PROBLEM:

SCORING ADAPTIVE TESTS

James R. McBride
University of Minnesota

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

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

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

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

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

Latent Trait Theory

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

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

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

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

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

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

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

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

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

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

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

Choosing Among Scoring Methods

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

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

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

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

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

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

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

Comparison of Maximum Likelihood and Bayesian Scoring

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

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

In order to generate data from which to compare the three scoring methods, we simulated administering a 20-item Bayesian sequential adaptive test to 3200 examinees of known ability--100 examinees at each of 32 trait levels (Θ) in the interval [-3.2≤Θ≤+3.0] . These trait level values can be thought of as standard deviation units. A pattern of 20 simulated item scores (1 or 0) was generated for each simulated examinee. Every such pattern was scored using each of the three scoring methods. For each scoring method, the mean and standard deviation of the 100 scores at each trait level Θ were calculated.

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

Figure 19

REGRESSION CURVES FOR BAYESIAN AND MAXIMUM LIKELIHOOD SCORING

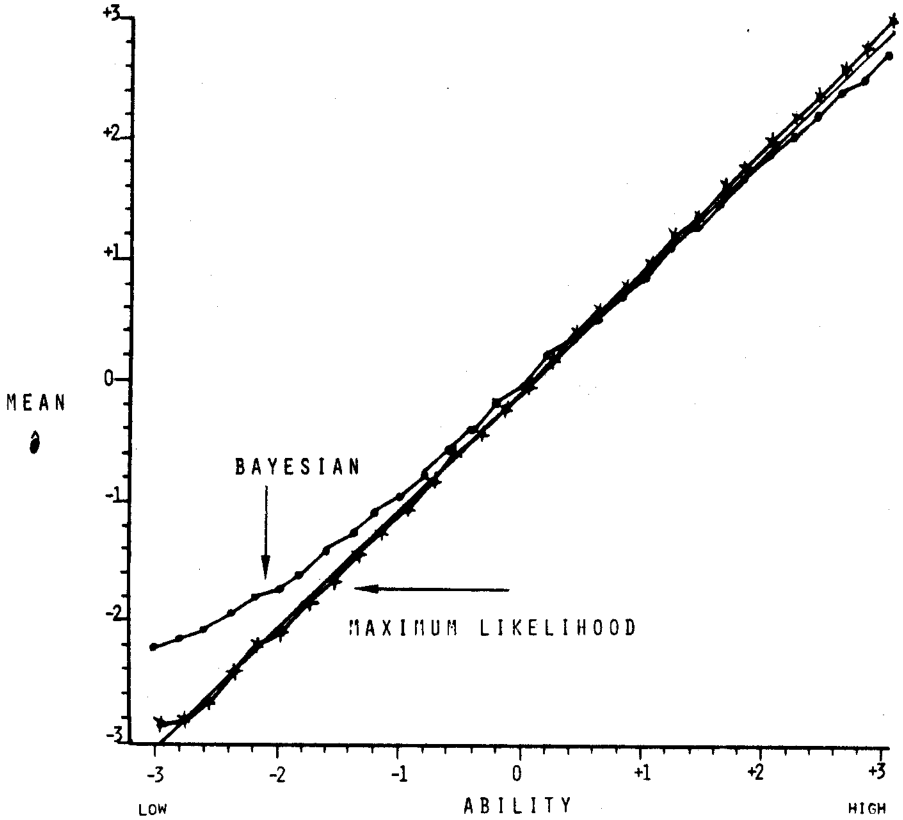


Figure 20

REGRESSION CURVE FOR NUMBER CORRECT SCORE

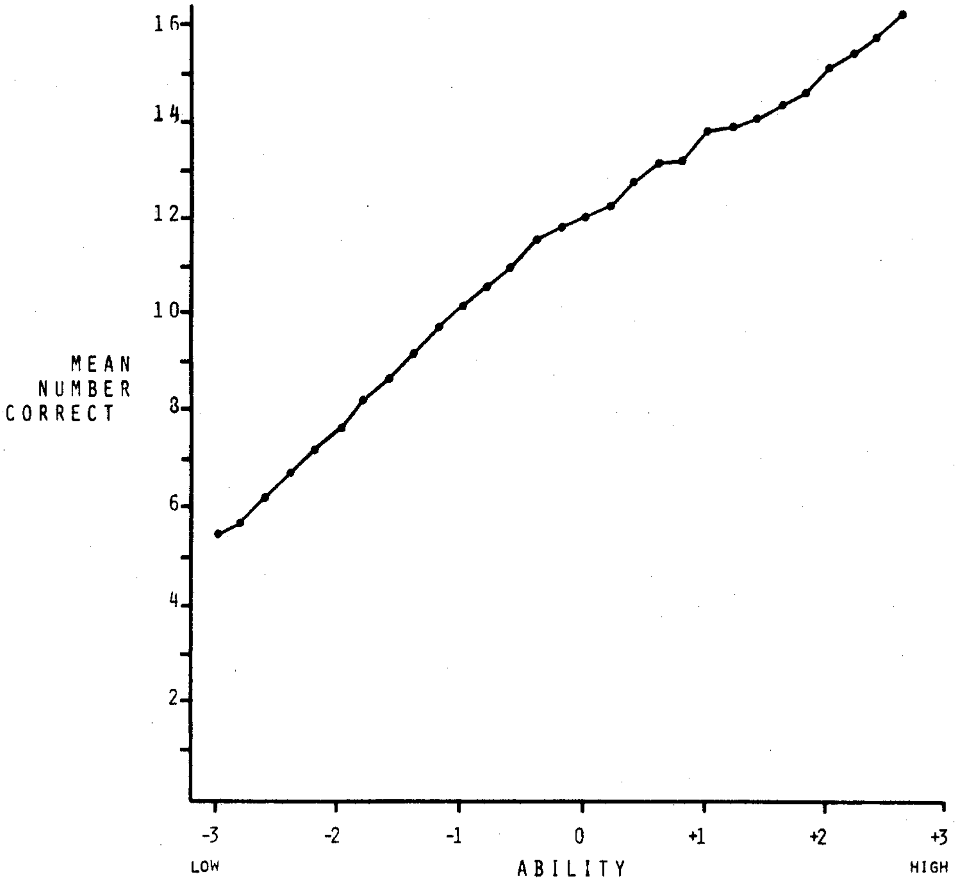
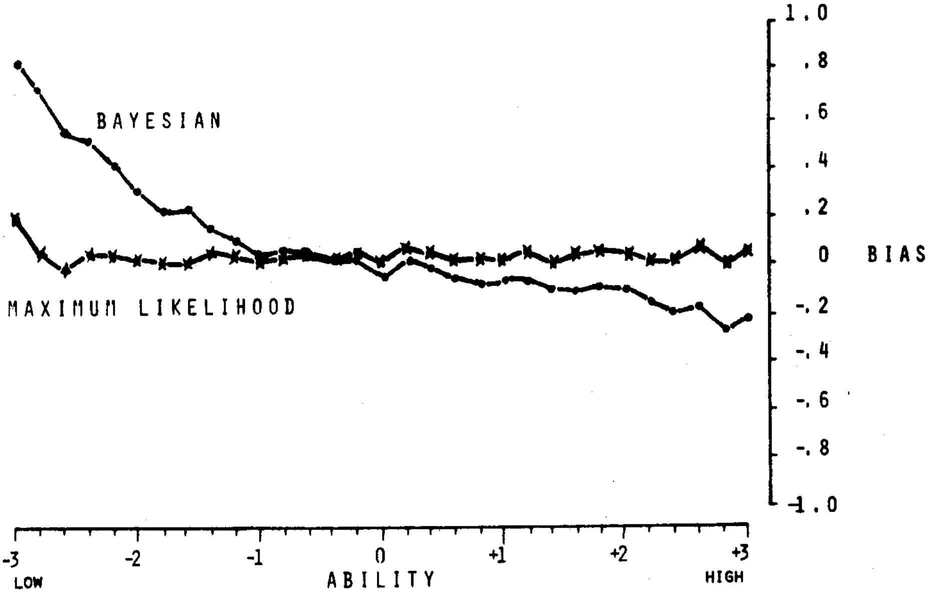


