in different regions of the ability continuum.

Test and Item Characteristic Curves

The frequency distribution of test scores for a fixed level of θ can be obtained using Equation [2] defined in the previous section. The curve connecting the means of these distributions represents the regression of test scores on ability θ. If the test is unidimensional, this curve is referred to as a test characteristic curve (or test characteristic function if the latent space is multidimensional).

It is also possible to develop item characteristic curves in a similar manner. The frequency distribution of a binary item score for fixed ability θ can be written

$$f_g(u_g|\theta) = P_g^{u_g} Q_g^{(1-u_g)} \quad , \qquad [3]$$

$$\text{i.e.,} \quad f_g(u_g|\theta) \equiv P_g \text{ if } u_g = 1$$

$$\equiv Q_g \text{ if } u_g = 0.$$

The curve connecting the means of the conditional distributions, represented by Equation [3], is the regression of item score on ability and is referred to as an <u>item characteristic curve</u> (or <u>item characteristic function</u> if the latent ability space is multidimensional). An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it. In simple terms, it is the non-linear regression function of item score on the latent trait measured by the test.

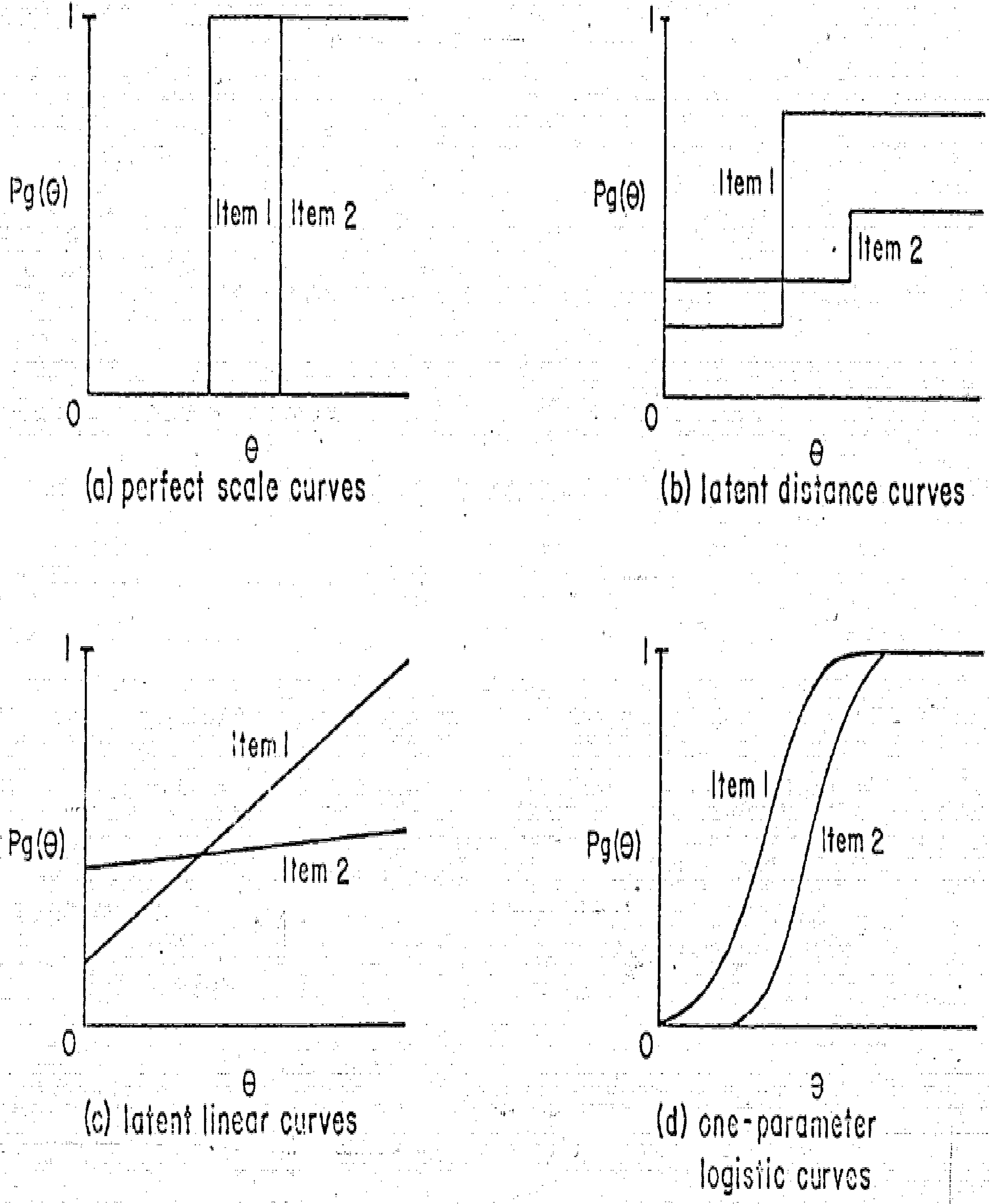
If the complete latent space is defined for the examinee populations of interest, the conditional distributions of item scores for fixed ability level must be identical across these populations. If the conditional distributions are identical, then the curves connecting the means of these distributions must be identical; i.e., the item characteristic curve will remain invariant across populations of examinees for which the complete latent space has been defined. Since the probability of an individual examinee providing a correct answer to an item depends only on the form of the item characteristic curve, it is <u>independent</u> of the distribution of examinee ability in the population of examinees of interest. Thus, the probability of a correct response to an item by an examinee will not depend on how many other examinees are located at the same ability level. In other words, the shape of an item characteristic curve does not depend on the distribution of ability in the examinee population. This invariance property of item characteristic curves and consequently the parameters describing the curves is one of the attractive characteristics of latent trait models. The invariance of latent trait item parameters has important implications for tailored testing, item banking, study of item bias, and other applications of latent trait models.

It is common to interpret $P_g(\theta)$ as the probability of an examinee answering item g correctly. Lord (1974b) questioned this interpretation and provided an example to show that this common interpretation of $P_g(\theta)$ leads to an awkward situation. Consider two examinees, a and b, and two items, i and j. Suppose examinee a knows the answer to item i and does not know the answer to item j. Consider the situation to be reversed for examinee b. Then, $P_i(\theta_a) = 1$, $P_j(\theta_a) = 0$, $P_i(\theta_b) = 0$, $P_j(\theta_b) = 1$. The first two equations suggest that item i is easier than item j. The other two equations suggest the reverse conclusion. One interpretation is that item i and j measure different abilities for the two examinees. Of course, this would make it impossible to compare the two students. One reasonable solution to the dilemma is to define the meaning of $P_g(\theta)$ differently. Lord suggests that $P_g(\theta)$ be interpreted as the probability of a correct response for the examinee across test items with near identical item parameters.

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. It is common, though, for the number of parameters to be one, two, or three. For example, the item characteristic curve of the latent linear model (Figure 1, c) has the general form $P_g(\theta) = b_g + a_g\theta$, where $P_g(\theta)$ designates the probability of a correct response to item g by an examinee with ability level θ. The function is described by two item parameters, item difficulty and item discrimination, denoted b_g and a_g respectively. An item characteristic curve is defined completely when its general form is specified and when the parameters of the curve for a particular item are known. Item characteristic curves of the latent linear model will vary in their intercepts (b_g) and slopes (a_g) to reflect the fact that the test items vary in "difficulty" and "discriminating power."

Item characteristic curves for Guttman's perfect scale model are shown in Figure 1 (a). These curves take the shape of step functions. Probabilities of correct responses are either 0 or 1. The critical ability level $\theta*$ is the point on the ability scale where probabilities change from 0 to 1. Different items lead to different values of $\theta*$. When $\theta*$ is high we have a difficult item, and when $\theta*$ is low, an easy item. Figure 1 (b) describes a variation on Guttman's "perfect scale" model. Item characteristic curves take the shape of step functions but the probabilities of incorrect and correct responses, in general, differ from 0 to 1. Figures 1 (d), (e), and (f) show "S" shaped curves representing logistic models, respectively. With the one-parameter logistic model, the item characteristic curves are non-intersecting curves that

Figure I. Seven examples of item characteristic curves.

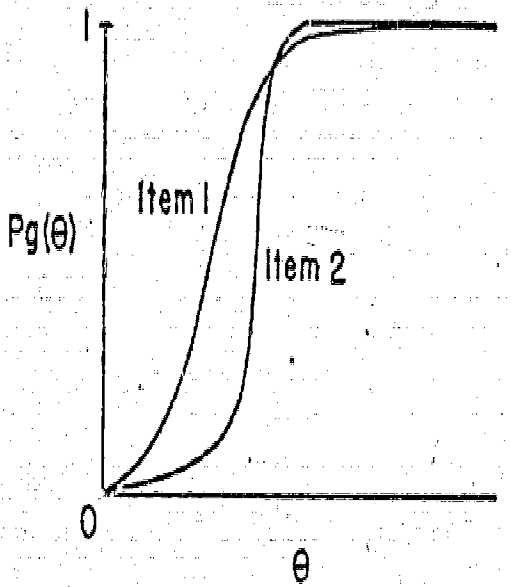


(a) perfect scale curves

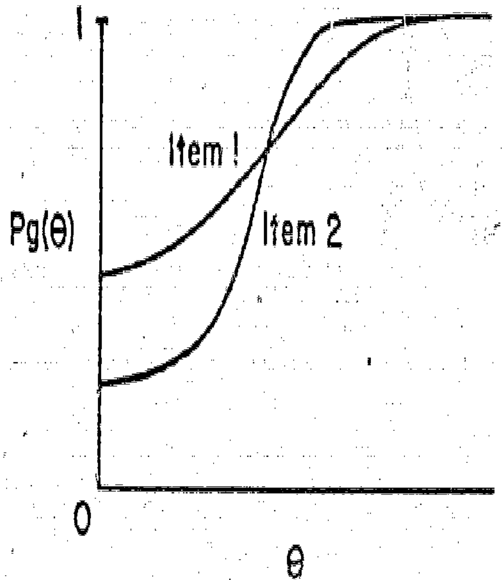
(b) latent distance curves

(c) latent linear curves

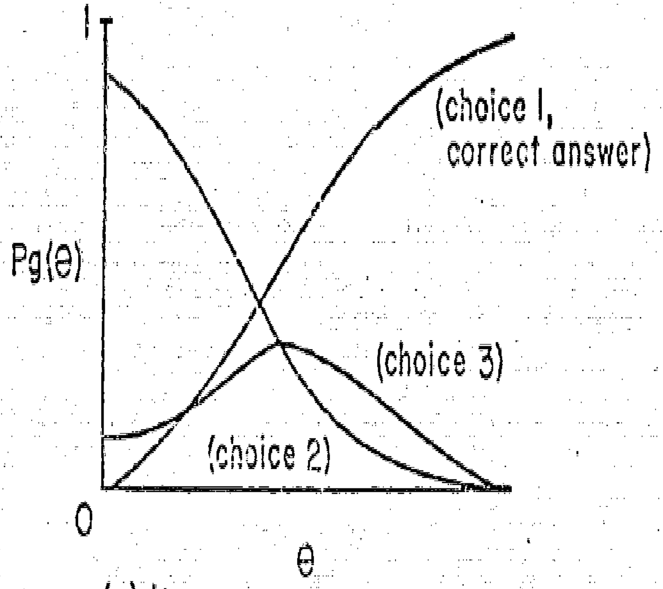
(d) one-parameter logistic curves



(e) two-parameter
logistic curves



(f) three-parameter
logistic curves



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

differ only by a translation along the θ scale. We say that items with such characteristic curves vary only in their difficulty. With the two-parameter logistic model, item characteristic curves vary in both slope (some curves increase more rapidly than others; i.e., the corresponding test items are more discriminating than others) and translation along the ability scale (some items are more difficult than others). Finally, with the three-parameter logistic model, curves may differ in slope, translation, and lower asymptote. With the one- and two-parameter logistic curves, the probabilities of correct responses range form 0 to 1. In the three-parameter model, the lower asymptote, in general, is greater than 0. When guessing is a factor in test performance, this feature of the item characteristic curve can improve the "fit" between the test data and the model. In other models (the nominal response model and the graded response model) there are item option characteristic curves. A curve depicting the probability of an item option being selected as a function of ability is produced for each option or choice in the test item. An example of this situation is shown in Figure 1 (g).

It is most common for a user to specify the mathematical form of the item characteristic curves before beginning his or her work. It is not easy to check on the appropriateness of the choice because item characteristic curves represent the regression of item scores on a variable (ability) that is not directly measurable. About the only way the assumption can be checked is to study the "validity" of the predictions with the item characteristic curves (Hambleton & Traub, 1973; Ross & Lumsden, 1968). More will be said about how to make these predictions later in the paper.

The Ability Scale

If we were to administer two tests, that measured the same ability, to the same group of examinees, and one test was more difficult than the other, we would obtain two different test score distributions. 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 true score, the expected test score for an examinee. What is the relationship between true scores and ability scores? Lord and Novick (1968) showed that the test characteristic curve, introduced earlier, provides the relationship. This is easily seen from the following argument. Consider the proportion-correct score, $Z = \frac{x}{n}$. Then

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

$$\mathrm{Var}\ (Z|\theta) = \frac{1}{n^2} \sum_{g=1}^{n} P_g(\theta)\ Q_g(\theta). \qquad [5]$$

$E(Z|\theta)$ 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, $E(Z|\theta) = T$, the true score. Also Var $(Z|\theta) \to 0$, as $n \to \infty$, and so T and θ will be related by a monotonic increasing transformation which is the test characteristic curve. Clearly then, the two concepts, T and θ, are the same, except for the scale of measurement used to describe each. One important difference is that true score is defined on the interval $[0, n]$ whereas ability scores are defined on the interval $[-\infty,+\infty]$.

There are other differences between true score and ability score. True score is defined for a particular test. It is the expected test score for an examinee. An examinee's true 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 true 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 larger "pool" of test items. Thus, while an examinee would have (in general) a different true 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.

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

$$E(X|\theta) = \sum_{g=1}^{n} P_g(\theta) \quad . \qquad [6]$$

Thus, ability scores provide a basis for content-referenced interpretations of examinee test scores. When the quantities in Equation [6] are scaled by $1/n$, $E(X/n|\theta)$ representsthe 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 norm-referenced interpretations as well.

Let us consider next how the metric for the ability sc.le is chosen. It is chosen so that the item characteristic curves have some specified mathematical form. On the basis of examinee test performance, examinees can be ordered on ability. The particular values of these abilities on the ability scale are chosen so as to maximize a criterion reflecting agreement between examinee item response data, predictions of the test data derived from the "best-fitting" item characteristic curves and optimally positioned ability scores on the ability scale. However, the origin and unit of measurement of the ability scale are arbitrary. Any linear transformation of the ability scores is permissible. Also, it has been suggested that when an external criterion measure with meaningful units can be located, a transformation be found to transform ability scores to this new scale. Such a transformation would enhance the interpretability of ability scores.

Lord (1975d) reported one rather distressing property of the ability scale observed in his work. Item parameters defined on this ability scale were found to be correlated in six sets of empirical data that he studied. Lord proposed a monotonic transformation of the ability scale to correct the problem. With the availability of computer programs, this operation could be routinely performed.

Latent Trait Models

The purpose of this section is to introduce several of the most commonly used latent trait models: The normal-ogive model, the one-, two-, and three-parameter logistic test models, the graded-response model, the nominal response model, and the continuous response model. All models assume that the principle of local independence applies and (equivalently) that the items in the test being fitted by a model measure a common ability. A significant distinction among the models is in the mathematical form taken by the item characteristic curves. A second important distinction among the models is the scoring.

Additional latent trait models are discussed by Lazarsfeld and Henry (1968), Lord and Novick (1968), and Torgerson (1958). Deterministic models (for example, Guttman's perfect-scale model) are of no interest to us here because they are not likely to fit most achievement and aptitude test data very well. Common test items rarely discriminate well enough to be fit by a deterministic model (Lord, 1974b).

(a) Normal-Ogive Model

Lord (1952, 1953b) proposed a latent trait model (although he was not the first psychometrician to do so) in which an item characteristic curve takes the form of the normal ogive:

$$P_g(\theta) = \int_{-\infty}^{a_g(\theta-b_g)} \phi\,(t)\;dt,\;(g = 1,2,\ldots,n) \qquad [7]$$

where $P_g(\theta)$ is the probability that an examinee with ability θ, answers item g correctly, $\phi(t)$ is the normal density function, and b_g and a_g are parameters characterizing item g. The parameter b_g is usually referred to as the index of item difficulty. It represents the point on the ability scale at which an examinee has a 50% probability of answering the item correctly. The parameter a_g, called item discrimination, is proportional to the slope of $P_g(\theta)$ at the point $\theta = b_g$.

The item difficulty parameter, b_g, is defined on the same scale as ability $[-\infty,+\infty]$. In practice though, the range of b_g is from about -2 to +2 (assuming the ability distribution has been scaled to be approximately on the range from -3 to +3). Values of b_g near -2 correspond to items that are very easy and values of b_g near +2 correspond to items that are very difficult for the group of examinees.

The item discrimination parameter, a_g, is defined, theoretically, on the scale $[-\infty,+\infty]$. However, negatively discriminating items are discarded from ability tests. Also, it is unusual to obtain a_g values larger than two. Hence, the usual range for item discrimination parameters is [0, 2]. High values of a_g result in item characteristic curves that are very "steep." Low values of a_g lead to item characteristic curves that increase gradually as a function of ability.

(b) Two-Parameter Logistic Model

Birnbaum (1968) proposed a latent trait model in which the item characteristic curve takes the form of a two-parameter logistic distribution function,

$$P_g(\theta) = \frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}} \quad (g = 1,2,\ldots,n). \qquad [8]$$

Birnbaum substituted the two-parameter logistic cumulative distribution function for the normal-ogive function as the form of the item characteristic curve. This model has the important advantage of being more mathematically tractable than the normal ogive model. $P_g(\theta)$, b_g, a_g, and θ have essentially the same interpretation as in the normal ogive model. The constant D is a scaling factor. It has been shown that when $D = 1.7$, values of $P_g(\theta)$ for the normal ogive and two-parameter logistic models differ absolutely by less than .01 for all values of θ (Haley, 1952).

