DISCUSSION

Robert L. Linn . . . . . University of Illinois

I'd like to start by commending David Weiss and the group of people involved in the Psychometric Methods Program at the University of Minnesota for the continued high quality work on issues related to adaptive testing. This work is praiseworthy in several regards but, in my view, most notably for its continued and systematic nature and for the use of multiple approaches combining theoretical, simulation and empirical techniques. These aspects give the work a cumulative quality that is too often missing.

Thanks largely to the continued work on problems in adaptive (or tailored) testing by Fred Lord and the work at the University of Minnesota there is by this time a pretty good understanding of the potential value of adaptive testing techniques, at least under idealized conditions. The best of the adaptive testing procedures provide the promise of measurement that is nearly of equal precision throughout a wide range of ability with only a small loss compared to a peaked conventional test at an ability level equal to the difficulty location of the peaked test. David Vale's results support this conclusion and indicate that several techniques have relatively good potential.

There are a couple of general questions, however, that need to be kept in mind in drawing conclusions from results such as Vale's. One of these issues is implicit in the limitations noted by Vale at the end of his paper. That is, will the results based on an overly simple model generalize to real items and real examinees? Items do differ in discriminating power. Furthermore, items with equal discriminating power are not apt to be uniformly distributed over item difficulty. Also, multiple-choice items are the backbone of most standardized testing and such items generally require another item parameter for the lower asymptote. For these reasons I'd like to see more simulation studies that are based on estimated item parameters (preferably with three parameters per item) for actual pools of items.

Fred Lord has recently released an ETS Research Bulletin which not only shows some very promising work of this type but includes an offer to make available item parameter estimates based on the three-parameter logistic model for 690 items from some fifteen forms of the SCAT and STEP tests. I think that the exploitation of this pool of items and parameter estimates in future simulation and empirical studies could be a great help in moving our understanding of adaptive testing forward.

A second type of question that needs to be addressed in considering the implications of results such as were presented by Vale and by McBride is whether the gain is worth the extra effort and precision. This is a pragmatic question and the answer will undoubtedly depend on a number of considerations. Important among these considerations, however, is the purpose of the testing. The procedures considered are of value where accurate measurement over a wide range is important. For many testing purposes equi-precision is

not very important. For example, in selection for a particular institution precision is needed near the cut point, not over a broad range, and here the peaked test does very well.

On the other hand, there are situations where precision over a wide range is needed. Some of these were discussed by Wood in his review article in the Review of Educational Research. Examples are for tests used by a wide variety of institutions for many purposes such as a college admissions exam. Another area is where there is an interest in plotting trends (growth) over extended periods of time. In the latter situation, however, the comparisons to conventional procedures might be fairer if the possibility of using prior information was allowed not only for the adaptive procedures but for the conventional procedures. For example, Vale showed nice results for the stradaptive test with different starting levels. Why not use prior information to select different peaked or other conventional tests? This is done in crude form all the time on educational achievement test batteries that have, so-called, vertically equated tests appropriate for different grade levels. However, current vertical equating is not based on an adaptive testing model, and there is reason to believe that current vertical equating procedures are rather inadequate.

The remainder of my comments are mainly on the paper by Nancy Betz and, to a lesser extent, the one by James Sympson. I think there is a need for considerably more research of the type reported by Betz under the heading of Psychological Effects. Feedback effects are not a necessary part of a computer-administered test but are an obvious possibility. The possible effects of feedback on the measurement characteristics of the instrument are many and mostly unknown. One might postulate that an adaptive test would be a less frustrating experience for low ability examinees because they would encounter fewer difficult items. On the other hand, many tests are arranged with easy items toward the beginning of the test and progressively more difficult items later in the test. Thus low ability examinees might have a less frustrating experience as the result of the very easy items early in the conventional test than on an adaptive test with a single middle difficulty entry point.

