DR. FREDERIC M. LORD
Educational Testing Service:

It is appropriate that my discussion should be expressed in the first person singular—to continually remind you that I am giving my own opinions, which may be biased, since I am not a disinterested party here. There have been many, many important points made during these sessions. I have chosen 14 points to emphasize in my discussion.

1. Cliff (Note 1) writes: "It is felt that our formulation will provide the framework for a test theory which is more appropriate to the interactive case than either the classical or traceline theories are." I am sure he would not want this challenge to ICC theory to go unanswered. Cliff proposes that the appropriate model for the item responses is the Guttman scale.

Since the Guttman scale is a special case of the more general logistic or normal ogive item characteristic curve, I cannot see how the Guttman scale can be called a more appropriate model than the logistic or normal ogive. If the Guttman scale were the correct model, the fitted logistic or normal trace lines would come out in the Guttman form.

The Guttman scale assumes that the tetrachoric correlation between any two items is 1.00. This value may be approximated for certain attitude test data, but for aptitude and achievement test data, typical tetrachoric item intercorrelations are usually less than 0.35. This is so *very* different from 1.00 that I cannot see how the Guttman model can be considered acceptable for aptitude and achievement tests.

2. Consider the problem of testing and assigning new armed forces recruits. One recruit, perhaps, should take a complete battery of tests to determine his suitability for officer training school. The next recruit, however, should be quickly extricated from this battery of tests and perhaps given a battery of mechanical aptitude tests. How can we use adaptive testing to route a new recruit through many such batteries of tests efficiently, with a minimum waste of time? Glenn Bryan raised this important question with me some years ago. It seems as if adaptive testing should be an excellent way to deal with this problem. Yet the situation is so multidimensional that current theory does not tell us how to proceed. Here is a very important unsolved problem.

3. Waters has pointed out and documented something that some of us had overlooked--that an adaptive test should be expected to take longer to administer than a conventional test with the same number of items. The reason is that the conventional test contains items that are too hard or too easy for each examinee—items that he can answer (or omit) without need for lengthy consideration. Studies of adaptive testing will have to take testing time into account.

4. There is one situation in which adaptive testing (or some other unconventional procedure) is really indispensable. Suppose it is necessary to have good measurement over an unusually wide range of ability. As a first step, one might build a conventional type of test with extra easy items added at one end and extra hard items at the other, so as to have some items that are appropriate in difficulty for each ability level. Of course, the easy items are a waste of time for the high-level examinees, but that is not the serious problem. The hard items are not merely a waste of time for the low-level examinees. The guessing of low-level examinees on the hard items adds so much noise that the measurement provided by the easy items is nearly drowned in random error.

In such situations, it can be shown that the test would be much improved as a measuring instrument for low-level examinees if we simply threw away (or refused to score) the more difficult half or two-thirds of the test. The situation cannot be remedied simply by adding more easy items. If we wish to obtain good measurement at low as well as at high ability levels, some kind of tailoring is necessary so that hard items are not administered to low-level examinees.

5. If total testing time is held fixed, adaptive testing leads to better measurement for some examinees. If accuracy of measurement is held fixed, adaptive testing leads to reduced testing time for some examinees. These two alternatives are not basically different.

Keeping the standard error of measurement fixed across examinees would be simple if the test were very long or if we knew the true parameter values, and if all items had identical characteristic curves. Otherwise there may be difficulty in finding a good small-sample theory and method. Gugel and Schmidt have given empirical evidence of this. This is a problem in sequential estimation (Wald, 1951; Robbins, 1959; Bickel & Yahav, 1968). Except perhaps for Bayesians, methods of sequential estimation are not as well settled as are methods of sequential hypothesis testing. Even sequential hypothesis testing poses unsolved problems when the items do not all have identical characteristic curves.

6. It is undoubtedly significant that most of the speakers here are using two- or three-parameter item characteristic curve models. No one here has urged that adaptive testing be limited to the one-parameter Rasch model.

It is sometimes asserted that the Rasch model is the only one that allows us to estimate examinee ability independently of the items administered. I would argue that all ICC models allow us to do this. The unique virtue of the Rasch model is that it provides a sufficient statistic for estimating examinee ability. Sufficient statistics are desirable, but they are not common in statistical work, outside of the usual normal-curve theory. Statistical inference still proceeds very effectively in the absence of sufficient statistics.

The objection usually cited against the Rasch model is that it assumes all items to be of equal discriminating power. I suspect that an even more serious objection is that it assumes there is no guessing. Any attempt to modify the Rasch model to take guessing into account would necessarily destroy the sufficiency properties of the Rasch model that make it attractive.

7. This brings us face to face with the question whether to use a two- or a three-parameter ICC model. Waters used a two-parameter normal-ogive model and the assumption that ability is normally distributed to estimate the a parameters (discriminating power) of the 50 verbal items in Form 2B of SCAT II. By chance, I had available estimates of the same parameters based on the three-parameter logistic model, computed by a program called LOGIST (available on request).