Careful inspection of the two-parameter normal ogive and logistic test models reveals an additional implicit assumption that is characteristic of most latent trait models: Guessing does not occur. This must be so since for all items with $a_g>0$ (that is, items for which there is a positive relationship between performance on the test item and the ability measured by the test), the probability of a correct response to the item decreases to zero as ability decreases.

(c) Three-Parameter Logistic Model

The three-parameter model can be obtained from the two-parameter model by adding a third parameter, denoted c_g. The mathematical form of the three-parameter logistic curve is written

$$P_g(\theta) = c_g + (1-c_g)\frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}} \qquad (g = 1,2,\ldots,n). \qquad [9]$$

The parameter c_g is the lower asymptote of the item characteristic curve and represents the probability of low ability examinees correctly answering an item. The purpose of including a parameter c_g in the model is to attempt to account for the misfit of item characteristic curves at the low end of the ability continuum, where among other things, guessing is a factor in test performance. It has been common to refer to the parameter c_g as the guessing parameter in the model.

It is perhaps surprising to note that the parameter c_g typically assumesvalues that are smaller than the valuesthat would result if examinees of low ability were to guess randomly on the item. As Lord (1974a) has noted, this situation 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 would seem to be desirable.

(d) One-Parameter Logistic Model (Rasch Model)

In the last decade, many researchers have become aware of work in the area of latent trait models by Georg Rasch, a Danish mathematician (Rasch, 1966), both through his own publications and the papers of others advancing his work (Anderson, Kearney, and Everett, 1968; Wright, 1968, 1977a, 1977b; Wright and Panchapakesan, 1969). Although the Rasch model was developed independently of other latent trait models and along quite different lines, Rasch's model can be viewed as a latent trait model in which the item characteristic curve is a one-parameter logistic function. Consequently, Rasch's model is a special case of Birnbaum's two-parameter logistic model, in which all items are assumed to have equal discriminating power and vary only in terms of difficulty. The equation of the item characteristic curve for this model can be written as

$$P_g(\theta) = \frac{e^{D\bar{a}(\theta-b_g)}}{1+e^{D\bar{a}(\theta-b_g)}} \qquad (g = 1,2,\ldots,n), \qquad [10]$$

in which $\bar{a}$, the only term not previously defined, is the common level of discrimination for all the items. Wright (1977a) prefers to write the model with $D\bar{a}$ incorporated into the θ scale. Thus, the right-hand side of the probability statement becomes $\frac{e^{\theta'-b_g'}}{1+e^{\theta'-b_g'}}$.

The assumption that all item discrimination parameters are equal is restrictive, and substantial evidence is available which suggests that unless test items are specifically chosen to have this characteristic, the assumption will be violated (e.g. Birnbaum, 1968; Hambleton & Traub, 1973; Lord, 1968; Ross, 1966).

While the Rasch model is a special case of the two- and three-parameter logistic test models, it does have some special properties that make it especially attractive to users. For one, since the model involves fewer item parameters, it is easier to work with. Two, the problem of parameter estimation is essentially solved. This point will be discussed in a later section.

There appears to be some misunderstanding of the ability scale for the Rasch model. Wright (1968) originally introduced the model this way: The odds in favor of success on an item, denoted O_{gi}, are given by the product of an examinee's ability θ_i^* and, the reciprocal of the difficulty of the item, $1/b_g^*$. Odds for success will be higher for brighter students and/or easier items. The odds of success are defined as the ratio of P_{gi} to $1-P_{gi}$, where P_{gi} is the probability of success by examinee i on item g. Therefore,

$$\frac{\theta_i^*}{b_g^*} = \frac{P_{gi}}{1-P_{gi}} \qquad [11]$$

or

$$P_{gi} = \frac{\theta_i^*}{b_g^* + \theta_i^*} \quad . \qquad [12]$$

Equation [10] can be obtained from Equation [12] by setting $\theta^* = e^{D\bar{a}\theta}$ and $b_g^* = e^{D\bar{a}b_g}$. In Equation [12], both θ^* and b_g^* are defined on the interval $[0,+\infty]$. If log ability and log difficulties are considered, then θ and b_g, and log θ^* and log b_g^* are measured on the same scale, $[-\infty,+\infty]$, differing only by an expansion transformation.

We return again to the point above regarding the odds for success on an item. Clearly, there is an indeterminancy in the product of θ_i^* and $1/b_g^*$. When odds for success are changed, we could attribute the change to either θ_i^* or $1/b_g^*$. For example, if odds for success are doubled, it could be because ability is doubled or because the item is half as difficult. There are several ways to remedy the problem. For one we could choose a special set or "standard set" of test items, and scale the b_g's, $g = 1,2,\ldots,n$ so that $\bar{b}_g = 1$. Alternately, we could do the same sort of scaling for a "standard" set of examinees such that the average of θ_i, $i = 1,2,\ldots,N$ is set to one. The final point is clear. When one item is twice as easy as another, a person's odds for success on the easier item are twice what they are on the harder item. If one person's ability is twice as high as another person's ability, the first person's odds for success are twice those of the second person (Wright, 1968). In what sense are item and ability parameters measured on a ratio scale? An examinee with twice the ability (as measured on the Rasch ability scale) of another examinee, has twice the odds of successfully answering a test item. Also, when one item is twice as easy as another item (again, as measured on the Rasch ability scale), a person has twice the odds of successfully answering the easier one. The other latent trait models do not permit this particular kind of interpretation of item and ability parameters.

(e) Nominal Response Model

The one-, two-, and three-parameter logistic test models can only be applied to test items which are scored dichotomously. The nominal response model, introduced by Bock (1972) and Samejima (1972), is applicable when items are multichotomously scored. The purpose of the model is to maximize the precision of obtained ability estimates by utilizing the information contained in each response (correct or incorrect) to an item. This approach represents another method in the search for differential scoring weights that improve the reliability and validity of mental test scores (Wang and Stanley, 1970). Each item option is described by an item option characteristic curve. Even the "omit" response can be represented by a curve. For the correct response, the curve should be monotonically increasing as a function of ability. For the incorrect options, the shapes of the curves depend on how the options are perceived by examinees at different ability levels.

There are, of course, many choices for the mathematical form of the item option characteristic curves (Samejima, 1972). For one, Bock (1972) assumed the probability that an examinee with ability level θ will select a particular item option k (from m available options per item) to item g is given by

$$P_{gk}(\theta) = \frac{e^{b^*_{gk} + a^*_{gk}\theta}}{\sum_{h=1}^{m} e^{b^*_{gh} + a^*_{gh}\theta}} \qquad (g = 1,2,\ldots,n;\ k = 1,2,\ldots,m). \qquad [13]$$

For any ability level θ, the sum of the probabilities of selecting each of the m item options is equal to one. The quantities b^*_{gk} and a^*_{gk} are item parameters related to the k^{th} item option. When m=2, the items are dichotomously scored and the two-parameter logistic model and the nominal response model are identical.

(f) Graded Response Model

This model was introduced by Samejima (1969) to handle the testing situation where item responses are made into two or more ordered categories. For example, with test items like those on the Raven's Progressive Matrices, one may desire to score examinees on the basis of the correctness (for example, incorrect, partially correct, correct) of their answers. Samejima (1969) assumed any response to an item g can be classified into $m_g + 1$ categories, scored $x_g = 0, 1, \ldots, m_g$, respectively. Samejima (1969) introduced the operating characteristic of a graded response category. She defines it as

$$P_{x_g}(\theta) = P^*_{x_g}(\theta) - P^*_{(x_g+1)}(\theta) \quad . \qquad [14]$$

$P^*_{x_g}(\theta)$ is the regression of the binary item score on latent ability, when all the response categories less than x_g are scored 0 and those equal to or greater than x_g are scored 1. $P_{x_g}(\theta)$ represents the probability with which an examinee of ability level θ receives a score of x_g. The mathematical form of $P^*_{x_g}$ is specified by the user. Samejima (1969) has considered both the two-parameter logistic and two-parameter normal-ogive curves in her work. In several applications of the graded response model, it has been common to assume that discrimination parameters are equal for $P^*_{x_g}(\theta)$, $x_g = 0, 1, \ldots, m_g$. This model is referred to as the <u>homogeneous case</u> of the graded response model. Further, Samejima defines $P^*_o(\theta)$ and $P^*_{(m_g+1)}(\theta)$ so that

$$P^*_o(\theta) = 1 \qquad [15]$$

and

$$P^*_{(m_g+1)}(\theta) = 0 \quad , \qquad [16]$$

Also, for any response category x_g,

$$P_{x_g}(\theta) = P^*_{x_g}(\theta) - P^*_{(x_g+1)}(\theta) > 0 \qquad [17]$$

The shape of $P_{x_g}(\theta)$, $x_g = 0, 1, \ldots, m_g$, will in general be non-monotonic except when $x_g = m_g$, and $x_g = 0$. (This is true as long as $P^*_{x_g}(\theta)$ is monotonically increasing, for all $x_g = 0, 1, \ldots, m_g$.)

(g) <u>Continuous Response Model</u>

The continuous response model can be considered as a limiting case of the graded response model. This model was introduced by Samejima (1973b) to handle the situation where examinee item responses are marked on a continuous scale. The model is likely to be useful, for example, to social psychologists interested in studying attitudes.

Estimation of Parameters

Once the various assumptions such as unidimensionality and local independence have been made regarding the latent variables, and the form of the item characteristic curve is specified, the problem of estimating the parameters of the latent trait model arises. If, say the two parameter normal ogive or the logistic model, is deemed appropriate, and if n items are administered to N examinees, the parameters that have to be estimated are the 2n item parameters pertaining to item difficulty and item discrimination, and the N parameters that correspond to the abilities of the examinees. Owing to the large number of parameters which may result when a large number of examinees are involved, the estimation of parameters in latent trait models present substantial statistical and numerical problems. The statistical problems that arise in the estimation of parameters are related to the nature and properties of the estimates. The numerical problems, on the other hand, arise in connection with the solution of the estimation equations and are related to the convergence of the algorithms employed to solve the equations.

The basic statistical problem associated with estimation of parameters in latent trait models arises when the item parameters have to be estimated simultaneously with the large number of ability parameters. In this situation, the item parameters are common to all the N observations and hence are called "structural parameters." The ability parameters, called "incidental parameters," on the other hand, are specific to the individual observations and hence increase with the number of observations. The problem of estimating structural parameters in the presence of incidental parameters has been studied by various authors. Neyman and Scott (1948) and Kendall and Stuart (1973, pp. 62) have shown that the maximum likelihood estimates of the structural parameters in the presence of incidental parameters are not consistent. More recently, Andersen (1973a) has demonstrated that consistent maximum likelihood estimates of the structural or item parameters in a one-parameter latent trait model do not exist when the ability parameters and the item parameters are estimated simultaneously.

The estimation of the parameters of the latent trait models requires the determination of the values of the parameters that maximize the likelihood function if maximum likelihood estimates are sought. The likelihood function, which will be defined a little later, is rather complex and is a function of a large number of variables. The problem of finding the extreme values of a function of several variables is not trivial and often requires numerical methods. These numerical procedures are iterative in nature, requiring some starting values for the parameters in question and these are then iterated upon until the sequence of values converges. Often, the convergence of the sequence may be rather slow, or, if the sequence does converge, it may not converge to the true solution. A case in point is the three-parameter logistic model. Samejima (1973a) has shown that the likelihood function for the estimation of ability parameters in a three-parameter logistic model (under the assumption that the item parameters are known) may not possess a unique maximum. In this case, since a unique maximum does not exist, depending on the starting value, the sequence may converge to a value that corresponds to a local maximum. Thus, the values of the parameters that maximize the likelihood function,or the estimates, will not be the true maximum likelihood estimates of the parameters. In the case of the three-parameter logistic model with known values for item parameters, Samejima (1973a) has provided conditions under which the likelihood function possesses a unique maximum. However, when the item parameters are not known and have to be estimated, the likelihood function which is a function of the item parameters as well as the ability parameters, may not possess a unique maximum, and hence, the values of the parameters that maximize the likelihood function may not correspond to the true maximum likelihood estimates.

Despite the statistical and numerical problems mentioned above, the literature in latent trait theory abounds with procedures for estimating the parameters that arise in latent trait models. These estimation procedures, which have been developed over the past thirty years range from heuristic procedures such as those given by Urry (1974) and Jensema (1976) to conditional as well as unconditional maximum likelihood procedures (Andersen, 1970, 1972, 1973a, 1973b; Bock, 1972; Lord, 1968, 1974b; Samejima, 1969; Wright and Panchapakesan, 1969; Wright and Douglas, 1977) and empirical as well as true Bayesian procedures (Birnbaum, 1969; Meredith and Kearns, 1973; Owen, 1975). These procedures are discussed next, and although there are severe problems with the estimation of parameters in latent trait models, in some instances these problems can be overcome.

Maximum Likelihood Estimation in Latent Trait Models

We assume that an examinee is administered n dichotomously scored items and that the underlying latent space is unidimensional. Let V be a vector of binary random variables such that

$$V = [U_1 \; U_2 \ldots \; U_g \ldots \; U_n],$$

and, v, a particular realization of V such that

$$v = [u_1 \; u_2 \ldots \; u_g \ldots \; u_n].$$

The random variable U_g takes on the value u_g where $u_g = 1$ if the examinee responds correctly to the item and $u_g = 0$ otherwise. We also denote

$$P_g(\theta) = \text{Prob } [U_g = 1|\theta]$$

and

$$Q_g(\theta) = 1-P_g(\theta) = \text{Prob } [U_g = 0|\theta] \quad .$$

Hence, the frequency distribution of the binary item score, for fixed θ can be written as

$$\begin{aligned} f_g(u_g|\theta) &= \text{Prob } [U_g = u_g|\theta] \\ &= P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \\ &\equiv \begin{cases} P_g(\theta) & \text{if } u_g = 1 \\ Q_g(\theta) & \text{if } u_g = 0, \end{cases} \end{aligned}$$

Thus, the conditional probability of a response vector, $V = v$, for fixed θ can be expressed as

$$\text{Prob } [V = v|\theta] = \text{Prob } [U_1 = u_1, U_2 = u_2. \ . \ ., U_n = u_n|\theta].$$

It then follows from the principle of local independence that

$$\begin{aligned} \text{Prob } [V = v|\theta] &= \prod_{g=1}^{n} \text{Prob } [U_g = u_g|\theta] \\ &= \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \ . \end{aligned}$$

If the n items are administered to a group of N examinees, then the likelihood function or the joint probability distribution of the response patterns for the N examinees for fixed ability levels $\theta_1, \theta_2, \ldots, \theta_N$, is given by

$$\begin{aligned} L &\equiv \text{Prob } [V_1 = v_1, V_2 = V_2, \ldots, V_n = v_n|\theta] \\ &= \prod_{k=1}^{N} \prod_{g=1}^{n} P_g(\theta_k)^{u_g} Q_g(\theta_k)^{1-u_g} \ . \end{aligned} \quad [18]$$

The function, $P_g(\theta)$, the probability that an examinee with ability θ responds correctly to item g, is the regression function of any item response

u_g on θ, and is more commonly referred to as the item characteristic curve. The item characteristic curve is a function of the item parameters such as the indices of item difficulty and discrimination as well as the ability, θ_k, of the k^{th} examinee. Once the form of the item characteristic curve is specified, the maximum likelihood estimates of the item parameters and the ability parameters can be determined as those values that maximize the likelihood function L given by Equation [18].

The item characteristic curve, $P_g(\theta)$ can take one of several forms as mentioned earlier. We shall discuss the procedure for obtaining maximum likelihood estimates only for the one-, two-, and three-parameter logistic models as these are the models thar are frequently used. The one-parameter logistic model better known as the Rasch Model, has the item characteristic curve $P_g(\theta)$, given by Equation [10]. In this case, $P_g(\theta)$ is a function of the item difficulty parameter, b_g, and the ability parameter, θ, more appropriately denoted as θ_k, corresponding to the ability of the k^{th} examinee. The two-parameter logistic model for which the item characteristic curve $P_g(\theta)$ given by Equation [8] is a function of a_g, the discriminating power of the item, b_g, the item difficulty index, and θ_k, the ability of the k^{th} examinee. Similarly, the item characteristic curve for the three-parameter logistic model given by Equation [9], in addition to being a function of the parameters a_g, b_g, and θ_k, is a function the parameter c_g. In general, if we let $\Psi(x)$ denote the function

$$\Psi(x) = \exp x/(1 + \exp x), \quad [19]$$

then the item characteristic curve for the one parameter model or the Rasch model $P_{1g}(\theta)$, is given by

$$P_{1g}(\theta_k) = \Psi(\theta_k - b_g). \quad [20]$$

The item characteristic curve for the two-parameter model is given by

$$P_{2g}(\theta_k) = \Psi[a_g(\theta_k - b_g)] \quad [21]$$

while the item characteristic curve for the three-parameter model is given by

$$P_{3g}(\theta_k) = c_g + (1-c_g)\ \psi[a_g(\theta_k-b_g)]. \qquad [22]$$

In order to obtain the maximum likelihood estimates of the parameters it is necessary to solve the likelihood equations

$$\partial \log L_1/\partial b_g = 0,\ \partial \log L_1/\partial \theta_k = 0\ ,\ g=1,\ldots,n;k=1,\ldots,N, \qquad [23]$$

for the Rasch model, the equations

$$\partial \log L_2/\partial a_g = 0,\ \partial \log L_2/\partial b_g = 0,\quad \log L_2/\partial \theta_k = 0 \qquad [24]$$

for the two-parameter logistic model, and the equations

$$\partial \log L_3/\partial a_g = 0,\ \partial \log L_3/\partial b_g = 0,\ \ \partial \log L_3/\partial \theta_k = 0, \qquad [25]$$

$$\partial \log L_3/\partial c_g = 0$$

for the three-parameter logistic model. The likelihood functions L_1, L_2, and L_3 are obtained by substituting $P_{1g}(\theta)$, $P_{2g}(\theta)$ and $P_{3g}(\theta)$ respectively for $P_g(\theta)$ in Equation [18]. Since the scale and origin of θ is not fixed, we have to solve simultaneously n+N-2 equations for the one-parameter model, 2n+N-2 equations for the two-parameter model, and 3n+N-2 equations for the three-parameter model. The exact form of these equations are not given here since they are well documented (Birnbaum, 1968; Wright & Douglas, 1977, in press).