The three-way interaction of race by feedback by order obtained by Betz is a tantalizing result. It clearly is one that is of sufficient potential importance to require replication. Assuming that the result can be replicated then many questions will need to be addressed, with the primary one being--Is the same trait measured under feedback and no feedback conditions?

In the second feedback study reported by Betz it was unclear to me why there was no group difference on the adaptive test. Doesn't this suggest a problem with the adaptive test?

The focus on feedback vs. no feedback is an example of looking at one of the components that Sympson wants to have separately evaluated. The logic of separating the components is good but there is also a possibility of interaction which requires evaluation of the composites.

I'd like to mention two other types of testing problems where adaptive procedures may be of value. One is in instructional uses of tests where frequent measures are needed for short-term dichotomous decisions. Adapting

test length to the examinee can yield savings in testing time. The second problem area is in multidimensional measurement problems where allocation of testing time for various dimensions might be adapted to the individual.

In summary, I would mainly like to encourage more work which uses as a base parameters that mirror existing item pools, using these both for simulation work and for corollary empirical work, and more efforts on psychological effects.

R. Darrell Bock . . . . . University of Chicago

As these excellent papers were being presented I made a few notes that I'll discuss in turn. Any discussant has to face the question of how much of the difficulties of the subject he is going to let the speaker assume away. In the case of David Vale's presentation, I find myself very reluctant to let him assume away the item heterogeneity and the possibility of correct responses due to guessing.

My experience has been that sets of items that are supposed to be homogeneous are often surprisingly heterogeneous. The Ravens Matrices Test, for example, is usually considered homogeneous and certainly scored as such. But David Thissen, one of the students at Chicago, has a paper to appear in Educational Measurement in which he reports an item analysis of the Ravens A, B, and C sections. He found that the discriminating powers of the items estimated in this procedure indicate that the main source of discrimination is a subset of items in Section B. They define the well-determined dimension underlying the test and the other items contribute little to it.

This is typical of the dilemma that may confront the test constructor: he has items whose discriminating powers vary a great deal, so that if he were to throw away items in order to obtain a set that is homogeneous in discriminating power, he would be in the embarrassing position of throwing away what appeared to be the best items.

I am uncertain how to deal with this problem. Perhaps it is actually risky to use these highly-discriminating items because their source of discrimination may be something peculiar to the particular data. The discriminations may be valid for the calibrating population, but not generalized to the population to which the test will be applied. If this is the case, we may be well advised to regard these highly-discriminating items with suspicion and to remove them, or at least to adjust downward the discriminating powers when estimating latent scores for subjects. The issue is difficult to resolve, but any proposals for test construction that assume them away cannot be considered ready for implementation.

The other matter is that of guessing. I am not enthusiastic about any solution to this problem that assumes all subjects are guessing in the same way. Willingness to guess is very much a personality characteristic that cannot be suppressed even by explicitly instructing all subjects to guess. Some will guess, but others will omit items rather than mark them randomly. If the item analysis procedure is based on the assumption that all subjects guess, then it will have to, in effect, assign random responses to omitted items. But such a practice

may so seriously degrade the information that the test gives about the subject that it is misinformation and is better ignored than included in the scoring procedure.

A better strategy is one in which there is an evaluation of the probability that a given subject is in fact guessing in his response to a given item. This is essentially the strategy taken by Michael Waller in dissertation work at Chicago recently reported in an Educational Testing Service Research Bulletin. On the basis of a provisional estimate of a subject's ability and a provisional estimate of the difficulty of the item, Waller sets up an objective rule for deciding whether or not that particular response should be deleted from the next stage of estimation of the item parameter and latent ability. This is very similar to the approach to data analysis advocated by Tukey, in which an observation is trimmed or censored if it is sufficiently improbable that it could have arisen from the main population being sampled. My preference would be to regard all item response data as potentially contaminated by random responding and to take steps in the analysis to distinguish, insofar as possible, between informative and non-informative item scores.