I have plotted Waters' values against the LOGIST values in Figure 1. Each point is shown as a digit representing item difficulty. The larger the digit, the more difficult the item and the more the examinees' responses are affected by guessing. Agreement is good only for the easy items where there is no guessing.

Many studies comparing different estimation methods should be carried out. Some should use real data; some should use artificial data, where the true parameters are known. I should be glad to run on LOGIST any suitable set of data that someone here may wish to use for making such comparisons.

8. In the three-parameter models, the ICC's have the form $c_i + (1 - c_i)F[a_i(\theta - b_i)]$. This mathematical form is not beyond challenge, as Samejima has pointed out, but it is relatively easy to defend as a versatile form that fits the data, so long as we do not suggest that examinees either know the answer to the item or else guess with probability of success c_i. We all know that examinees do not respond this way. If ICC theory were based on the dichotomy, knowledge or random guessing, it would not be credible. For this reason, it may be best not to refer to c_i as a 'guessing parameter.' (I confess to violating this good advice.)

9. When working with real answer sheets, it becomes necessary to deal with the problem of omitted responses. If we require the examinee to answer all items, we are purposely introducing random error into our data. In addition, we are forcing an examinee who has demonstrated a certain level of performance by his responses to gamble on some possibily random events, which may, if he is unlucky, destroy all the positive evidence of ability that he has displayed.

If we permit the examinee to omit items, we cannot properly treat such responses as wrong. To do so would penalize the examinee who omits, in comparison to the examinee who guesses.

It seems at first thought that we might simply treat omitted items as if they had not been administered at all. This cannot be correct, however. If we ignore omitted items, an examinee could win a very high estimate of ability simply by answering items only when he was completely sure of his answer.

The fact that an examinee has omitted an item carries information about his level that cannot be ignored. A method for using this information efficiently, under certain assumptions, is outlined in a *Psychometrika* paper (Lord, 1974).

10. I want to take this opportunity to make a correction. In a 1968 paper (Lord, 1970), I wrote:

> If $a_i = 0.333$, under the assumptions already made [the] reliability for a 60-item test will be 0.80; if $a_i = 0.5$, this reliability will be 0.90; if $a_i = 1.0$, this reliability will be 0.97. In view of this, we shall choose $a_i = 0.5$ as a typical value and shall address most of our attention to it.

After seven years of experience with the a parameter, these reliabilities sound high. Actually, they are correct, but, as the assumptions stated, they are for free response, not multiple-choice items. Urry made this same point this morning. Since most of the cited paper dealt with multiple-choice items, it was a mistake to suggest $a_i = .50$ as a typical value. Although the diagrams presented in that paper required the reader to supply his own values of a_i, the general impression given was one of only limited enthusiasm for adaptive testing.

Current results show that when $a_i = 0.9$, a peaked test composed of 40 five-choice items should have a KR$_{20}$ reliability of .90. When a_i is 0.9, the conclusions supplied by the diagrams in the cited paper are quite encouraging for the future of adaptive testing.

11. The purpose of the cited paper was to evaluate adaptive tests in comparison to conventional tests. To do this, the situation considered had to be a simple one. This was the reason for the use of a fixed-step-size up-and-down branching procedure. Such a procedure is *not* to be recommended for practical testing.

When the item parameters have been estimated and a computer is available for making the calculations, the choice of the item to be administered next should be made by checking all unused items (perhaps within a specified item type) and selecting the item that is expected to give the most information about the examinee.

If a Bayesian prior distribution of ability is being used, and if this distribution is normal, this is Owen's (in press) procedure, frequently used today. In such a procedure, except for certain approximations each step is locally optimal. We cannot expect local optimality to produce overall global optimality, but the difference may not be of great importance.

12. When we select the next item to be administered on other considerations besides item difficulty, we no longer

have an up-and-down branching procedure. The next item administered after a correct response might be an easier item, not a harder item.

The recommended procedure means that items with high a_i will be used very frequently and items with low a_i will be used seldom or not at all. The gain from this use of the best items will probably more than double the gain from any procedure, such as the up-and-down procedure, that selects items solely on item difficulty.