Solutions to the likelihood equations discussed above are, unfortunately, not available in closed form. Hence, numerical procedures have to be employed to obtain the solutions of the likelihood equations. Procedures for solving these equations have been suggested by various writers. Birnbaum (1968, pp. 422) suggests a heuristicprocedure that involves specifying starting values for the parameters, substituting these in the likelihood equations and iterating until convergence takes place. Although this is an appealingly simple procedure, it is inefficient and convergence to the true solution is not guaranteed. A more satisfactory procedure is the

Newton-Raphson procedure suggested by Bock and Liebermann (1970), Bock (1972), and Wright and Douglas (1977). In general, if the equations to be solved are of the form $f(\underline{\alpha}) = \underline{0}$ where $\underline{\alpha}$ is a vector of unknowns, and $f'(\underline{\alpha})$ is the matrix of the derivatives of $f(\underline{\alpha})$ with respect to the vector of parameters, then the (i+1)th approximation to the solution of the system $f(\underline{\alpha}) = \underline{0}$, $\underline{\alpha}_{i+1}$, is given by

$$\underline{\alpha}_{i+1} = \underline{\alpha}_i - [f'(\underline{\alpha}_i)]^{-1} f(\underline{\alpha}_i) \qquad [26]$$

where $\underline{\alpha}_i$ is the ith approximation to the solution of $f(\underline{\alpha}) = 0$. Thus, the Newton-Raphson procedure, in this case, requires the evaluation of the matrix of second derivatives, or the Hessian, of the logarithm of the likelihood function with respect to the parameters. Although the Newton-Raphson procedure is more tedious than the simpler procedures, the convergence of the Newton-Raphson algorithm is quadratic, at least in the neighborhood of the solution vector. In addition, the Hessian evaluated at the maximum of the log-likelihood function yields the inverse of the asymptotic dispersion matrix of the maximum likelihood estimates (the expression for the asymptotic dispersion matrix is given in a later section). Alternatively, the Method of Scoring (Rao, 1965, p. 302) can be employed to solve the likelihood equations. The Method of Scoring is essentially the Newton-Raphson procedure, but employs the asymptotic dispersion matrix in place of the inverse of the Hessian in the iterative sequence. Although this procedure can be slow in convergence when compared to the Newton-Raphson procedure, it is computationally simpler, since the asymptotic dispersion matrix does not have to be updated at each iteration. In addition, the asymptotic dispersion matrix is positive definite while the Hessian may become indefinite at some stages of the iteration, a fact that causes convergence problems in some instances.

The maximum likelihood procedure discussed above has been employed for the simultaneous estimation of item parameters and ability parameters by such authors as Birnbaum (1968), Lord (1968), Wright and Panchapakesan (1969), and Wright and Douglas (1977). An excellent discussion of some of the estimation problems that are encountered in practice for the three-parameter model is given by Lord (1968). He points out that the iterative procedure fails to converge unless the number of items and the number of examinees is large. If either n or N is small, the estimates of the item discrimination indices may increase without bound. Wright (1977a) has suggested a reason why this may happen by using an argument based on the traditional method of estimating item discrimination. In order to estimate the item parameters, estimates of the abilities of the examinees are first obtained. Birnbaum (1968) has shown that a sufficient statistic for estimating the ability θ_i for the ith examinee is given by $\sum_j a_j u_{ij}$, where u_{ij} is the score of the ith examinee on the jth item. Since $a_g = \rho_g/(1-\rho_g^2)^{\frac{1}{2}}$ (assuming guessing is minimal and ability is normally distributed), where ρ_g is the correlation between θ and u_g (Lord and Novick, 1968, p. 378), an initial value for $\hat{a}_g$ can be obtained when ρ_g is known and the assumptions are met by the test data. The item scores are then weighted by these values to yield an estimate for the ability of an examinee. This procedure is iterated until a stable value of $\hat{a}_g$ is obtained. However, during the iteration, the ρ_g that was the largest gets larger until it dominates the weighted combination of item scores, and in the next step, results in a ρ_g that approaches unity, thus driving the value of $\hat{a}_g$, beyond bounds. Lord (1968) suggests imposing an upper limit for $\hat{a}_g$ based on the largest permissible correlation between the item and the ability. This suggestion which is not unlike that of incorporating prior beliefs into the estimation procedure, produces estimates that are reasonable.

Similar problems arise when estimating the ability of the examinee. Unlike the a_g, infinite values (positive or negative) are permissible for θ_i, and can be expected whenever an examinee obtains a perfect score on the test or fails to score correctly even on one item. These infinite values can be avoided by using a bounded function of θ_i instead of θ_i itself. However, as the occurrence of infinite values for θ_i itself causes no theoretical problems, it is not necessary to compensate for this.

Another problem, noted by Lord (1968), is that the entire iterative procedure may fail to converge or may converge extremely slowly. Lord (1968) employed the Method of Scoring for the solution of the likelihood equations. Although the procedure converges quadratically in the neighborhood of the maximum, the convergence is rather slow when the starting values given are far from the maximum. In some instances, poor starting values cause the procedure to diverge. One possible solution, obviously, is to provide good initial values, or, employ a linear search procedure like the method of steepest ascent, and then switch over to the Method of Scoring (or the Newton-Raphson) when the linear procedure slows down in convergence. The linear search process does not seem to have been incorporated in the existing algorithms and its efficacy needs to be investigated further.

The major statistical problem that remains with the simultaneous estimation of item parameters and the ability parameters is that these maximum likelihood estimates do not enjoy the properties they are usually accorded. Andersen (1973b) points out that maximum likelihood estimates of the item parameters and the ability parameters, when estimated simultaneously, are not consistent. This is true in general when structural parameters are estimated in the presence of incidental parameters. Thus, the procedure advocated by Wright and Panchapakesan (1969), Birnbaum (1968), and Lord (1968) may not yield consistent estimates of the parameters. Since the estimates may not be consistent, they may not even be asympototically unbiased.

Wright and Douglas (1977) provide a correction for the asymptotic bias of the estimates of the parameters in the Rasch model. However, it should be pointed out that this correction does not necessarily guarantee the consistency of the estimates.

The likelihood function given by Equation [18] is, in the strict sense, a conditional likelihood function of the item parameters and ability parameters, i.e.,

$$L \equiv L\ (u_{11},\ u_{12},\ldots,\ u_{1n},\ldots\ldots,\ u_{Nn}|\underline{\gamma},\ \theta_1,\ldots,\theta_N)$$

where γ is the vector of item parameters, θ_1, $\theta_2,\ldots,$ θ_N are the abilities of the examinees, and u_{ij} is the score of the ith examinee on the jth item. As the sample size increases and approaches infinity, the number of ability parameters or incidental parameters increases without bound. Instead of becoming stable as the sample size increases, the maximum likelihood estimates become ineffective--in fact they are not even consistent (Neyman and Scott, 1948; Kendall and Stuart, 1973, p. 63; Andersen, 1973b). Thus, the likelihood function, when expressed as a function conditional upon the item and the ability parameters, does not yield estimators with desirable properties. The problem can be overcome if it is possible to express the conditional likelihood function in terms of only the item parameters. When this is possible, the item parameters can be estimated without reference to the ability parameters and the estimates can be expected to have the desirable properties that maximum likelihood estimates usually possess.

The likelihood function involving the item parameters can be expressed independently of the ability parameters if a minimal sufficient statistic T_i for θ_i exists such that T_i does not depend on the item parameters. Then, the conditional maximum likelihood estimator of $\underline{\gamma}$, the item parameters, is defined as the value of $\underline{\gamma}$ that maximizes

$$L\ (u_{11},\ u_{12},\ldots\ldots,\ u_{Nn}|\underset{\sim}{\gamma},\ t_1,\ t_2,\ldots,\ t_N).$$

Since, by definition of a minimal sufficient statistic, the likelihood function conditional on $T_i = t_i$ is independent of the ability parameters $\theta_1, \theta_2,\ldots, \theta_N$, the vector of item parameters $\underset{\sim}{\gamma}$ can be estimated without any reference to $\theta_1, \theta_2,\ldots, \theta_N$. Andersen (1970) has shown that such conditional maximum likelihood estimators are consistent and asymptotically normally distributed. Conditional maximum likelihood estimators that are consistent and that are asymptotically normally distributed have been obtained for the Rasch model (Andersen, 1972, 1973a, 1973b). For the Rasch model, $T_i = \sum_j u_{ij}$, the total score for individual i, is a sufficient statistic for θ_i (Birnbaum, 1968, p. 429) and is independent of the item parameters. Thus the conditional likelihood function is given by

$$L\ (u_{11},\ u_{12},\ldots,\ u_{Nn}\Big|\ t_1,\ t_2,\ldots,\ t_N;\ b_1,\ b_2,\ldots,\ b_n)$$

$$= \exp(-\sum_{i=1}^{N}\sum_{j=1}^{n} b_j u_{ij}) / \prod_{i=1}^{N} r(t_i;\ b_1,\ b_2,\ldots,\ b_n),$$

where

$$r(t_i;\ b_1,\ b_2,\ldots,\ b_n) = \sum_{t_i} \exp(-\sum_{j=1}^{n} u_{ij}\ b_j).$$

The summation $\sum_{t_i}$ is over all response vectors with $\sum_{j=1}^{n} u_{ij} = t_i$. The conditional likelihood function given above is independent of the ability parameters, θ_i, and hence the item parameters b_j can be estimated without any reference to the ability parameters. The resulting likelihood equations (Andersen, 1970) cannot be solved in the closed form. A numerical procedure for the solution of the likelihood equations, based on the Method of Scoring, is given by Andersen (1972) and the reader is referred to this paper for details of the procedure.

In the two-parameter model, Birnbaum (1968) has shown that a sufficient statistic for θ_i is $\sum_j a_j u_{ij}$. However, this statistic is a function of the unknown parameters a_j, and hence it is not possible to express the conditional likelihood function as a function of only the item parameters. However, it is possible to express the likelihood function as a function of the item parameters alone if it is possible to view the examinees as a random sample from a known population. If we denote the density function of the ability parameter θ_i as $q(\theta)$, and the jth pattern of item responses by the vector

$$\underline{v}_j = (u_{1_j}, u_{2_j}, \ldots, u_{n_j}),$$

then

$$\text{Prob } [\underline{v}_j | \underline{\gamma}] = \int_{-\infty}^{\infty} \prod_{g=1}^{n} P_g^{u_g} Q_g^{1-u_g} q(\theta)\, d\theta$$
$$\equiv \pi_j$$

where $\underline{\gamma}$ is the vector of item parameters (Lord and Novick, 1968, p. 362). When the items are dichotomously scored, there are in all 2^n score patterns. If N examinees are randomly sampled from the population, the number of examinees with response pattern j is r_j, where $r_j = Np_j$, and $E(p_j) = \pi_j$. Thus, the number of examinees with the jth response pattern are distributed multinomially with parameters N and π_j, whence we obtain the likelihood function conditional only on the item parameter, as

$$L = N! \prod_{j=1}^{2^n} \pi_j^{r_j} / \prod_{j=1}^{2^n} r_j! \quad .$$

On maximizing this likelihood function with respect to the item parameters, we obtain the maximum likelihood estimates of the parameters.

Bock and Liebermann (1970) and Bock (1972) have named the estimates obtained by maximizing the likelihood function given above, the unconditional maximum likelihood estimates, since the likelihood function is not conditioned on the ability parameters. The term "unconditional" estimates has been used in a different sense by Wright and Douglas (1977) and should not be confused with the usage of the term in this paper. Wright and Douglas (1977) term their procedure for the simultaneous estimation of item and ability parameters as the unconditional procedure, in contrast to the conditional procedure provided by Andersen (1972). In the present usage, the estimates obtained by Wright and Douglas (1977) are conditional, since the likelihood function employed by them is conditional on the ability parameters.

Bock and Liebermann(1970) and Bock (1972) have outlined a procedure for the unconditional maximum likelihood estimation of the parameters. The procedure introduces a further complication to the already complex estimation procedure. It is necessary to integrate the likelihood function with respect to θ. As this integral cannot be evaluated in the closed form, numerical integration procedures have to be employed. In addition to this, the likelihood function requires the evaluation of 2^n response patterns, a tedious task when a large number of items is involved. Finally, the problem of specifying the density function of the latent variable θ has to be faced. Bock and Liebermann(1970) and Bock (1970) assumed that θ is distributed normally with zero mean and unit variance, an assumption that may not be realistic.

Despite these problems, the unconditional procedure has theoretical advantages over the conditional procedure in the two- and three-parameter models. Kiefer and Wolfowitz (1956) have shown that in structural models, if the incidental parameters are independently and identically distributed, then the maximum likelihood estimates of the structural parameters are consistent under regularity conditions. In the unconditional approach the ability parameters are assumed to be independently and identically distributed and hence the unconditional maximum likelihood estimator can be expected to be consistent. Thus, as Bock and Liebermann (1970) point out, the unconditional procedure provides a standard to which other solutions, can be compared. A further justification is that when calibrating items, it is not necessary to estimate the ability parameters and item parameters simultaneously. Hence, a sample of individuals can be randomly selected from a desired population and the item parameters estimated without any reference to the ability parameters. The estimatesof the item parameters, since they have some of the optimal properties, can be treated as known entities when estimating the ability of a group of examinees to whom the items are later administered. This procedure is particularly attractive since the estimation of ability parameters, when item parameters are known, is relatively straightforward and the ability estimates, in this case, possess the properties that are usually accorded the maximum likelihood estimates.

The estimation of ability parameters, when the item parameters are known, has been discussed by various authors (Lord, 1974a; Samejima, 1969, 1972, 1973a). The likelihood function for estimating the ability θ_i of the ith individual is given by

$$L_i(u_{i1}, u_{i2}, \ldots, u_{in}|\theta_i) = \prod_{j=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{(1-u_{ij})} \quad .$$

The maximum likelihood estimator of θ_i is sufficient, and efficient (Birnbaum, 1968, p. 455-459). Lord (1974a) has shown further that the estimates are consistent (Lord's proof of consistency, though valid for a different case, can be adapted to the two-parameter case readily). In addition, the likelihood function for the two-parameter ιogistic model possesses a unique maximum. However, with respect to the three-parameter logistic model, Samejima (1973a) has shown that the likelihood function may not have a unique maximum, if sample size is small and if the range of θ is unrestricted. Samejima (1973a) goes on further to show that the problem can be solved by considering the subdomain of the latent trait, θ, such that

$$\max\ (\theta_g^*) \leq \theta < \infty$$

where

$$\theta_g^* = b_g + (\log c_g)/\ 2\ Da_g.$$

In the subdomain, the likelihood function possesses a unique maximum, and hence the maximum likelihood estimators of θ exist with their usual properties.

Properties of maximum likelihood estimators

Let $\hat{\underline{\tau}}$ be the maximum likelihood estimate of the vector $\underline{\tau}$ obtained by maximizing the likelihood function L. Then, under general conditions (not satisfied, as we have seen, by the maximum likelihood estimators when item parameters and ability parameters are estimated simultaneously) the maximum likelihood estimator, $\hat{\underline{\tau}}$, is asymptotically consistent, unbiased, efficient, and a function of the sufficient statistic if a sufficient statistic exists. In addition, $\hat{\underline{\tau}}$ is asymptotically multivariate normally distributed with mean $\underline{\tau}$, and dispersion matrix $-\ \{E\ (\partial^2 \log L/\partial\hat{\underline{\tau}}'\partial\ \hat{\underline{\tau}}\ )\}^{-1}$. The expression $-\ E\ (\partial^2\ \log L/\partial\hat{\underline{\tau}}'\partial\hat{\underline{\tau}})$ is commonly known as the information matrix (Kendall and Stuart, 1973, p. 55), and is denoted by $I\ (\hat{\underline{\tau}})$. As pointed out earlier, the information matrix is the expected value of the Hessian at the maximum point of the likelihood function. The information function can be expressed in one of several ways, i.e.,

$$I\ (\hat{\underline{\tau}}) = -E\ (\partial^2 \log L/\partial \hat{\underline{\tau}}' \partial \hat{\underline{\tau}})$$

$$= E\ (\frac{\partial \log L}{\partial \hat{\underline{\tau}}} \frac{\partial \log L}{\partial \hat{\underline{\tau}}'}).$$

The last form is particularly suitable for evaluating the information matrix of complex likelihood functions.

The usefulness of the information matrix is evident. Since the estimates are multivariate normally distributed asymptotically, the inverse of the information matrix has along its diagonal the asymptotic variances of the estimates. It is then possible to construct confidence intervals for individual parameters and test hypotheses concerning the parameters jointly or individually.

The information function that is usually of interest is that of the estimates of the ability parameters, $I(\hat{\theta}_i)$. In particular, if $\hat{\theta}_i$ is the maximum likelihood estimate of θ, then

$$\hat{\theta}_i \sim N\ \{\theta_i\ ,\ 1/I(\hat{\theta}_i)\}$$

The inverse of the asymptotic variance, $I(\hat{\theta}_i)$, is given by

$$\begin{aligned} I\ (\hat{\theta}_i) &= -\ E\ (\partial^2 \log L/\partial\ \hat{\theta}_i^2) \\ &= E\ (\partial \log L/\partial\ \hat{\theta}_i)^2 \\ &= n\ E\ (\partial \log f\ (\theta_i)/\partial\ \hat{\theta}_i)^2 \end{aligned}$$

where

$$f(\theta_i) = P_{ij}^{\ u_{ij}}\ Q_{ij}^{\ (1-u_{ij})}$$

and $P_{ij}(\theta)$ is the item characteristic curve.

Expressions for the information function for the various models are given by Birnbaum (1968, p. 460-462). Thus, it is possible to obtain an estimate of the standard error associated with each ability estimate $\hat{\theta}_i$.

For details of an application, the reader is referred to Lord (1953a) where the confidence interval for an examinee's ability, θ_i, is constructed. We shall return to a detailed discussion and use of the information function in a later section.

Heuristic estimation procedures

The maximum likelihood estimates, as pointed out in the previous section, do have desirable properties, at least asymptotically. However, these procedures are costly and time consuming in some situations. When cost and time are of concern, heuristic estimation procedures (Urry, 1974; Jensema, 1976) that provide rough and ready estimates of the item parameters, may be employed.