Figure 21

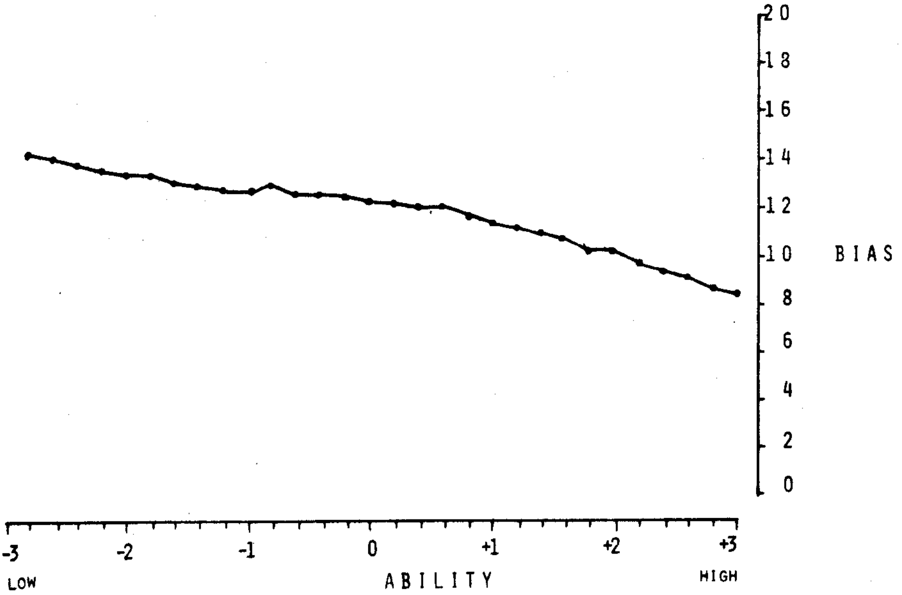
BIAS CURVES FOR BAYESIAN AND MAXIMUM LIKELIHOOD SCORING



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

Figure 22

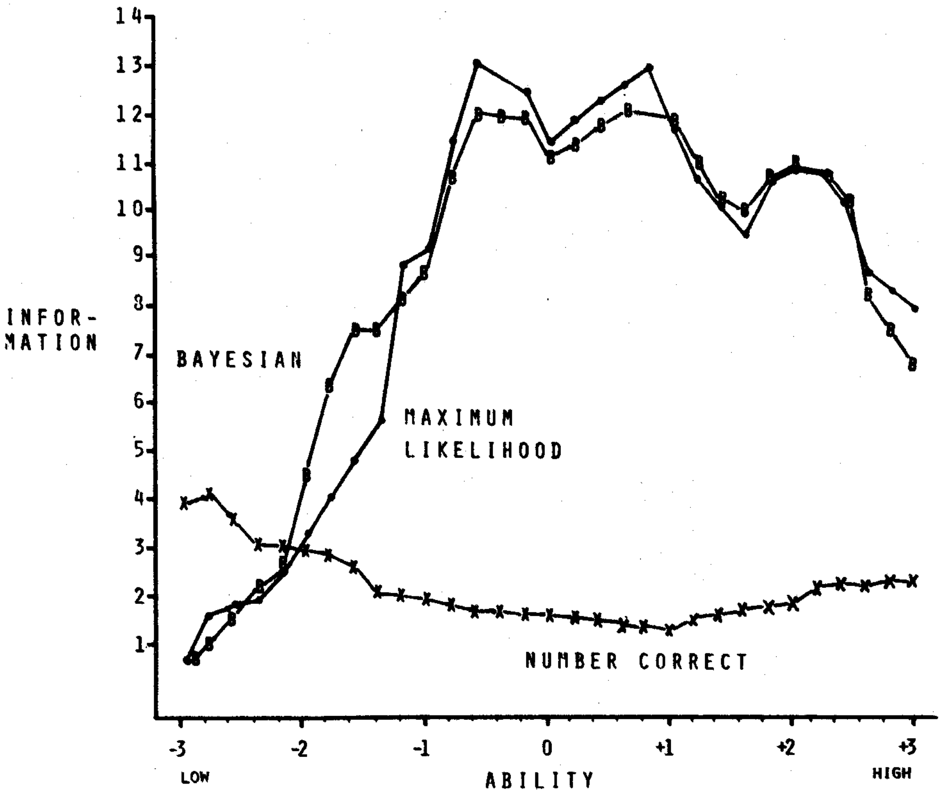
BIAS CURVE FOR NUMBER CORRECT SCORE



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

Figure 23

INFORMATION CURVES FOR THREE SCORING METHODS



Limitations of the Scoring Methods

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

But of course other things are not equal. The maximum likelihood estimation method is about three times more expensive than the Bayesian one. On the other hand, the Bayesian method of scoring is subject to non-linear bias. If unbiased measurement were a goal of the test, the expense of the maximum likelihood procedure might be justified. If simple ordering of persons with respect to trait level were all the tester required, the Bayesian scores seem preferable[1]. Other than that, no simple prescription is advisable.

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

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