Furthermore, the larger the item pool, the greater the gain. This is not surprising. We always knew that if we

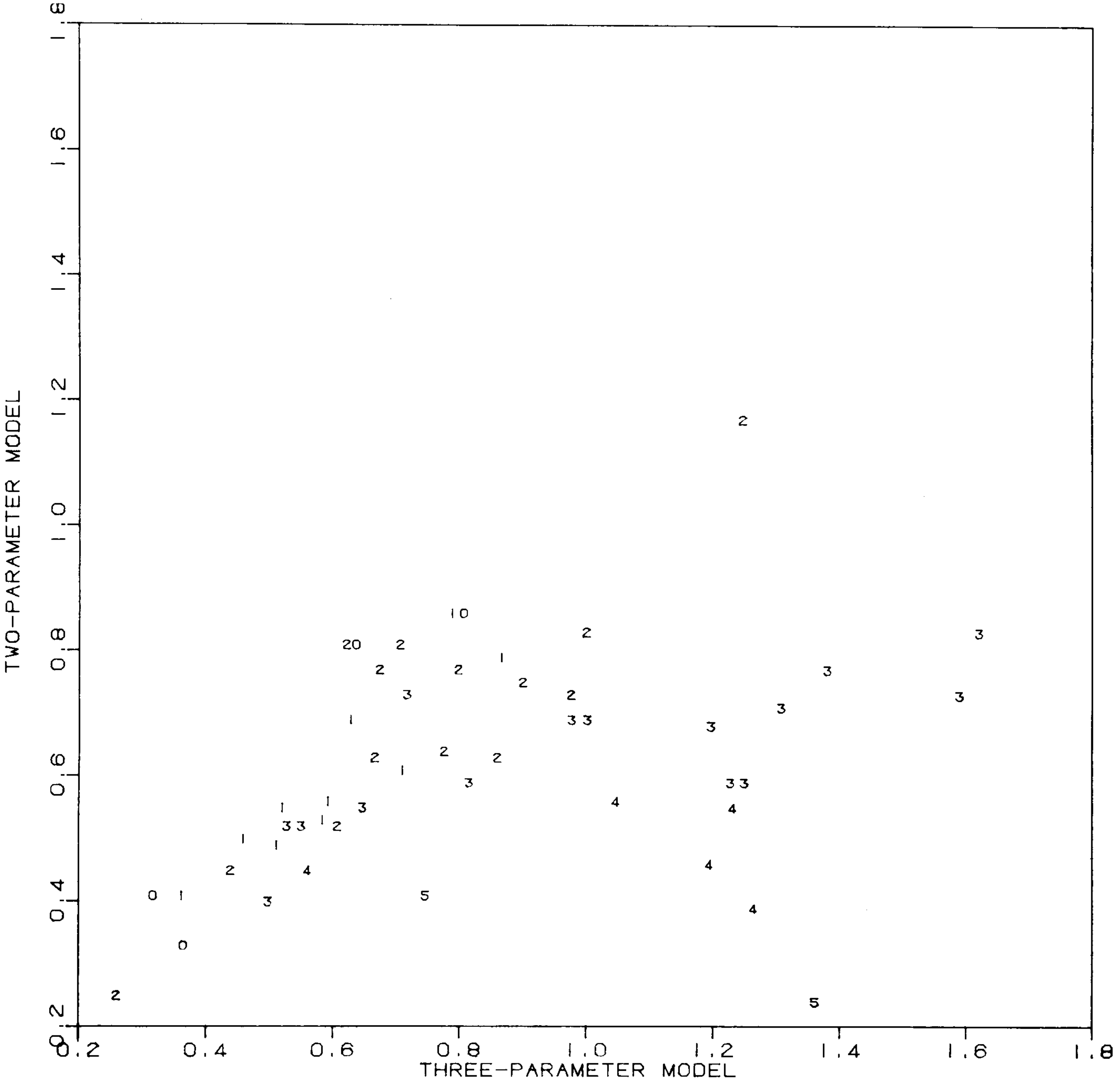

Figure 1. SCAT 2B. A comparison of estimated a_i parameters. The two-parameter model assumes a normal distribution of ability. Each item in the plot is located by a digit which represents item difficulty (b_i + 3). The easiest items are indicated by a 0, the hardest by a 5.

selected the best items from ten tests, we could build a single test that would be much more reliable than any of the original tests.

13. My last point concerns the use of Bayesian inference in adaptive testing. When we are testing large numbers of examinees all coming from a single source, we are in a really exceptionally good position to obtain and use a prior distribution describing the examinees. It would seem negligent not to obtain and use such a readily available prior distribution.

On the other hand, I would like to make a simple point not often expressed. Bayesian inference based on a prior distribution will give correct results when the prior corresponds, in some sense, to reality. It is likely to give incorrect results if the prior itself is incorrect.

In most Bayesian work, it is usually not practicable to determine whether the prior is correct or incorrect. In our work, on the contrary, it is fairly easy to do so. We need only estimate the ability of each person tested and then look at the distribution of estimated abilities.

If we were testing unselected school children in grade school, a normal distribution of ability might possibly be found. When we are testing highly selected groups in college or elsewhere, it seems unlikely that we will find a normal distribution.

Bayesians point out that the effect of an assumed prior becomes unimportant as the number of observations becomes large. In our context, this means that our ability estimates will not be spoiled by an incorrect prior distribution of ability provided the test administered is long enough.