In the case of dichotomously scored items, under the assumption that the ability is normally distributed with zero mean and unit variance, and that the item characteristic curve is the two-parameter normal ogive, Lord and Novic (1968, p. 377-378) have shown that the correlation ρ_g between the score on item g, u_g, and the underlying ability, θ, is given by

$$\rho_g = a_g/\{1+a_g^2\}^{\frac{1}{2}} .$$

They have also shown that the difficulty of item g for the group, π_g, is given by

$$\pi_g = \Phi(-\gamma_g)$$

where $\gamma_g = b_g \rho_g$, and $\Phi(-\gamma_g)$ is the area under the unit normal curve from $-\gamma_g$ to infinity. Since ρ_g is the correlation between the score on item g and the latent ability θ, ρ_g is given as the factor loading of the item on the common factor obtained by a factor analysis of the matrix of sample tetrachoric correlations. The item difficulty, π_g, is estimated by the proportion of examinees who answered item g correctly. Thus, once

ρ_g and π_g are determined, $\hat{a}_g$ and $\hat{b}_g$ can be obtained readily. Of course, the appropriateness of these estimates will depend on the assumptions made in the estimation procedure.

A further parameter, the guessing parameter, c_g, has to be estimated in the three-parameter latent trait model. Jensema (1976), following Lord (1968), suggests obtaining a proportion of the examinees passing an item at each of the lower item-excluded subtest scores, and using this as an estimate of c_g. Once a value for c_g is obtained, the method suggested in the preceding paragraphs can be employed to estimate a_g and b_g.

Although the above procedure is relatively simple to implement, there are several problems with the procedure. The estimates $\hat{a}_g$, $\hat{b}_g$, and $\hat{c}g$ obtained by this method do not have any known sampling properties. Secondly, factor analysis of a matrix of tetrachoric correlations presents theoretical problems. The matrix of sample tetrachoric correlations is not necessarily positive definite and hence, cannot, in the strict sense, be factor analyzed.

However, despite these problems, Jensema (1976) reports that the correlations between these estimates and the maximum likelihood estimates of the parameters are relatively high. Hence, these heuristic procedures can be taken to provide quick and cost-saving estimates of the item parameters when these issues are of major concern.

Bayesian estimation of parameters

When prior information (or belief) about a parameter is available, it is conceivable that incorporation of this information in the estimation procedure would increase the "accuracy" or the meaningfulness of the estimates. An example of this was encountered earlier, where in order to prevent the estimates of the item discrimination parameter from drifting out of bounds, it was necessary to impose limits on the range of values the parameter could take. Similarly, the distribution of ability, q (θ), or the prior information about θ, was incorporated into the unconditional estimation procedure. Despite these efforts, relatively little is known about the feasibility of applying Bayesian procedures for the estimation of parameters in latent trait models.

It may be instructive to review the logic of the Bayesian estimation procedure briefly (for a detailed account, the reader is referred to Novick and Jackson, 1974). Let T be a parameter of interest and x_1, x_2, ..., x_N, denote N values of observable random variable X whose probability density function $f\ (x|\tau)$ depends upon the value of the parameter $T = \tau$. Supposing further that the N observationsare independent, the joint probability of the observations, or the likelihood function, $L(x|\tau)$, is given by

$$L(x|\tau) = \prod_{i=1}^{N} f(x_i|\tau).$$

If prior information or belief about the parameter τ can be expressed as g(τ), where g(τ) is the probability density function of τ, then the *posterior* distribution of τ given the observation, $h(\tau|x_1, x_2, \ldots x_N)$ can be expressed as (Kendall and Stuart, 1973, p. 159)

$$h(\tau|x_1, x_2, \ldots, x_N) = k\ L(x|\tau)g(\tau)$$

where k is a constant of proportionality. The posterior distribution of τ is thus an expression of the investigator's revised belief about the parameter once the data are obtained.

The procedure for obtaining a Bayes estimator employing prior belief has been advocated by, among others, Lindley and Smith (1972) and Novick and Jackson (1974), and has been applied to latent trait models by Birnbaum (1969) and Owen (1975). This approach employs the "subjective" notion of probability as opposed to the classical, or, frequency theory of probability. A compromise between these two views of probability is obtained by employing the empirical Bayes procedure in which the prior distribution of the parameter is "estimated" from the data. This procedure which yields empirical Bayes estimators, is exemplified by the works of Lord (1971b), and Meredith and Kearns (1973).

Birnbaum (1969) obtained Bayes estimates for the ability parameters in the one- and two-parameter logistic models under the assumption that the item parameters are known. He chose, for mathematical tractability, the prior probability density function of θ_i to be the logistic density function, i.e.

$$g(\theta_i) = e^{-D\theta_i} / (1 + e^{-D\theta_i})^2$$

where D = 1.7 is a scaling factor. The likelihood function in this case is given by

$$L(u_{i1}, \ldots, u_{in} | \theta_i) = \prod_{g=1}^{n} P_g(\theta_i)^{u_g} Q_g^{1-u_g}(\theta_i)$$

where $P_g(\theta_i)$ is the item characteristic curve for the one- or two-parameter logistic model. The posterior density function of θ_i is then given as

$$h(\theta_i | u_{i1}, u_{i2}, \ldots, u_{in}) \propto L(u_{i1}, \ldots, u_{in} | \theta_i)\, g(\theta_i).$$

The Bayes estimator of θ_i, $\bar{\theta}_i$, is taken as the mean of the posterior distribution, i.e.

$$\bar{\theta} = E(\theta_i | u_{i1}, \ldots, u_{in})$$

$$= \int_{-\infty}^{\infty} \theta_i h(\theta_i | u_{i1}, u_{i2}, \ldots, u_{in}) d\theta_i \quad .$$

For a discussion and further details of the procedure, the reader is referred to Birnbaum (1969).

The procedure advocated by Birnbaum (1969) is not general enough to permit the estimation of item parameters and ability parameters simultaneously. In addition, there is no provision for incorporating available information about the "hyperparameters" that specify the prior distribution completely.

The procedure suggested by Lindley and Smith (1972) for the estimation of parameters in the general linear model can be applied to estimate the parameters in the latent trait models. The likelihood function of the observations for fixed θ_i, and item parameters a_g and b_g (for the two-parameter model) is expressed as

$$L(u_{11}, u_{12}, \ldots\ldots, u_{Nn} | \theta_1, \theta_2, \ldots, \theta_N; a_1, a_2, \ldots, a_n; b_1, b_2, \ldots, b_n).$$

In order to obtain the posterior distribution of the parameters $\underline{\theta}$, $\underline{a}$, and $\underline{b}$, it is necessary to specify prior distributions. We assume that our prior beliefs about a θ_i are no different than about any other θ_j, i.e., the prior information is "exchangeable" (Lindley and Smith, 1972, Novick and Jackson, 1974). This implies that the θ_i have the probability structure of a random sample from some common distribution. Thus, we can assume that $\theta_i \sim N(\mu_\theta, \phi_\theta)$. In turn, we assume that μ_0 and ϕ_0, the mean and variance of the prior distribution, are independent a priori and that the density function of μ_θ is $f(\mu_\theta)$ and that of ϕ_θ is $h(\phi_\theta)$.

The assumption that the prior information about $b_1, b_2, \ldots, b_n$ and $a_1, a_2, \ldots, a_n$ is exchangeable may appear to be implausible. However, it is not unreasonable to assume a distribution for the item difficulty parameters. Birnbaum (1968, p. 466) considers the case where the item difficulty parameters are distributed normally. Thus, we may assume that $b_i \sim N(\mu_b, \phi_b)$ and in turn assume that μ_b and ϕ_b are independently distributed with known prior distributions. Although it seems unreasonable to assume that the a_g's have the probability structure of a random sample from a common distribution, we may assume that the prior information on the a_g's are identical with density function $p(a_g)$. Thus, the posterior distribution of $\theta_1, \theta_2, \ldots, \theta_N$, $b_1, b_2, \ldots, b_n$, $a_1, a_2, \ldots, a_n$, $\mu_\theta, \mu_b, \phi_\theta, \phi_b$, given the observations is

$$P(\theta_1,\theta_2,\ldots,\theta_N, b_1,b_2,\ldots,b_n, a_1,a_2,\ldots,a_n,\mu_\theta,\phi_\theta,\mu_b,\phi_b | u_{11},u_{12},\ldots,u_{Nn})$$

$$\propto L(u_{11},u_{12},\ldots,u_{Nn}|\theta_1,\theta_2,\ldots,\theta_N,b_1, b_2,\ldots, b_n, a_1,a_2,\ldots,a_n)$$

$$\times \{ \prod_{i=1}^{N} \prod_{g=1}^{n} \phi_\theta^{-\frac{1}{2}} \phi_b^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\{\frac{1}{\phi_\theta}(\theta_i-\mu_\theta)^2 + \frac{1}{\phi_b}(b_g-\mu_b)^2\}\right] p(a_g)\} f(\mu_\theta) f(\mu_b) h(\phi_\theta) h(\phi_b) .$$

In this case μ_θ, ϕ_θ, μ_b, and ϕ_b are nuisance parameters and could be removed by integrating the posterior density function. The resulting posterior density function is only a function of the parameters $\theta_1, \theta_2, \ldots, \theta_N$, $b_1, \ldots, b_n$, $a_1, \ldots, a_n$. Joint modal estimates of these parameters can be then obtained by differentiating the posterior density function, setting these derivatives equal to zero, and solving the resulting "Lindley Equations." Alternatively, the θ's could be estimated independently of the a_g's and the b_g's by integrating with respect to these parameters. The joint modal estimates of $\theta_1, \theta_2, \ldots, \theta_N$ can be obtained by solving the resulting Lindley Equations, or alternatively, the marginal density function of say, θ_i, can be obtained by integrating out the other ability parameters. The mode or the mean of the marginal distribution of θ_i could be then taken as the Bayes estimate of θ_i. The same procedure may then be applied to each of the remaining parameters.

The procedure outlined above requires specification of prior distributions for the parameters θ, b, and a, and also for the hyperparameters μ_θ, μ_b, ϕ_θ, and ϕ_b. In the case of prior ignorance, we may take $f(\mu_\theta)$ and $f(\mu_b)$ to have uniform distributions. Prior ignorance on ϕ_θ and ϕ_b implies that $h(\phi_\theta) \propto \phi_\theta^{-1}$ and similarly for ϕ_b. At this point, specification of $p(a_g)$, the prior distribution of a_g, is unclear, but could be specified in terms of a rapidly decaying exponential function.

Alternatively, an empirical Bayes procedure could be used to estimate the parameters in the latent trait models. This approach requires the specification of prior distributions for the parameters, but the hyperparameters that specify the prior distributions are, in general, estimated from the data. Meredith and Kearns (1973) have applied this procedure to the Rasch model and obtained empirical Bayes estimates of the ability parameter by expressing the likelihood function in terms of the sufficient statistic, the total score of an examinee.

The Bayesian procedures discussed above are obviously more complex than the estimation procedures discussed in the preceding sections. In addition, Bayes procedures require specification of prior information on the parameters of interest and thus involve a subjective view of probability as opposed to the classical or frequency theory of probability. However, when applicable, Bayesian procedures yield more satisfactory solutions, in that improper solutions do not usually occur. Moreover, it is well known that the Bayes procedures, with the exception of the empirical Bayes procedure, are in general, *admissible*, in the sense that they minimize the expected loss (Meredith and Kearns 1973). Furthermore, Bayesian credibility intervals may be more meaningful than conventional confidence intervals. In addition, as demonstrated by Owen (1975) and Meredith and Kearns (1973), the Bayes estimates converge in probability to the true value with increasing sample size. Finally, the Bayes procedures have the potential of offering solutions to the estimation problems in the latent trait models when the sample size and the number of items are small. In these situations prior beliefs assume importance, while with increasing sample size they tend to lose their importance.

Despite these advantages, further investigation is necessary regarding the Bayesian procedures. The estimation procedure based on the approach of Lindley and Smith (1972), outlined earlier, has not yet been implemented and its usefulness has to be further documented. In particular, little is known about the families of priors that are appropriate, especially since natural conjugate priors are not available for the latent trait models of interest. Finally, the effect of specifying poor priors on the estimates has to be studied carefully. In conclusion we note that while Bayesian procedures hold the promise for solving the estimation problems in latent trait theory, considerable research is required before definitive statements can be made regarding the efficacy of these procedures.

Estimation in the nominal response, graded response, and continuous response models

As opposed to the dichotomous response model, in the nominal response model it is assumed that each of N examinees responds to n multiple-choice items of which the jth item has m_j response categories. In this case, the probability that an examinee of ability θ will respond to item j by choosing category k_j is given by

$$P_{jk_j}(\theta) = \exp\,[z_{jk_j}(\theta)] \;/\; \sum_{h=1}^{m_j} \exp\,[z_{jh}(\theta)] \quad .$$

Andersen (1972) chose $z_{jh}(\theta)$ to be of the form

$$z_{jh}(\theta) = b_{jh} + \theta,$$

Bock (1972), on the other hand, chose z_{jh} to be of the form

$$z_{jh}(\theta) = b_{jh} + a_{jh}\theta \;.$$

Since responses to the m_j-1 categories fix the response to the m_jth response category, we have the restrictions $\sum_h b_{jh} = 0$ for the one-parameter model, and $\sum_h b_{jh} = 0$, $\sum_h a_{jh} = 0$ for the two-parameter logistic model. The simplest way to incorporate this restriction is to take $b_{jm_j} = 0$ and $a_{jm_j} = 0$, or alternatively, reparameterize the model as indicated by Bock (1972).

Andersen (1972) obtained conditional estimates for his one-parameter nominal response model by maximizing the likelihood function, conditional on the sufficient statistic. As indicated earlier, these maximum likelihood estimates of parameters are consistent. Bock (1972) obtained unconditional estimates of the item parameters and also the conditional estimates of the item, as well as ability, parameters in the manner described in an earlier section.

The graded response model and its natural extension, the continuous response model were introduced and studied by Samejima (1969, 1972, 1973b, 1974). Although Samejima does not discuss the estimation of parameters in detail for these models, she derives important results concerning these estimates. She shows that, unlike in the dichotomous response case, both the normal ogive and the logistic models yield sufficient statistics for the ability parameter. In addition, she shows that the amount of information increases by shifting from dichotomous scoring to graded and continuous scoring. Hence, the graded and continuous response models offer advantages over the nominal and dichotomous response models in that the information available increases. Furthermore, the problem of estimating ability parameters in the graded and continuous response models appears to be solved. We may, however, expect difficulties when estimating item parameters and ability parameters simultaneously. It appears that these problems may be solved by employing the procedures discussed earlier, but further research is needed to establish this.

Testing Assumptions and Goodness of Fit of Latent Trait Models

Assumptions

How reasonable is the assumption of unidimensionality or (as has been shown to be equivalent) the assumption of local independence? Lumsden (1976) was particularly distressed that more researchers do not attend to this assumption. Testing the assumption of unidimensionality takes precedence over other goodness of fit tests of a latent trait model since, if the assumption of unidimensionality is untenable, the results of the other tests are more difficult to interpret. For example, if tests of goodness of fit of the model indicate that a particular latent trait model does not fit the data, and if unidimensionality was previously established, then at least this potential explanation of the misfit between the model and the data can be ruled out.

The simplest way to ascertain unidimensionality is to factor analyze the matrix of inter-item correlations. Existence of a single factor would imply unidimensionality. Lord (1968) reported that various researchers have factor analyzed matrices of tetrachoric item intercorrelations to determine if a set of test items measure more than a single factor. He noted that the residuals after extracting one factor were often near the size that one would expect from sampling fluctuations. For example, Coffman (1966) extracted 11 factors for the SAT Verbal Test but most of the variance could be accounted for by the first factor. On the other hand, Hambleton and Traub (1973) were less successful in locating unifactoral tests, but in the three aptitude tests that they studied, they did find a "dominant" first factor.

Another assumption of latent trait models concerns the particular choice of a mathematical form of the item characteristic curves to describe the test data. Since latent traits are not directly measurable, we find ourselves in a situation where it is quite difficult to separate the inappropriateness of a particular choice of mathematical form for item characteristic curves from violations of other assumptions of the model. One solution offered by Lord (1970a) is to compare item characteristic curves derived from a direct method (where the mathematical form of item characteristic curves does not have to be prespecified) with estimated item characteristic curves of the form specified by the user. The "closeness" of the two sets of item characteristic curves provides a basis for checking the appropriateness of the assumption. (Incidentally, when Lord attempted this comparison with SAT test data, he found close agreement between a direct method of item characteristic curve estimation and three-parameter logistic curves.)

A second possible test of the assumption is to check the "accuracy" of various predictions with the estimated item characteristic curves of specified form. Accurate predictions provide evidence of the suitability of the model for the particular data set and, of interest here, the assumption concerning the mathematical form of item characteristic curves. Of course, if the predictions are not good, pinpointing the problem could be difficult. Several researchers (for example, Hambleton and Traub, 1973; Ross, 1966) have attempted to study the appropriateness of different mathematical forms of item characteristic curves by using them, in a comparative way, to predict various test score characteristics. Hambleton and Traub (1973) obtained item parameters for one- and two-parameter logistic curves with three aptitude tests. Assuming a normal ability distribution and using test characteristic curves obtained from both the one- and two-parameter logistic curves, they were able to obtain predicted score distributions for each of the three aptitude tests. A χ^2 measure of goodness of fit was used to compare actual test score distributions with predicted test score distributions from each test model. The "relative" appropriateness of the two mathematical forms of item charácteristic curves was studied by comparing the χ^2 statistics. A likelihood ratio test for comparing the "relative" appropriateness of two mathematical forms of item characteristic curves will be discussed later in this section. In all three cases, substantially improved predictions were obtained with the two-parameter logistic curves. The Hambleton-Traub results also suggest, not surprisingly, that the two-parameter logistic model will provide the greatest improvements over the one-parameter logistic model when applied to data from short tests where the variability of discrimination parameters is substantial.

Goodness of Fit

Statistical tests of goodness of fit of the various latent trait models have been given by several authors (Andersen, 1973; Bock, 1972; Mead, 1976; Wright, Mead, and Draba, 1976; Wright and Panchapákesan, 1969). The procedure advocated by Wright and Panchapakesan (1969), for testing the fit of the Rasch model, essentially involves examining the quantity f_{ij} where f_{ij} represents the frequency of examinees at the ith ability level answering the jth item correctly. Then, the quantity y_{ij}, where

$$y_{ij} = \{f_{ij} - E(f_{ij})\}/\{\text{Var } f_{ij}\}^{1/2}$$

is distributed normally with zero mean and unit variance. Since f_{ij} has a binomial distribution with parameter p_{ij}, the probability of a correct response is given by $\theta^*_i/(\theta^*_i + b^*_j)$ for the Rasch model, and r_i, the number of examinees in the score group. Hence, $E(f_{ij}) = r_i p_{ij}$, and Var $(f_{ij}) = r_i p_{ij}(1-p_{ij})$. Thus a measure of the goodness of fit, χ^2, of the model can be defined as

$$\chi^2 = \sum_{i=1}^{n-1} \sum_{j=1}^{n} y_{ij}^2 .$$

The quantity, χ^2, defined above has the χ^2 distribution with degrees of freedom (n-1) (n-2) since the total number of observations in the matrix $F = \{f_{ij}\}$ is n(n-1), and the number of parameters estimated is 2(n-1). Wright and Panchapakesan (1969) also defined goodness of fit measure for individual items as

$$\chi_j^2 = \sum_{i=1}^{n-1} y_{ij}^2$$

where χ_j^2 is distributed as χ^2 with degrees of freedom, (n-2). This general method of determining the goodness of fit of overall test data can be extended to the two- and three-parameter latent trait models. The reader is referred to Hambleton and Traub (1973) for an example of a test of goodness of fit applied to two- and three-parameter logistic models.

There are several problems associated with the chi-square tests of fit discussed above. The χ^2 test has dubious validity where any one of the $E(f_{ij})$ terms, $\underline{i} = 1, 2, ..., \underline{n} - 1$; $\underline{j} = 1, 2, ..., \underline{n}$, have values less than one. This follows from the fact that when any of the $E(f_{ij})$ terms are less than one, the deviates y_{ij}, $i = 1, 2, \ldots, n-1$; $j = 1, 2, \ldots, n$, are not normally distributed and a χ^2 distribution is obtained only by summing the squares of normal deviates. Another problem encountered in using the χ^2 test is that it is sensitive to sample size. If enough observations are taken, the null hypothesis that the model fits the data will always be rejected using the χ^2 test. However it should be pointed out that this is an inherent weakness of all statistical tests.

Alternately, Wright, Mead, and Draba (1976) and Mead (1976) have suggested a method of test of fit for the one parameter model which involves conducting an analysis of variance on the variation remaining in the data after removing the effect of the fitted model. This procedure allows not only a determination of the general fit of the data to the model but also enables the investigator to pin-point guessing as the major factor contributing to the misfit. This procedure for testing goodness of fit of the one parameter model involves computing residuals in the data after removing the effect of the fitted model. These residuals are plotted against $(\theta_i - b_g)$. According to the model, the plot should be represented by a horizontal line through the origin. For guessing, the residuals follow the horizontal line until the guessing becomes important. When this happens the residuals are positive since the person is doing better than expected and in that region have a negative trend. If practice or speed is involved, the items which are affected display negative residuals with a negative trend line over the entire range of ability. Bias for a particular group may be detected by plotting the residuals separately for the two groups. It is generally found that the residuals have a negative trend for the unfavored group and a positive trend for the favored group.

Mead (1976) concludes by saying "All of the disturbances considered represent some form of multidimensionality; they would violate any model that assumes unidimensionality. Since the effect of the disturbances often appears as a change in the slope of the item characteristic curve, any model which includes item discrimination as a parameter would appear to fit the data".

When maximum likelihood estimates of the parameters are obtained, likelihood ratio tests can be obtained for hypotheses of interest. Likelihood ratio tests involve evaluating the ratio, λ, of the maximum values of the likelihood function under the hypothesis of interest to the maximum value of the likelihood function under the alternate hypothesis. If the number of observations is large, $-2 \log \lambda$ is known to have a chi-square distribution with degrees of freedom given by the difference in the number of parameters estimated under the alternate and null hypotheses. An advantage possessed by likelihood ratio tests over the other tests discussed earlier is apparent. Employing the likelihood ratio criterion, it is possible to assess the fit of a particular latent trait model against an alternative.

Andersen (1973) and Bock and Liebermann (1970) have obtained likelihood ratio tests for assessing the fit of the Rasch model and the two-parameter normal ogive model respectively. Andersen (1973) obtains a conditional likelihood ratio test for the Rasch model based on the within score group estimates and the overall estimates of item difficulties. He shows further that -2 times the logarithm of this ratio is distributed as χ^2 with degrees of freedom, (n-1) (n-2).

Based on the work of Bock and Liebermann (1970), likelihood ratio tests can be obtained for testing the fit of the two-parameter normal ogive model. It should be pointed out that these authors have obtained both conditional and unconditional estimates of the parameters. For the likelihood ratio test, it would be more appropriate if the unconditional model is used since with this model ability parameters are not estimated, and hence the likelihood ratio criterion can be expected to have the chi-square distribution. This procedure can be extended to compare the fits of one model against another (Andersen, 1973).

The major problem with this approach is that the test criteria are distributed as chi-square only asymptotically. When large samples are used to accommodate this fact, the chi-square value may become significant owing to the large sample size! Further investigation is clearly needed in this area in order to resolve this dilemma.

Test and Item Information and Efficiency Curves

The precision with which examinee ability can be estimated is of considerable importance. When the maximum likelihood estimate of ability is obtained, the precision of the ability estimate can be conveniently expressed in terms of the information function, referred to here as the test information curve. The standard error of maximum likelihood estimates is given by the square root of the inverse of the information curve. Birnbaum (1968) defined information as a quantity inversely proportional to the squared length of the confidence interval around an examinee's ability. Thus, 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 test information curve is a function of ability, 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

$$I_y(\theta) = \frac{(\sum_{g=1}^{n} w_g P_g')^2}{\sum_{g=1}^{n} w_g^2 P_g Q_g} \quad . \qquad [27]$$

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

$$y = \sum_{g=1}^{n} w_g u_g \quad ; \qquad [28]$$

The variable u_g takes on values 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, $g=1, 2, \ldots, n$.

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

$$I(\theta) = \sum_{g=1}^{n} \left(\frac{P_g{}'^2}{P_g Q_g}\right) \quad . \qquad [29]$$

The maximum value of the information curve of a given scoring formula is obtained when the scoring weights, w_g, are given by

$$w_g = \frac{P_g{}'}{P_g Q_g} \quad . \qquad [30]$$

In order to obtain the test information curve for a particular set of test items, and consequently minimize the widths of confidence bands about examinee ability, it has been shown that the scoring weights for the one-, two-, and three-parameter logistic test models should be chosen to be 1, Da_g, and $\frac{Da_g\ (P_g-c_g)}{(1-c_g)\ P_g}$, respectively (Lord and Novick, 1968). (Test information curves and the best scoring weights for several other latent trait models are given by Samejima [1969, 1972].) It should be noticed that only for the three-parameter model are the scoring weights a function of ability level. The scoring system in the three-parameter model has the effect of reducing the weight assigned to correct answers on items where the values of the lower asymptote (c_g) of the item characteristic curves are large. It can be seen that the weights for such items are smaller for low-ability examinees than for either middle- or high-ability examinees. These weights reflect the fact that low-ability examinees are most likely to be answering the items by guessing. For high ability examinees, the optimum scoring weights of items approach the quantity, $Da_g(g=1, 2, \ldots, n)$.

The quantity $P_g'^2/P_gQ_g$ in Equation [29] is the contribution of item g to the information curve of the test. For this reason it is called the item information curve.

Item information curves 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 item scores at each ability level θ. The steeper 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.

Figure 2 shows item information curves for five verbal test items, and the test information curve for a test composed of these items.[1] The logistic parameters of the five items are shown below:

Item	b_g	a_g	c_g
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

[1] We are grateful to Frederic Lord for allowing us to reproduce this figure from (Lord, 1968).

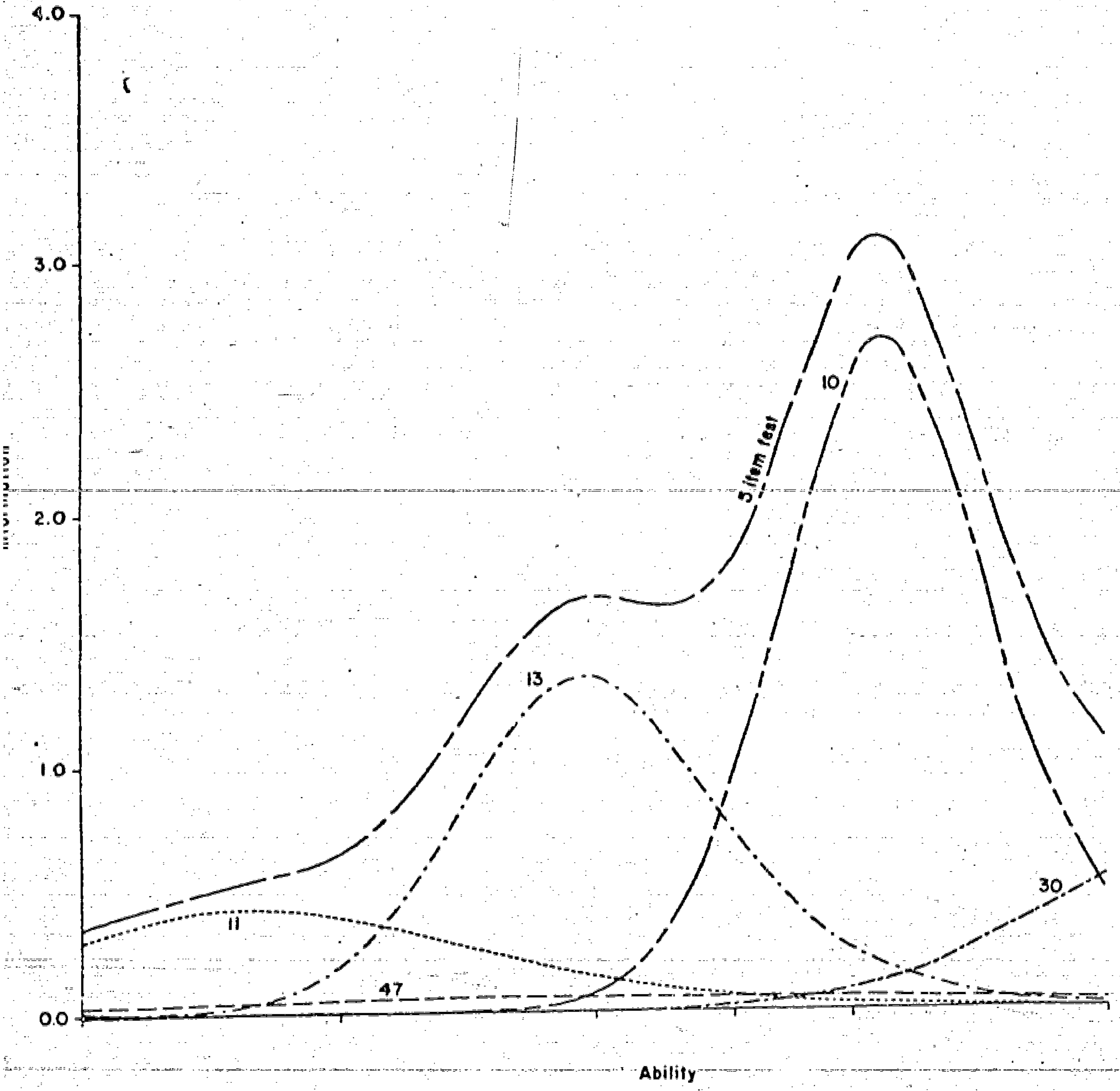


Figure 2. Information curves estimated for five items and a five-item test. The items are from the verbal section of the SAT. This figure is reproduced by permission from Lord (1968).

The information curve for the test composed of the items is obtained by summing the ordinates of the five item information curves. This curve reveals that the five-item test provides the most information for high ability students. This means that abilities are more precisely estimated at the high end of the ability continuum. The height of the test information curve is misleading though because of its dependence on the metric of the ability scale (Lord, 1975d). At a particular ability level, θ, the item information curve is given by P'^2/PQ. But the slope of the item characteristic curve P', is a function of the ability scale. If the ability scale is compressed in the region of θ, P' will increase; and, when the scale is stretched in the region of θ, P' will decrease. A good example of the effect that a monotonic transformation on the ability scale has on the test information curve is seen in Lord (1975d). The effect is substantial.

From Equation [29] it is clear that items contribute independently to the test information curve. Birnbaum (1968) has also shown that with his three-parameter model, an item provides maximum information at an ability level θ, where

$$\theta = b_g + \frac{1}{1.7\ a_g} \log_e .5\ (1 + \sqrt{1+8c_g}) \quad . \qquad [31]$$

If guessing is minimal, then $c_g = 0$, and $\theta = b_g$. When $c_g > 0$, the point of maximum information is shifted to the right of the item difficulty value, b_g.

If non-optimal scoring weights are used with a particular logistic test model, the information curve derived from Equation [27] will be lower, at all ability levels, than one that would result from the use of optimal weights. Birnbaum (1968) used the term efficiency to refer to the information loss due to the use of less than optimal scoring weights. Efficiency is studied by calculating the ratio of the values of the information curve of a given scoring formula a'd the test information curve at each ability level (Hambleton and Traub, 1971). Hambleton and Traub (1971) found that when there was no guessing (i.e., c_g=0, g=1, 2, ..., n), the efficiency of unit scoring weights was quite high (over 85%) for typical levels of variation in the item discrimination parameters (.20 to 1.00, which very roughly translates into a range of biserial correlations from .20 to .70). When guessing is introduced, the situation changes dramatically. For low ability examinees, the efficiency of unit scoring weights dropped to the 60%-70% range. The authors concluded that when a test is being used to estimate ability across a broad range of the ability scale and when guessing is a factor in test performance, the scoring system of the three-parameter logistic model is to be preferred. Unit scoring weights lead to efficient estimates of ability when there is little or no guessing and when the range of discrimination parameters is not too wide.

Lord (1968) investigated the efficiency of unit scoring weights on the verbal section of the SAT. Under the assumption that the three-parameter model was the correct test model to explain the data, he found that the efficiency of unit-scoring weights varied from 55% at the lowest ability level to a maximum of 90% at the highest ability level. Using unit scoring weights was equivalent to discarding about 45% of the test items for the low-ability examinees!

On other occasions, one may be interested in comparing the relative efficiency with which two different tests measure the same ability at various points on the ability scale. It may also be of interest to know the relative efficiency of two different scoring methods at various ability levels. This can be determined by calculating the ratio of the values of two test information curves at each ability level. (It should be mentioned that the notion of relative efficiency is an important one in assessing the merits of a test for measuring ability along an ability continuum but Lord [1974c, 1974d] has also produced a way of studying relative efficiency without introducing the concepts of latent trait theory).

It has been shown by Birnbaum (1968) that relative efficiency is directly proportional to the test length of a "base-line" test. That is, if relative efficiency = 1.50 at a particular ability level, then it would take 1½ times as many items in the baseline test to yield the discriminating power at that ability level, provided by the other test under consideration.

Applications of Latent Trait Models

In this section we will consider several applications of latent trait models.

Differential Weighting of Response Alternatives

It is a common belief among test developers that it ought to be possible to construct alternatives for multiple-choice test items that differ in their degree of correctness. An examinee's test score could then be based on the degree of correctness of his or her response alternative selections, instead of simply the number of correct answers, possibly corrected for guessing. However, with few exceptions, the results of differential weighting of response alternatives have been disappointing Wang & Stanley, 1970). Despite the intuitive beliefs of test developers and researchers, from past research it would appear that differential weighting of response alternatives has no consistently positive effect on the reliability and validity of the derived test scores. However, our view is that using correlation coefficients to study the merits of any new scoring system is less than ideal. This is because correlation coefficients will not reveal any improvements in the estimation of ability at different regions of the ability scale. A concern for the precision of measurement at different ability levels is important. There is reason to believe that the largest gains in precision of measurement to be derived from a scoring system that incorporates scoring weights for the response alternatives will occur with low ability examinees. High ability examinees make relatively few errors on their test papers and therefore would make little use of differentially weighted incorrect response alternatives. The problem with using a group statistic, like the correlation coefficient, to reflect the improvements of a new scoring system is that any gains at the low end of the ability continuum will be "washed out" when combined with the lack of gain in information at other places on the ability continuum. One way of evaluating a test scoring method is in terms of the precision with which it estimates an examinee's ability: The more precise the estimate, the more information the test scoring method provides. Birnbaum's concept of information introduced earlier provides a much better criterion than do correlation coefficients for judging the merits of new scoring methods.

Thissen (1976) applied the nominal response model to a set of data from Raven's Progressive Matrices Test (a test where the options to each item can be logically ordered according to their degree of correctness). The model provides a measure of the precision of ability estimation (Birnbaum's "information") at each ability level. Thissen's results were clear and impressive. The nominal response model produced substantial improvements in the precision of ability estimation in the lower half of the ability range. Gains in information ranged from 1/3 more to nearly twice the information derived from 0-1 scoring with the logistic test model. According to Bock (1972), most of the new information to be derived from weighted response scoring comes from distinguishing between examinees who choose plausible or partly correct answers from those who omit the items.

In a study of vocabulary test items with the nominal response model, Bock (1972) found that below median ability, there was 1½ to 2 times more information derived from the nominal response model over the usual 0-1 test scoring method. In terms of test length, the scoring system associated with the nominal response model had, for about one-half of the examinee population, produced improvements in precision of ability estimation equal to the precision that could be obtained by a binary-scored test 1½ to 2 times longer than the original one with the new method of scoring. Also, encouraging was that the "curve" for each response alternative (estimated empirically) was psychologically interpretable. The Thissen and Bock studies should encourage other researchers to go back and reanalyze their data using the nominal response model and the measure of "information" provided by the logistic latent trait models. The Thissen and Bock studies indicate that there is "information" that can be recovered from incorrect examinee responses to a set of test items and provide interesting applications of test information curves to compare different test scoring methods.