This is not the whole story, however. The assumption of a normal distribution of ability, if false, may lead to unsatisfactory estimates of item parameters. The usual formula for biserial r can give absurd results if the continuous variable, in this case examinee ability, unknown to the statistician, is far from normally distributed. Unlike some other effects of Bayesian priors, this difficulty does not diminish as sample size becomes large.

Two different estimates of the distribution of examinee ability for one set of data are shown in Figure 2, reproduced here from Lord (1974). The agreement between the two estimates, obtained from very different assumptions, gives me some confidence in these results. My empirical results from other sets of data (including a representative sixth-grade group) are similar. When the ability scale is chosen so that all item characteristic curves are three-parameter normal ogives, or logistic curves, it turns out, for my data, that ability is not normally distributed.

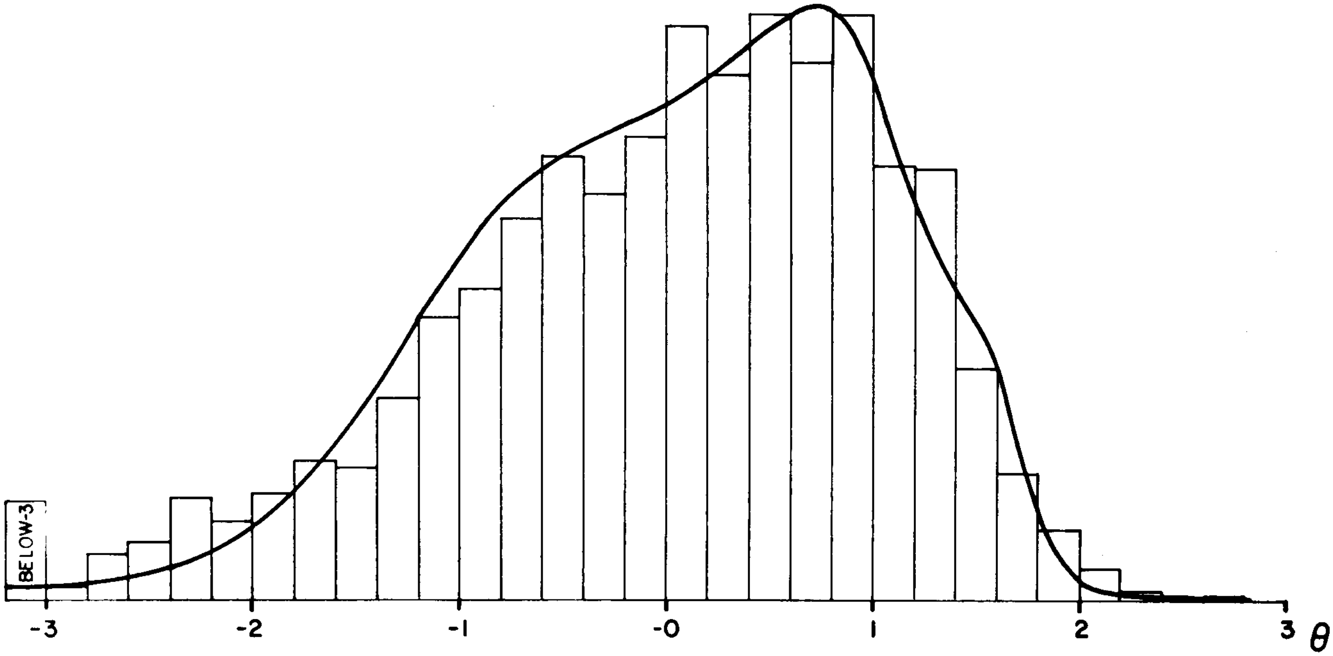

Figure 2. Distribution of estimated θ (histogram) and estimated distribution of θ (curve). Reproduced from Lord (1974) with permission of *Psychometrika.*

14. Although I an not a market analyst, I will without much risk venture two assertions. Computer costs–if they have not already done so–will come down to the point where computer-based adaptive testing is economical. When this happens, adaptive testing will come into wide use. The

McKillip and Urry paper provides important details on this subject.

REFERENCE NOTE

1. Cliff, N. *Complete orders from incomplete data: Interactive ordering and tailored testing.* Mimeographed paper. Conference on Computerized Adaptive Testing, Washington, D.C., June 1975.

REFERENCES

Bickel, P. J., & Yahav, J. A. Asymptotically optimal Bayes and minimax procedures in sequential estimation. *The Annals of Mathematical Statistics,* 1968, *39,* 442-456.

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

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

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 1975, in press.

Robbins, H. Sequential estimation of the mean of a normal population. In U. Grenander (Ed.), *Probability and statistics.* New York: Wiley, 1959.

Wald, A. Asymptotic minimax solutions of sequential point estimation problems. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1951.