Criterion-Referenced Testing

Latent trait models provide an excellent underpinning for a theory and practice of criterion-referenced testing. Much has been written on the topic of criterion-referenced testing, but the area is suffering because of a great many disconnected contributions, confusion over many basic problems such as test development and test score use, and the existence of unique problems such as the establishment of cutting scores (Hambleton & Novick, 1973).

A criterion-referenced test is constructed by sampling items from a well-defined domain of items measuring an instructional objective

(Millman, 1974). (Typically, a criterion-referenced test will include sets of test items measuring several instructional objectives. When several objectives are measured, the steps described below are repeated for each set of items measuring a single objective.)

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) could 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 (Hambleton, 1977).

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 can not 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 would 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.

Hambleton (1977) conducted an extensive study of criterion-referenced test designs in various testing situations and has reported substantial gains in test efficiency when the proper test design for a particular group of examinees is used.

Test Development

In this section we will attempt to describe a few of the areas of test development to which latent trait theory has been applied and shown to have decided advantages over standard test construction technology. It can be anticipated that as more is discovered about the properties of the latent trait models and as more psychometricians begin to use these models in the test development process, greater insight into the process will accrue.

Latent trait theory offers two advantages to the psychometrician interested in developing tests: (1) invariant item parameters that facilitate the test development process as well as make possible the development of tests for a variety of applications, and (2) item characteristic curves that provide valuable insights into how examinees perform on specific test items.

The first step in the test development process is the determination of test specifications. One of these specifications is the type of test item to be employed. The worker using latent trait theory has two options open to him/her. Either items can be developed to fit a specific test model or a test model can be chosen to "fit" the derived test data. For example, one may select the three-parameter logistic test model if the items are of the multiple-choice type. However, if he/she felt strongly that the test should be developed using the one parameter logistic model, he/she would include in the test specifications that the items be constructed to minimize guessing and to have equal discriminating parameters.

After the test specification process is completed, the actual construction of the items is generally the next step. In many instances previously constructed items may exist that are appropriate for test usage.

Suppose that an appropriate pool of pretested items does exist. If these items were characterized by classical test theory parameters describing item difficulty and item discrimination , the usefulness of the item statistics in test development would depend on the match between the characteristics of the pretest sample of examinees and the population of examinees in which the test will be used. Another shortcoming is that the size of item discrimination indices depends both on the number, and the particular items included in the pretest. When items are placed in a test which has test items different from those in the pretest, the usefulness of the discrimination indices is unknown. Because of the invariant properties of the latent trait item parameters this problem is circumvented.

How does one select items from an existing item pool in order to construct a test that meets a set of previously determined specifications? If standard test development technology is employed, there are a series of calculations that can be carried out to predict the mean, standard deviation, and test reliability (Lord & Novick, 1968). Input data for the calculations are the pretest item statistics.

Lord (1977b) outlined a method for predicting the mean, squared standard error of measurement and the test reliability based on any set of items characterized by latent trait theory parameters. The procedure involves specifying the ability level of the group for which the test is intended. The following expressions can then be used to determine the test statistics of interest:

$$(1)\quad \hat{\mu}_x = 1/N \sum_{a=1}^{N} \sum_{g=1}^{n} P_{ga}$$

$$(2)\quad \hat{\sigma}^2_{x|t} = 1/N \sum_{a=1}^{N} \sum_{g=1}^{n} P_{ga}\, Q_{ga}$$

$$(3)\quad \hat{\rho}_{xx'} = 1 - \hat{\sigma}^2_{x|t} / \hat{\sigma}^2_x \quad .$$

The predicted score variance ($\hat{\sigma}^2_x$) can be computed from the predicted test score distribution (Lord, 1977b).

Lord (1977b) concluded by saying that ". . .if we have a pool of pretested items all measuring the same trait or ability, we can predict the mean, variance, reliability and raw-score frequency distribution of any test constructed from these items once we know the ability levels in the group to be tested." When the shape of the ability distribution for a population of examinees can be specified, Lord (1953a) has shown how to use latent trait parameters to select items so as to produce desired test score distributions.

To summarize, when a psychometrician is selecting items characterized by classical test theory parameters to construct a test, he/she is forced to use a heuristic process that depends a great deal on the ability to estimate, from previous experience, the average item test correlation and also on the similarity of the pretest group and the group of interest. When using latent trait theory, only knowledge of the ability distribution of the group of examiness of interest is necessary to make accurate predictions of the test statistics.

Test information curves may also be used as a means of selecting items from a previously established pool of items characterized by latent trait theory parameters. The useful feature is that the contribution of each item to the test information curve can be determined without knowledge of the other items in the test. In conventional testing technology, 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 (1977b) discussed Birnbaum's (1968) procedure for building a new test. This procedure operates on a pool of calibrated items (so that an item information curve is available for each item). The procedure outlined by Lord is:

1. Decide on the shape of the desired test information curve. Lord (1977b) calls this the <u>target information curve</u>.
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 curves approximates the target information curve to a satisfactory degree.

It is obvious that the use of item information curves in the manner described above will allow the test developer to produce a test that will very precisely fulfill any set of desired test specifications.

Latent trait models not only allow the test developer to examine the contribution of individual items to a test information curve, but they also allow for the comparison of test information curves. It is possible for a psychometrician to form different combinations of items (tentative tests) in the initial stages of test development and compare the information curves of different sets of items at specific ability levels, thus allowing him/her to choose the set of items most suited for the purpose of the test. Marco (1977) used this technique to study the effect of lowering the difficulty of the Scholastic Aptitude Test. Item parameters (b_g, c_g, a_g) were determined on item data from about 3,000 students who took the mathematical part of the SAT in December, 1970 and the verbal part of the SAT on January, 1971. He then selected items to form four tests:

1. a test composed mostly of moderately difficult and easy items;
2. a middle difficulty test having a bimodal distribution of item difficulties;
3. a middle difficulty test with no easy or difficult items, and
4. a very easy test composed of all easy items.

Examination of the four test information curves showed clearly that if the test was made easier, discrimination in the upper or middle part of the ability range suffered.

To summarize, the test information curves obtained from tests developed using latent trait models make it possible to obtain some indication of the probable results of combining various subsets of items and also allow for comparisons among these subsets. No such feature exists for tests developed by conventional test construction methods. Thus, much of the combining of test items or altering of existing tests with conventional procedures must be done on an intuitive basis.

Once the final test forms are assembled, the next step in the test development process is usually test norming. In conventional testing technology, this is an expensive and time consuming process involving testing large samples of examinees similar to the population the test is intended for. Because latent trait models provide ability estimates that are independent of the items selected for administration, the norming process can be simplified considerably. It is not necessary for all individuals to take all of the test items. The test can be broken up into subtests with different groups of students taking different subsets of items. A successful application of this type of norming was made to the Key Math Diagnostic Arithmetic Test published by American Guidance Service.

Tailored Testing

Since the beginning of formal testing some 60 years ago, almost all testing has been done in a conventional fashion; that is, a group of individuals all take the same test. Since these individuals will vary in terms of the ability that is being measured by the test, some will find the test too difficult and others too easy. Those who find the test too difficult may experience frustration and negative reactions, while those who find the test too easy will not be sufficiently motivated to put forth maximum effort. In short, the test will do a good job of measuring for those individuals whose ability is at or near the median ability of the test. For such individuals, the difficulty level will be such that they will answer half the questions correctly and half incorrectly. A logical extension of this line of reasoning dictates that the test would measure maximally the ability of all individuals in the group if it presented questions to each individual that that individual could answer correctly half the time. This, of course, is not possible using one test.

In tailored testing, an attempt is made to "tailor" the difficulties of the test items to the ability of the examinee being measured. This demands the existence of a large pool of items whose statistical characteristics are known so that suitable items may be drawn. The procedure does not lend itself easily to paper and pencil testing situations, and hence the tailoring process is typically done by computer (exceptions to this rule are presented in the work of Lord [1971c, 1971d]). According to Lord (1974b), a computer must be programmed to do the following in order to tailor a test to an examinee:

1. Predict from the examinee's previous responses how the examinee would respond to various test items not yet administered.

2. Make effective use of this knowledge in picking the test item to be administered next.

3. Assign at the end of testing a numerical score that somehow represents the ability of the examinee tested.

Tailoring a test to examinees will circumvent the psychological problems mentioned earlier. Also, from a psychometric point of view, tailored testing can insure that the standard error of measurement will be the same throughout the ability continuum. This is not true of conventional tests where the standard error tends to enlarge for individuals at the extremes of the ability continuum.

(a) Classical Testing Theory and Tailored Testing

Early work on tailored testing, making use of classical test theory, tended to focus on concerns somewhat removed from the notion of ability estimation for an individual. Because of this different focus, classical methods functioned adequately. These studies (for example, Cleary, Linn, & Rock, 1968; Linn, Rock, & Cleary, 1972) focused on two areas: Allocation of examinees to extreme ability groups and the capacity of the tailored test to reproduce, using fewer test items, the rank ordering of examinees supplied by the conventional group test. The results of these studies tended to support the use of tailoring strategies, and the sorts of questions addressed allowed the use of traditional item indices.

The use of traditional item indices no longer suffices at the individual examinee level when the problem of interest is ability estimation. Here, based upon the set of test items an examinee encounters, we want to make an inference as to expected performance on a large set of questions like those encountered (Lord, 1974b). This expected performance is the ability of the examinee measured by the test items. Since we are concerned now with ability estimation on a single individual, this precludes the use of traditional item indices in selecting items, because these statistics are based upon a particular norm group. A set of calibrated items that is free from the norm or calibrating group is necessary.

Tailoring test items to an examinee dictates that different examinees take different test items. What is needed are examinee ability estimates that are independent of the particular choice of test items, if there is interest in comparing one examinee with another. The solution to this problem and the one mentioned previously, is provided by latent trait theory. Classical methods are of no value here.

(b) Latent Trait Theory and Tailored Testing

In order to perform the three tasks discussed by Lord, it is necessary to introduce the notion of item characteristic curves. This will allow us to predict how an examinee will perform on a new item, even if the item has a different difficulty level from the one previously responded to. The two- and three-parameter logistic curves have most often been selected as the mathematical formsof item characteristic curves used in tailored testing research.

(c) Tailored Testing Strategies

Research done on tailored testing, whether based upon latent trait theory or classical theory, has been built upon the following rule: If an examinee answers an item correctly, the next item should be more difficult; if an examinee answers incorrectly, the next item should be easier. Based upon this general rule, certain branching strategies have been devised.

These strategies can be broken down into two-stage strategies and multi-stage strategies. The multi-stage strategies are either of the fixed branching variety or the variable branching variety.

In the two stage procedure, all examinees take a routing test and based upon scores on this test, are directed to one of a number of measurement tests lying at various points along the ability continuum. Ability estimates are then arrived at through a suitable combination of scores from the routing test and the measurement test, which is usually peaked at a particular difficulty level. Lord (1971a) uses a maximum likelihood procedure that combines ability estimates for both the routing and measurement test in a fashion such that each estimate is weighted inversely by its estimated variance. Other combinations would also appear suitable.

Whereas the two-stage strategy requires only one branching solution, from the routing to the measurement test, multi-stage strategies involve a branching decision after the examinee responds to each item. If the same item structure is used for all individuals, but each individual can move through the structure in a unique way, then it is called a fixed branching model. Considering how much item difficulty should vary from item to item leads to involvement with constant step size structures (usually represented as pyramids) or decreasing step size pyramids. If guessing should become a consideration, then a possible solution would be to make step size in the positive direction less than that in the negative direction (Lord, 1970b).

For these multi-stage fixed branching models, all examinees start at an item of median difficulty on the continuum ($b_1 = 0$) and based upon a correct or an incorrect response, start to pass through a set of items that have been arranged on the basis of item difficulty. After having completed a fixed set of items, either of two scores are used to give an estimate of ability. One score is the difficulty of the item that would have been administered to the examinee after the nth (last) item. The other score is the average of the item difficulties, excluding the first item that everyone takes, but including the hypothetical n+1st item. Lord (1971a, 1971b, 1974b) has demonstrated that different scores should be used to estimate ability depending upon the strategy used. For co stant step size procedures (up and down methods), average difficulty score is preferred, while for variable step-size procedures (Robbins-Monro methods), the final difficulty score should be used.

The variable branching strategies are multi-stage strategies that do not operate with a fixed item structure. Rather, at each stage of the process, an item in the established item pool is selected for a certain examinee in a fashion such that the item will maximally educe the uncertainty of the examinee's ability estimate, if administered. After administration of the item, the ability estimate is either recomputed using Bayes Theorem (Owen, 1975), or recalculated using the maximum likelihood procedure. A normal prior on ability is assumed for the Bayesian method and the administration of items is terminated when $\sigma_m^2 <$ assigned value, where σ_m^2 is the posterior variance of the ability estimate after m items have been administered. For the maximum likelihood procedure, item administration ceases at a set number or when the standard error of the estimate is < a prescribed value for the last item administered.

(d) Studies Using Latent Trait Theory

As discussed in the previous section, there are a number of ways in which examinees can be presented with items tailored to their ability, and there are also a number of ways of computing scores to estimate ability, based upon item difficulties. What is also needed is a means of evaluating results obtained from various procedures. The mechanism for evaluation should not be based on group statistics such as correlation coefficients because the crux of the situation is to determine the accuracy with which we can measure ability for a single examinee. Most of the studies on tailored testing to date have made use of test information curves.

Lord (1971a) compared the test information curves obtained for various two-step procedures with a test information curve provided by a conventional peaked test (which he calls the standard test). His conventional test provided maximum information for scores at median ability level of the continuum (b=0), and decreasing information for scores deviant from median ability. Specific results of the study and others that he did (for example, Lord, 1970b, 1971b) cannot be summarized briefly because of the multitude of test designs and strategies that he studied. What is clear is this: The tailored procedures provide more information at the extremes of the ability distribution than does the standard test, and provide adequate information at the median difficulty and ability level (b=0), where the standard test cannot be surpassed.

Studies using the variable branching models will not be discussed in this paper. This is because it is very difficult to compare the results from these strategies among themselves, let alone with the fixed branching models. Readers are referred to Owen (1975) and Wood (1973, 1976a). Weiss (1974, 1976) and Vale and Weiss (1974), in their reviews of the strategies and relevant studies, also summarize some of the results of these procedures.

(e) Final Comments

The work by Lord and others in introducing latent trait models to explain or predict examinee performance in individualized testing situations represents one of the most successful of the applications of latent trait theory to date. Of course, much work remains to be done. For example,

1. It is unclear as to which of the various scoring methods, be it final difficulty score, average difficulty score, or any of the other possibilities, gives the best statistical approximation of ability. This is especially a problem when the number of test items administered is small.

2. The present models do not deal well with the effects of guessing. Since tailoring strategies minimize the number of items too difficult for an examinee, guessing should be reduced and any guessing that goes on probably can't be considered random. What is needed is an investigation of the exact effects of guessing on tailoring strategies for ability estimation.

The above list of problems and/or research areas are not meant to be all-inclusive. Wood (1973), Green (1970), Lord (1977a), and Weiss (1974) all offer further suggestions for research.

Item Banking

Interest in individualized instruction and testing, has brought to light the need for item banking (Choppin, 1976; Wood, 1976b). An item bank is a collection of test items, "stored" with known item characteristics and made available to test constructors. According to the intended purpose of the test, items with the desired characteristics can be drawn from the bank and used to construct a test with known properties.

Unfortunately, classical item statistics (item difficulty and discrimination) are of limited value for describing the test items in the bank because they are dependent on the group of examinees from which they came. On the other hand, latent trait item parameters do not have this limitation and therefore are more useful for describing test items in the bank.

A practical problem facing many test constructors is that of building, over a period of several years, a pool of items to be used in constructing test forms. Because of the time span, newly written items will need to be pretested on groups of examinees different from groups used to pretest other items in the pool. Because of the invariance property of the latent trait item parameters, even though two pretest groups may be quite dissimilar in ability, there are few problems in obtaining item parameters that are comparable across these groups. Let us assume that we are interested in describing items by the two item parameters in the two-parameter logistic test model. The one serious problem 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 items 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 common item parameters in the two groups, a prediction equation can then be used to predict item parameters for the new items had they been administered to the first group. In this way, all item parameters can be equated to a common group of examinees and corresponding ability scale. No such linear relationship exists between the classical model parameters.

Item Bias

The notion that certain items in a test may be biased toward certain minority groups is becoming a matter of concern for the testing community. The concern for test bias and therefore item bias essentially stems from litigation involving the use of tests to classify minorities for employment and educational opportunities. The problem here is properly called test fairness, but in order for a test to be fair (in usage), it is necessary but not sufficient that the items be unbiased. Test bias refers to the psychometric properties of a set of test items or scores; test fairness is concerned with the way the test is used in a particular situation. Thus, it would seem that a first step in investigating how tests are being or not being used fairly with minorities is to investigate item bias.

Investigations of item bias using classical test theory have not been successful. One reason for this has been offered by Pine (1976). Bias in testing is caused by the inability of tests to consider individual difference variables, such as motivation and ethnic background. Investigations of these variables using classical test theory will further perpetrate the problem; namely that we are using a group based approach, whether in the test or in the bias study, to try to investigate individual difference variables. We create a situation of bias and then try to use the mechanism that created the situation in the first place to investigate it.

What is item bias and why have traditional explanations for item bias led to procedures of minimal usage? The most extreme stance on item bias is that a test is biased to the extent that the means of the two populations considered are different. The problem here is that other variables besides item bias contribute to these mean differences.

As Hunter (1975) says, it is not so much that the test is biased, but that there is bias in the learning environments that help determine the test score. The notion of matching will not help; it would be impossible to list all relevant variables upon which to match. Noteworthy is that, from a latent trait point of view, this lack of educational equality of experience can be viewed as a problem of dimensionality. Experiences, that one group has had benefit of, expand the dimensionality of the underlying structure for that group in comparison to the other.

Taking the mean difference notion one step further doesn't help. If we suppose that we have a perfect unidimensional test without bias, then the difference between the means of groups should be consistent over items. There would be no group by item interaction. If in an analysis of variance, a group by item interaction should prove to be significant, it has been advanced that this fact is a demonstration that the items are biased. However, Hunter (1975) has clearly pointed out that a perfectly unbiased test can show such interaction. Items of varying difficulty demonstrate an item by group interaction. Thus, it would seem that dealing with item difficulties would be the next step, but there are problems with using the classical definition of item difficulty as an indicant of item bias.

A classical definition of item difficulty would refer to the proportion of correct answers given to an item. If the item difficulty were the same for both groups, it has been advanced that this would be a demonstration that the item was unbiased. Lord (1976) has noted that one could plot these proportions for items on a test for both groups, and fit the resulting scatterplot with a straight line. Departure from linearity would then seem to be a good indicant of test, and one step further, individual item bias. Lord clearly points out that the failure of points to fall on a straight line does not mean that there is test and item bias. He states the following reasons for his stand:

1. There is no good reason for the points to lie on a straight line in the first place. If one group consistently outperforms the other, the relationship must be curved. Further, while straightening the line of relationship by using the inverse normal transformation (and perhaps further transforming to Δ values) does straighten the line, there are still further causes for problems.

2. If the questions can be answered by guessing, even using the inverse normal transformation is not going to assure that the points will lie on a straight line unless the groups performed equally well on the test.

3. If guessing weren't a problem, the discrimination index of an item would be. More discriminating items would produce more of a difference between groups than less discriminating items. Items of the same discrimination would lie along the same line, but there is no assurance, without building equal discrimination into the situation or model, that this is the case.

Thus, while we would want no other variables to keep points from lying along a straight line than item bias, using proportion correct will not assure that the situation will be so. Lord (1976) then demonstrates in a quite clear and simple fashion that the proportion of correct answers (classical item difficulty) is not really a measure of item difficulty. Stated simply, we would want the item difficulty to be independent of the people used to determine the index; this is not possible using "proportion-correct" as the index.

Wright, Mead, and Draba (1976) and Hunter (1975) offer further discussions about the problems inherent in using group-based statistics as indicants of item bias or test bias. Factor analytic approaches, whereby factor structures for the groups are compared, or the use of item-test point biserials suffer from the same problem as proportion correct; the indices are dependent upon the group from which the measures were established.

If all of the traditional indices, which not only describe the test item, but also the group tested, are of questionable use in dealing with test bias, what can be done? A useful index would have to be free of the group used for defining it. This "sample-invariant" property does exist for latent trait model parameters.

Using latent trait theory rather than classical test theory, we can formulate a definition of item bias in a different fashion. According to Pine (1976):

> A test item is unbiased if all individuals having the same underlying ability have an equal probability of getting the item correct, regardless of subgroup membership.

This means that item characteristic curves which provide the probabilities of correct responses must be identical across different sub-populations of interest. Taking this one step further, if the item characteristic curves are the same, then the item parameter(s) have the same values for the subgroups, up to a determinable linear transformation. If the subgroups upon which the parameters are calibrated differ in means and variances, then a linear transformation will be necessary to equate scales. If the transformation is not applied, the parameters will be linearly related for the subgroups (assuming no bias).

How does one proceed? At least three solutions are currently being studied. Lord (1976) is developing a statistical test for deciding whether the item characteristic curve for an item is the same for the subgroups involved. Pine (1976) discusses first a test for unidimensionality, and also describes a possible method for correcting for item bias by adjusting item parameter estimates. Pine and Weiss (1976) take items of varying item bias and look at how this affects three test fairness models—the Cleary model, the Thorndike model, and the model based upon a validity correlation with an external criterion. Wright, Mead and Draba (1976) and Mead (1976), utilizing the Rasch model, develop, through the use of residuals, an ANOVA approach to detecting item bias.

A study described by Lord (1976) is now in progress at ETS. Noteworthy is that he advances a two step approach to the detection of item bias:

1. Plot item difficulties for the subgroups on the same graph, and fit the plotted points with a straight line. This will put all items on the same reference scale, and aberrant items will demonstrate significant departures from linearity.

2. Test the hypothesis that the aberrant item has the same item characteristic curve for the subgroups of interest.

Pine (1976) suggests a two step procedure like Lord's, but he adds one additional step; namely testing for unidimensionality of a set of test items. If the trait dimensions are the same, any variability in parameter values can be attributed to item bias. However, it was mentioned earlier in the paper that factor analysis of tetrachoric correlation matrices has problems associated with it. It remains to be seen how useful in practice this step will be.

Wright, Mead, and Draba (1976) and Mead (1976) utilize the "simpler" nature of the Rasch model to develop a very interesting approach to studying test and item bias. They first form a residual, i.e. the difference between observed outcome on an item, and expected outcome based upon the model, and then transform metrics from the proportion metric to the ability metric. Using residuals on the ability metric, they are able to set up a weighted least squares ANOVA for testing shifts in item difficulty across subgroups, which in the Rasch model, would be the sole indicant of item bias. Mead (1976) also discusses a graphical method whereby residuals are plotted against the ability scale. The residuals plotted against the ability scale fall along a horizontal line through the origin. Any disturbance, such as guessing, discrimination differences caused by practice or speed, or most important here, item bias, will appear as a departure from the horizontal. The shapes of departures would then indicate the sort of disturbance present.

In summary, the application of latent trait models for the detection of item bias is just now beginning. As with the field of tailored testing, classical test theory will not solve the problem of interest. Certain transformations of the classical indices will help in curtailing some problems, but one can never escape the dependency upon population characteristics. As such, any indication of item bias can't be read in a pure fashion; it could also be the result of another variable, such as guessing, which classical indices cannot control for.

The areas for further expansion and research have been well defined elsewhere. These include:

1. Development of a method for correcting item parameter values to account for bias in the item. This would seem to be of value in eliminating the effect of bias in the item rather than eliminating the item itself. Pine is presently working on such techniques.

2. A further study of the effects of item bias on other test fairnes models other than those investigated by Pine and Weiss (1976).

3. Further study and documentation of the ANOVA of residuals method developed by Wright et al. (1976).

In conclusion, while the use of item characteristic curves for detecting item bias is in the beginning stages, it appears that the critical areas of concern are now being investigated. The months ahead will bring evidence as to the feasibility and practicality of the use of these methods.

Test Equating

Large scale testing situations often dictate the need for multiple and interchangeable forms of the same test. Test construction techniques do not assure that two (or more) forms of a test can be made equivalent in level and range of difficulty, and hence there is a necessity for test-score equating. In equating the forms, the system of units of one form is converted to the system of units of the other, so that scores derived from the two forms, after conversion, will be equivalent (Angoff, 1971).

The advantages of equating test scores is that one can study and measure growth, using equated forms, can merge data when the data is derived from different forms of a test, and perhaps most importantly, equating allows comparison of performance of two individuals who have taken different test forms.

Two sorts of stipulations or restrictions involving the equating process can be exclamated:

1. The tests that are to be equated must be measures of the same characteristic. Tests measuring different traits or abilities cannot be equated.

2. If equating is to be a transformation of only systems of units, the transformation must be unique (except for a random error component). By this is meant that the transformation must not be situation specific, but be independent of the individuals from which the data were drawn to perform the conversion, and be applicable to other situations.

The extant literature in this field can be roughly separated into three areas: Angoff's explication of the field (prior to the use of latent trait theory), the Rentz and Bashaw work (1975, 1977) on equating using the Rasch model, and Lord's work (1975a, 1977b).

The methods described by Angoff are adequate for handling parallel tests that are to be equated. Lord's work essentially deals with non-parallel equating situations, and his 1975 study contrasts situations where raw score methods using equipercentile equating can be used, to the use of item characteristic curves for the same situations.

There are essentially three distinct ways of collecting data for an equating project: (1) Administer the two tests to the same group of individuals. (2) Administer the two tests to two equivalent groups of individuals, where the groups are set up by random sampling or (3) Administer the two tests along with an anchor test to two groups that need not be equivalent. The anchor test, which is administered as a part of both of the non-parallel forms to be equated, measures differences from equivalence in the two groups. The anchor test should demonstrate a high correlation with the two tests to be used in an equating study.

Besides the three methods of data collection as mentioned above, there are also two methods of non-linear equating. One method is the equipercentile method using raw scores (Angoff, 1971). For non-parallel tests, the true scores on the two tests will have a non-linear relationship, and because of this, the standard error of measurement for the equated test will probably not be equal to the standard error of measurement of the test being equated to, for the entire score scale. This is critical to equating, and if the standard errors are not the same, raw scores cannot be equated with the assurance of strict interchangeability.

The other method of equating is based upon ability estimates using item characteristic curves. Lord (1975a) points out that if we are willing to equate on ability estimates $\hat{\theta}$, and if the theory holds, it models the non-linear relationship exactly. This means only linear relationships would need to be dealt with in equating.

Using data from the Anchor Test Study (Loret, Seder, Bianchini, Vale, 1974), based upon a single group of individuals who took both tests to be equated, Lord demonstrated that equating using item characteristic curves formulated an equating line that closely coincided with the line developed by equipercentile methods. Using the LOGIST program (Wood et al., 1976), item parameters and a single ability estimate for each individual were obtained by combining forms. Then estimated true scores $\hat{T}$ were found for each test form from the relation:

$$\hat{T} = \sum_{g=1}^{n} \hat{P}_g(\theta)$$

where $\hat{P}_g(\theta)$ is the three-parameter logistic curve estimated by LOGIST. These estimated true scores were then equated, and the method was found to closely coincide with equipercentile methods using raw scores.

In sum, Lord's studies, involving a single group, demonstrate that true score equating and equating using the estimated distribution of observed scores closely coincide with the conventional method of equipercentile equating of raw scores. It remains to be seen which method is practically most advantageous, but from a computer time point of view, the conventional method would seem more practical if item parameters have to be estimated for each item. If this had already been done, the decision would be less clear.

Using another data set from the Anchor Test Study, where representative and equivalent samples took one of the two tests, Lord was able to equate the testsusing a number of methods. These included:

1. Because there were no overlapping students (as in the single group) or overlapping items (as in the Anchor Test), there was no way to get a single ability estimate across both tests for an individual. Therefore an ability estimate was gotten for each examinee on the respective test he/she took, and the ability estimates were equated using the equipercentile method. An advantage of such a method is that when the two tests measure the same ability, the ability estimates have a straight line relationship under the latent trait model used (the raw scores would not). This allows easier extrapolation at the extremes of the distribution, where data is often scarce.

2. Using the straight line plotted to the ability estimates and using an inverse transformation twice (see Lord, 1975a), the curvilinear relationship between true scores may be obtained and the scores equated.

Thus, for equivalent groups, equating using ability level offers a distinct advantage in that the line for equating will be straight. Other ways of equating (using estimated true scores, estimated distributions of observed scores, or equipercentile equating using raw scores) have curvilinear equating lines. It is as if by using ability estimates for equating, we are reducing the equating problem for non-linear tests to one of linear (parallel) tests.

The third and final method of data collection for equating tests involves an Anchor Test. Because items overlap, one ability estimate can be obtained for each examinee and then estimated true scores $\hat{T}$ are computed and equated as in the second design. Also, an estimated frequency distribution of raw scores can be obtained and equated as with the first method. Both methods were compared to the equipercentile method using raw scores, and there was less coinciding of the equating lines derived from item characteristic curves with the raw scores than before. Lord offers an explanation:

> The conventional equipercentile equating of two tests to an anchor test is an inefficient, and strictly speaking, a biased and inadequate equating procedure for groups that differ in ability level.

Thus, it would appear that in this situation equating using item characteristic curves is a necessity. When the tests are not parallel, and the groups are not equivalent, it would appear that item characteristic curve methods are the only adequate way of ascertaining equality of two tests.

In summary, when a single group takes both tests, it would appear that the use of latent trait theory would be advantageous only if item parameters have already been estimated. Results appear to coincide for raw score equating and equating using item characteristic curves and the decision about method will probably be based upon computer use.

When equivalent groups take the two tests, latent trait theory equating offers a distinct advantage if the equating is done using ability estimates, for the equating line will be straight and extrapolation problems minimized. Any other method of equating using item characteristic curves seems to offer no advantage over conventional methods.

When an anchor test is used for non-equivalent groups, item characteristic curve equating is the only justifiable method to use.

So far, we have said little about the use of the Rasch model in equating tests (Brigman & Bashaw, 1976; Rentz & Bashaw, 1975). The following points can be made:

1. The papers by Lord deal with the use of general item characteristic curves; that is, item parameters are not restricted. From this point of view, use of the Rasch model can be viewed as a special case of Lord's work.

2. The items must fit the assumptions of the Rasch model. If they do not, it would seem a necessity that a discussion of the uses of other latent trait models be presented.

3. The Rasch procedure is based upon obtaining equating constants for the two tests (see, Rentz & Bashaw, 1975). Two methods exist for doing this, the item difficulty method and the ability method. In either case, it is necessary that the same group of individuals take both tests. Thus, the procedures can be viewed as a subset of our discussion of data collection method one above. While the simplicity of the Rasch equating procedure would seem to warrant its use, it can only be used for test items that fit the model and under situations where the same group takes both tests.

In Rentz and Bashaw (1975), the authors conclude that equating using the Rasch model involved an equating line that closely coincided with the conventional method. They also mentioned that the Rasch procedure involved less time, effort, and money (discussed as savings). Two comments seem appropriate: (1) The results confirm the results of Lord's study using a single group, and (2) The mentioned savings may have been partially an artifact of the complexity of the equating study. The Rasch study was a reanalysis of the data from the Anchor Test Study, which is of a complex nature, involving multiple equatings. It is not really known at present whether equating using item characteristic curves on a single group, using the Rasch model or otherwise, always affords a savings over conventional methods. The mentioned savings may in fact be situation specific.

The present state of test equating would seem to be well explicated. Unlike some of the other applications of latent trait theory, like tailored testing, there are conventional methods, not using latent trait theory, that work well in a variety of situations. Those areas where latent trait models offer explicit advantages have been discussed. Lord (1975a, 1977b) does, however, briefly indicate two areas that need further work:

1. There needs to be more studies done using item characteristic curves in equating, and particularly in the comparison of equating methods using different item characteristic curve models to conventional methods.

2. If two tests are not parallel to begin with (i.e. have a non-linear equatin curve), one is forced into the logic that the tests are not equally reliable for all subgroups of examinees. Thus, by definition, it is not proper to equate raw scores. Faced with a choice of exact true score equating or inexact raw score equating, one finds no criterion for choosing which to use. A set of criteria would need to be developed for this and other situations when a procedural choice must be made.

Estimation of Power Scores

A speeded test is defined as one for which examinees do not have time to respond to some questions for which they know the answers. A power test is one for which examinees have sufficient time to show what they know. Most academic achievement tests are more speeded for some examinees than for others.

Occasionally the situation exists when a test, that is intended to be a power test, becomes a speeded test. An example of this situation is a test that has been mistimed, i.e., examinees are given less than the specified amount of time to complete the test. In this situation, it would be desirable to estimate what an examinee's score would have been if the test had been properly timed. This score is referred to as an examinees' power score.

Power scores are not difficult to obtain if the test items are all of equal difficulty and equal discriminating power. An examinee's expected item score on each unanswered item would equal the examinee's proportion-correct score on the items that were attempted. However, if items vary in difficulty or discrimination, another method is needed. Lord (1973) has discussed a method using the three-parameter logistic model and applied it to the estimation of power scores for 21 examinees who had taken a mistimed verbal aptitude test.

Lord's method requires not only the usual assumptions of the three-parameter logistic model, but also it assumes that the students answer the items in order and that they respond as they would if given unlimited time, i.e., if given more time, they would not go back and change any of their answers.

If the test score, x, is the number of correct answers, the expected power score for an examinee with ability level, θ, for a set of n items is equal to the sum of the examinee's probability of answering each item correctly. The probabilities are obtained from the item characteristic curves. Therefore, if there is sufficient data to estimate an examinee's ability score, and the item characteristic curve parameters are known (or can be estimated) an examinee's power score on the n test items can be estimated: It is equal to the examinee's test score on the attempted items plus the examinee's expected score on the unanswered items (found by summing the examinee's probabilities of answering each unanswered item correctly). Suppose k is used to designate the last item attempted by an examinee, x is the examinee's score, n is the number of items in the test, and $\hat{\theta}$ is the examinee's estimated ability derived from the k items attempted by the examinee. The examinee's estimated power score is given by

$$x + \sum_{g=k+1}^{n} P_g (\hat{\theta}) .$$

Lord (1973) reported the following application of his method. Item parameters of the 90 verbal aptitude items comprising the mistimed test were estimated using responses obtained from 944 students including the 21 mistimed students. Abilities were estimated for 21 students from their responses to the items excluding responses to any unanswered items at the end of the test. Power scores were estimated using the method described above.

Lord felt his method could be justified empirically if the following properties of the estimates could be demonstrated:

1. Estimates of item parameters from one group of examinees closely approximate estimates of the same item parameters from other groups of examinees.
2. Estimates of ability parameters from part of a test closely approximate estimates obtained from the entire test.
3. The power score of an examinee on a test can be accurately approximated from his ability estimate as estimated from the same test.

In Lord's judgment, the available evidence has been quite favorable:

1. Lord (1970a) showed good agreement between estimates of item characteristic curves from two different groups of examinees.
2. Correlations over .94 were obtained between ability estimates derived from different subsets of items in one study of SAT response data.
3. The correlation between power scores and number right scores has exceeded .98 in two different studies.

Lord cautioned that a wide variety of empirical checks would have to be carried out before one could be sure of all the circumstances under which the three properties of the estimates listed above would hold.

Computer Programs

How can test practitioners use latent trait models in their work? Fortunately, there are a number of computer programs available for estimating ability and item parameters (Hambleton and Rovinelli, 1972; Kolakowski and Bock, 1970; Wood and Lord, 1976; Wood, Wingersky, and Lord, 1976; Wright and Mead, 1976a, 1976b; Wright and Panchapakesan, 1969). Some details on four of the computer programs, LOGIST, CALFIT, BICAL, and DATAGEN, will be provided next.

LOGIST (Wood and Lord, 1976) allows the user to estimate examinee abilities and all parameters of the three-parameter logistic model. A maximum likelihood method is used to obtain estimates of the item and ability parameters (Lord, 1974a). The item and ability parameters are estimated simultaneously. For the estimates of the parameters to converge, various restrictions are placed on the parameters being estimated. Ability estimates are scaled to have a mean of zero and a standard deviation of one.

The following statistics, reported by Wood et al. (1976), give some idea of the computer time required for running on an IBM 360-65. A test of 60 items and 5305 examinees took approximately 230 seconds per complete stage. A complete stage involves the estimation of both ability and item parameters. A test with 85 items and 2269 examinees took approximately 130 seconds per complete stage. Convergence was obtained after 10-15 stages. To achieve convergence, certain restrictions are imposed: For example, (1) abilities for examinees with zero scores, perfect scores, and those who answered less than 1/3 of the items, are not estimated; and (2) an upper bound value is imposed on the estimated discrimination parameters.

Wood, et al. (1976) provide a complete description of the output from the program after each stage and after the job is completed. The output after the final stage is completed includes: (1) Final item and ability estimates; (2) a summary containing various statistics for each stage; and (3) the total time for the run.

According to a write-up (Wright and Mead, 1976a) on BICAL, "The BICAL program estimates the parameters of the Rasch model when the underlying response process is binomial . . . The algorithms used for estimating item difficulties and person abilities are the corrected unconditional maximum likelihood procedure and a normal approximation . . . In addition to estimates of difficulty and ability, and tests of item fit, output includes the standard errors associated with these estimates, residual indices of item discrimination and the degree of convergence of the estimation procedures." BICAL contains a data simulator which can be used to verify the functioning of the program or to provide an appropriate random background for the Monte Carlo analysis of unusual data.

The CALFIT program has also been described by Wright and Mead (1976b). This program performs 4 major tasks: (1) Data input and description; (2) data editing; (3) estimation of parameters; and (4) analysis of fit.

The output includes: (1) The distribution of examinees by total score; (2) the results of the estimation process; (3) the number of iterations required for convergence; (4) the analysis of the fit of the data to the Rasch model; (5) a summary of the fit information in three sequences; serial order, difficulty order, and fit order; (6) a plot of the Z^2 statistics, used in the fit analysis, against the probability of a person in an ability group answering the item correctly; (7) a plot of the item fit mean squares against item difficulty; (8) a plot of the item fit mean squares against the index of item discrimination; and (9) a plot of the item discrimination index against item difficulty.

Hambleton and Rovinelli (1972) have produced a computer program (DATAGEN) to simulate examinee item response data from logistic test models. One purpose of the computer program is to allow users the opportunity to study relationships among item and examinee ability parameters, logistic test models, and test score characteristics. A second purpose of the computer program is to produce test data with known characteristics, so that robustness studies, studies of estimation methods, studies of scoring methods, and so on, can be conducted.

The program is designed to produce a set of response patterns and test scores to represent the performance of N examinees on n binary-scored items. By appropriate choice of item and ability parameters in the program, it is possible to produce a set of response patterns with a distribution of test scores approximating desired mean, variance, kurtosis and skewness values. Description of the item parameters in the logistic test models used to generate the test data are described by Lord and Novick (1968) and Hambleton and Traub (1971).

The user reads in specifications for the distribution of item difficulty, discrimination, and guessing parameters and ability parameters. Parameters may be selected from either a uniform distribution with specified upper and lower bounds, or a normal distribution with a specified mean and standard deviation. The user also specifies the desired number of examinees and items, and starting numbers for the random number generator.

Output from the program includes desired descriptive statistics on the item parameters and estimated values on the basis of sample data; a listing of the item parameters and estimated conventional item parameters calculated from the generated test data. Also reported is a complete set of summary statistics on the generated response patterns and test scores. Response patterns may be either saved on a data tape or punched out on computer cards.

The program is currently designed to generate response patterns on up to 100 items although the number of items can easily be increased by changing a few dimension statements. The program is practically machine independent except for the random number generator.

Final Comments

The goal of this paper has been to review the developments in latent trait theory to date, to demonstrate the applicability of latent trait theory models to specific measurement problems, and finally, to point out the advantages of the latent trait theoretical approach over the classical approach for the solution of mental measurement problems. However, the latent trait theoretical models are, in general, mathematically more complex than the classical test models, require strong assumptions that may limit their applicability to mental data sets, and, in some cases, pose problems that are, as of yet, unresolved.

As pointed out in the paper, the latent trait models have numerous advantages over the classical test models. Perhaps the most important advantage of latent trait models is that it is possible to estimate an examinee's ability on the same ability scale from any subset of items that have been fitted to the model. This implies that the ability of an examinee can be estimated independently of the particular choice or the number of items and hence represents a major breakthrough in the area of mental measurement. A consequence of this fact is that examinees may be compared with each other even though they may have taken quite different subsets of items. This feature makes latent trait models indispensable to the field of tailored testing where examinees receive test items that are matched to their ability level. In such situations the items administered to different examinees will not be matched on difficulty, and hence the usual test score metric will not permit meaningful comparisons of examinees. Latent trait models take into account the difficulty level of the items and reflect this in the estimates of the ability. Inus, the estimates of the abilities of two examinees, who receive identical scores on easy and difficult subtests, may differ, and hence a meaningful comparison of the examinees is possible. A further consequence of the fact that ability can be estimated independently of the choice of items is that, equating scores of tests that measure the same ability is possible. In addition, the problem of constructing parallel forms of tests is eliminated.

Another advantage of latent trait models is that the item parameters are invariant across subgroups of examinees chosen from a population of examinees. Item parameters, such as item difficulty and discrimination, derived from classical test theory models are not invariant across subgroups. They are defined for a particular group of interest and will depend on the average ability of the group being tested. Hence, despite their computational ease, classical item parameters do not permit meaningful comparisons across different populations of interest. Item parameters based on latent trait models, on the other hand, permit comparisons across different populations of interest and consequently are of immense value to test developers. In particular, invariant item parameters are of fundamental importance in the development of item banks and in detecting item bias.

A further property inherent in latent trait models not exhibited by classical test models, is that it is possible to measure the precision of the ability estimates at each ability level. Thus, instead of providing a standard error of measurement that applies to all examinees regardless of test scores, separate estimates of error for each examinee or at each ability level are available through the latent trait models.

Despite these advantages, there are several unresolved issues which need further investigation. Since latent trait models require strong assumptions, the question that naturally arises is that of the robustness of the latent trait models. Robustness refers to the extent that data can deviate from underlying assumptions of a latent trait model and still be fit by the model. The studies reported to date have often produced different conclusions (see for example, Hambleton [1969] and Panchapakesan [1969]). Researchers have reached different conclusions because they have used subjective methods to interpret the results of robustness studies. It is obvious that the assumptions of any latent trait model will never be completely satisfied by any data set. Hence, the important questions are whether latent trait analyses provide useful summaries of test data, lead to better test score interpretations, and can predict appropriately chosen criteria. When the last question was studied by Lord (1974a), he obtained excellent predictions. However, the issue of robustness is not completely resolved as of yet and further work is clearly needed to resolve these issues.

The major problem that remains to be solved is that of estimation of parameters in latent trait models. As pointed out earlier, the simultaneous estimation of item and ability parameters in latent trait models leads to difficulties. In addition, the estimates of the item parameters, especially that of the guessing parameter, c_g, will not be stable if examinees with a wide range of abilities are not used. Furthermore, current estimation procedures require a large number of examinees and items before stable estimates can be obtained, a problem similar to that of estimating parameters in regression models. The numerical problems associated with the estimation procedures present another area of concern.

Further research is clearly needed in the above areas. Although it may not be possible to show that the maximum likelihood estimates of item and ability parameters possess optimal properties, these estimates may approximate the ideal estimates in some situations. For instance, the comparison of the unconditional estimates and the conditional estimates of the item parameters in the Rasch model (Wright and Douglas, in press) has provided a meaningful insight into the nature of the estimates. These comparisons can be carried out for the two- and three-parameter logistic models. (In this connection, it should be pointed out that unconditional estimates in the sense of Bock [1972] have not been obtained for the three-parameter logistic model.) Finally, the feasibility of Bayesian procedures should be investigated more fully. Incorporation of prior information in the estimation procedure may provide improved estimates of the parameters and may also permit estimation of parameters with a small sample size and a small number of items. However, poor specification of priors may adversely affect the estimates and hence a careful study of appropriate priors would be necessary.

In conclusion, we note that latent trait theorv offers the promise for solving the problems that arise in mental measurement. The advantages of the latent trait theoretic approach over the classical test theoretic approach are obvious. It appears that the major factors that have hindered wide spread use of latent trait theoretic methods are the lack of familarity with these methods on the part of practitioners and the lack of user oriented computer programs. These problems have been overcome in recent years and hence we can expect latent trait theoretic procedures to emerge as methods of the future for the measurement of mental abilities.

References

Andersen, E.B. Asymptotic properties of conditional maximum likelihood estimates. The Journal of the Royal Statistical Society, Series B, 1970, 32, 283-301.

Andersen, E.B. The numerical solution of a set of conditional estimation equations. The Journal of the Royal Statistical Society, Series B, 1972, 34, 42-54.

Andersen, E.B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140. (a)

Andersen, E.B. Conditional inference in multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 1973, 26, 31-44. (b)

Anderson, J., Kearney, G.E., & Everett, A.V. An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement. Washington: American Council on Education, 1971.

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.

Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. Journal of Mathematical Psychology, 1969, 6, 258-276.

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

Bock, R.D., and Liebermann, M. Fitting a response model for n dicotomously scored items. Psychometrika, 1970, 35, 179-197.

Bock, R.D., and Wood, R. Test theory. Annual Review of Psychology, 1971, 22, 193-224.

Bradley, J.B. Distribution-free Statistical Tests. Englewood Cliffs, NJ: Prentice-Hall, 1968.

Brigman, S.L., and Bashaw, W.L. Multiple test equating using the Rasch model. A paper presented at the annual meeting of AERA, San Francisco, 1976.

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.

Cleary, T.A., Linn, R.L., & Rock, D.A. An exploratory study of programmed tests. *Educational and Psychological Measurement*, 1968, *28*, 345-360.

Coffman, W.E. A factor analysis of the verbal sections of the scholastic aptitude test. *Research Bulletin 66-30*. Princeton, NJ: Educational Testing Service, 1966.

Fischer, G.H. Some probabilistic models for measuring change. In D. DeGruijter & L.J. Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement*. New York: Wiley, 1976.

Green, B.F. Comments on tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted Instruction, Testing, and Guidance*. New York: Harper and Row, 1970.

Haley, D.C. Estimation of the dosage mortality relationship when the dose is subject to error. *Technical Report No. 15*. Stanford, CA: Applied Mathematics and Statistics Laboratory, Stanford University, 1952.

Hambleton, R.K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R.K. Contributions to criterion-referenced test theory: On the uses of item characteristic curves and related concepts. *Laboratory of Psychometric and Evaluative Research Report No. 51*. Amherst, MA: School of Education, University of Massachusetts, 1977.

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

Hambleton, R.K., and 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., and Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science*, 1973, *18*, 74.

Hambleton, R.K., and Traub, R.E. Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 273-281.

Hambleton, R.K., and Traub, R.E. Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 195-211.

Hambleton, R.K., and Traub, R.E. The robustness of the Rasch test model. Laboratory of Psychometric and Evaluative Research Report No. 42. Amherst, MA: School of Education, University of Massachusetts, 1976.

Hunter, J.E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, Maryland, 1975.

Jensema, C.J. An application of latent trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 29-48. (a)

Jensema, C.J. The validity of Bayesian tailored testing. Educational and Psychological Measurement, 1974, 34, 757-766. (b)

Jensema, C.J. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 1976, 36, 705-715.

Keats, J.A. Test theory. Annual Review of Psychology, 1967, 16, 217-238.

Kendall, M.G., and Stuart, A. Advanced Theory of Statistics, Vol. II. New York: Hafner Publishing Co., 1973.

Kiefer, J., and Wolfowitz, J. Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics, 1956, 27, 887-890.

Kolakowski, D., and Bock, R.D. A FORTRAN IV program for maximum likelihood item analysis and test scoring: Normal ogive model. Educational Statistics Laboratory Research Memo No. 12. Chicago: University of Chicago, 1970.

Lawley, D.N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, 273-287.

Lawley, D.N. The factorial analysis of multiple item tests. Proceedings of the Royal Society of Edinburgh, 1944, 62-A, 74-82.

Lazarsfeld, P.F. The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al., Measurement and Prediction. Princeton: Princeton University Press, 1950.

Lazarsfeld, P.F., and Henry, N.W. Latent structure analysis. New York: Houghton Mifflin, 1968.

Lindley, D.V., and Smith, A.F.M. Bayesian estimates for the linear model. Journal of the Royal Statistical Society, 1972, 34, 1-41.

Linn, R.L., Rock, D.A., & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1972, 32, 85-95.

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

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, *18*, 57-75. (a)

Lord, F.M. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, *13*, 517-548. (b)

Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F.M. Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika*, 1970, *35*, 42-50. (a)

Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), *Computer-Assisted Instruction, Testing, and Guidance*. New York: Harper and Row, 1970. (b)

Lord, F.M. A theoretical study of two-stage testing. *Psychometrika*, 1971, *36*, 227-242. (a)

Lord, F.M. Robbins - Monro procedures for tailored testing. *Educational and Psychological Measurement*, 1971, *31*, 3-31. (b)

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

Lord, F.M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 1971, *31*, 805-813. (d)

Lord, F.M. Power scores estimated by item characteristic curves. *Educational and Psychological Measurement*, 1973, *33*, 219-224.

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

Lord, F.M. Individualized testing and item characteristic curve theory. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology, Vol. II*. San Francisco: Freeman, 1974. (b)

Lord, F.M. Quick estimates of the relative efficiency of two tests as a function of ability level. *Journal of Educational Measurement*, 1974, *11*, 247-254. (c)

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

Lord, F.M. A survey of equating methods based on item characteristic curve theory. Research Bulletin 75-13. Princeton, NJ: Educational Testing Service, 1975. (a)

Lord, F.M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. Research Bulletin 75-33. Princeton, NJ: Educational Testing Service, 1975. (b)

Lord, F.M. Relative efficiency of number-right and formula scores. British Journal of Mathematical and Statistical Psychology, 1975, 28, 46-50. (c)

Lord, F.M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 44, 205-217. (d)

Lord, F.M. A study of item bias using item characteristic curve theory. Paper presented at the Third International Association for Cross-Cultural Psychology Congress, Tilburg University, Tilburg, the Netherlands, 1976.

Lord, F.M. A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1977, 1, 95-100. (a)

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, in press. (b)

Lord, F.M., and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Loret, P.G., Seder, A., Bianchini, J.C., & Vale, C.A. Anchor test study: Equivalence and norms tables for selected reading achievement tests (Grades 4, 5, 6). Washington, US Department of Health, Education, and Welfare, US Office of Education, 1974.

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.

Marco, G. The application of item characteristic curve methodology to practical testing problems. Journal of Educational Measurement, 1977, in press.

McDonald, R.P., and Ahlawat, K.S. Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 1974, 27, 82-99.

Mead, R. Assessing the fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Fracisco, 1976.

Meredith, W., and Kearns, J. Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. Psychometrika, 1973, 38, 533-554.

Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in Education: Current Practices. Berkeley, CA: McCutchan Publishers, 1974.

Mulaik, S.A. The Foundations of Factor Analysis. New York: McGraw Hill, 1972.

Neyman, J., and Scott, E.L. Consistent estimates based on partially consistent observations. Econometrika, 1948, 16, 1-5.

Novick, M.R., and Jackson, P. Statistical Methods for Educational and Psychological Research. New York: McGraw Hill Book Co., 1974.

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

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Pine, S.M. Applications of item response theory to the problem of test bias. Unpublished manuscript. Minneapolis: Department of Psychology, University of Minnesota, 1976.

Pine, S.M., and Weiss, D.J. Effects of item characteristics on test fairness. Research Report 76-5. Minneapolis: Department of Psychology, University of Minnesota, 1976.

Rao, C.R. Linear Statistical Inference and Its Application. New York: Wiley, 1965.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Rentz, R.R., and Bashaw, W.L. Equating reading tests with the Rasch model, Volume I final report, Volume II technical reference tables. Athens, GA: University of Georgia, Educational Research Laboratory, 1975.

Rentz, R.R., and Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, in press.

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

Ross, J., and Lumsden, J. Attribute and reliability. British Journal of Mathematical and Statistical Psychology, 1968, 21, 251-263.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, 1969, No. 17.

Samejima, F. A general model for free-response data. Psychometric Monograph, 1972, No. 18.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233. (a)

Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219. (b)

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

Thissen, D.M. Information in wrong responses to Raven's Progressive Matrices. Journal of Educational Measurement, 1976, 13, 201-214.

Tinsley, H.E.A., and Dawis, R.V. An investigation of the Rasch simple logistic model: Sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Torgerson, W.S. Theory and Methods of Scaling. New York: Wiley, 1958.

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

Vale, C.D., and Weiss, D.J. A study of computer-administered stradaptive ability testing. Research Report 74-4. Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

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

Weiss, D.J. Strategies of adaptive ability measurement. Research Report 74-5. Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Weiss, D.J. Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C.L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, DC: United States Civil Service Commission, 1976.

Whitely, S., and Dawis, R.V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

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

Wood, R. Adaptive testing: A Bayesian procedure for the efficient measurement of ability. *Programmed Learning and Educational Technology*, 1976, *13*, 34-48. (a)

Wood, R. Trait measurement and item banks. In D. DeGruijter and L.J. Th. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement*. New York: Wiley, 1976. (b)

Wood, R.L., and Lord, F.M. A user's guide to LOGIST. *Research Memorandum 76-4*. Princeton, NJ: Educational Testing Service, 1976.

Wood, R.L., Wingersky, M.S., & Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. *Research Memorandum 76-6*. Princeton, NJ: Educational Testing Service, 1976.

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*, in press. (a)

Wright, B.D. Misunderstanding of the Rasch model. *Journal of Educational Measurement*, 1977, *14*, in press. (b)

Wright, B.D., and Douglas, G.A. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1977, *1*, in press.

Wright, B.D., and Douglas, G.A. Conditional versus unconditional procedures for sample-free analysis. *Educational and Psychological Measurement*, in press.

Wright, B.D., and Mead, R.J. BICAL: Calibrating rating scales with the Rasch model. *Research Memorandum No. 23*. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976. (a)

Wright, B.D., and Mead, R.J. CALFIT: Sample-free item calibration with a Rasch measurement model. *Research Memorandum No. 18*. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976. (b)

Wright, B.D., Mead, R., & Draba, R. Detecting and correcting item bias with a logistic response model. *Research Memorandum No. 22*. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

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