Turning now to McBride's paper, I found myself having difficulty accepting at face value the bias curves comparing the Bayesian procedure and the maximum likelihood estimate of test score. In order to obtain biases in the Bayes estimates like those shown in Figure 21, McBride must be assuming a normal prior distribution of ability, which in effect restricts the Bayes estimate, especially at the ends of the distribution. In that case the word "bias" seems unduly prejudicial--both the Bayes and maximum likelihood estimates are valid inferences from the data starting from different assumptions. To plot curves of these estimates conditional on ability, as in the graph, is unfair to the Bayes estimate since, in effect, it takes into account the assumed probability density at each point on the trait continuum.

I also think it is somewhat misleading to show the information supplied by the ordinary test score based on an item-sequential test administration procedure in which all subjects are expected to obtain the same score (but from items of differing difficulty). The information of such scores, which is itself an expected value, is actually zero everywhere. This seems a little too trivial to plot on a graph. The non-zero values shown presumably reflect the imperfect working of the sequential procedure.

In Sympson's paper, I certainly agree with him that there is a need, before plunging headlong into latent trait assumptions and models, to give considerable thought to the plausibility of the assumption that the behavior in question is under the control of an unobservable and continuous latent trait. There may be good reason to do so, but some sort of theoretical justification is needed.

Consider, for example, a vocabulary test. We could estimate vocabulary size in terms of a sample of the number of words that a person has available in his personal lexicon. That would be a perfectly objective, direct way of describing the trait. How does one then justify switching from that intuitively direct concept to an abstract conception of a latent verbal ability? A possible justification that I can think of is that, if vocabulary is to serve as an index of cognitive development generally, we might wish to think in terms of capacity for

acquisition of vocabulary as a developing latent trait of which personal lexicon is a consequence. If so, it is a measure of that continuously developing capacity that we're trying to capture, and the vocabulary itself is just a symptom of that growth. Another justification might be that the latent trait estimate has greater generality and power to predict and account for behavior in other areas. If so, some of that generality and power should be demonstrated and not merely assumed as is so often the case in theoretical presentations of the subject.

Concerning specific criteria for evaluating items, I wish that Sympson had not chosen to omit discussion of some of the preliminaries. I think that, at an early stage in working with an item domain, it is advisable to look at some form of factor analysis of the item intercorrelations. In the past there have been objections to this because of the type of correlation coefficient used. Phi coefficients introduce spurious "difficulty" factors and should be avoided. Tetrachoric correlation coefficients may give a non-positive definite matrix of correlations and thus rule out rigorous factor analysis with a statistical test of the number of factors (although an approximate analysis goes through without difficulty).

But recently, in the March 1975 issue of Psychometrika, Anders Christoffersson has published a technique for a general factor analysis of dichotomous data that overcomes all of these objections and is reasonably practical computationally. His procedure could be used at a first level to verify that the item domain is unidimensional, or to classify items according to dimension. Once the set of items is narrowed down to a unitary domain under the control of a single latent trait, then the psychometric procedures for estimating item parameters and trait values will provide good statistical tests of whether or not the latent trait model holds. There are tests that would distinguish between the homogeneous case where a Rasch model would apply and a heterogeneous case where a more general model--the normal ogive or logistic model--would be required. The technical procedures for making these decisions are in fact available in the form of likelihood ratio tests of alternative models.

Finally, a comment on Nancy Betz's paper. Bob Linn has already made a number of points about that paper, so I will pick up just one aspect of it. I strongly support the point of view that traditional testing is too limited in terms of the sources of information that it exploits in order to assess abilities. But I am also concerned that, in an effort to expand these sources by a shift to computer terminal test administration, we will cut off a very important area of item content, namely the graphics. Much graphic material--half-tone or color pictures, for example--cannot be presented even on CRT displays under computer control. But these visual and non-verbal ways of communicating information should nevertheless be part of the evaluation of ability.

In recent years it has become increasingly clear that cognition is by no means limited to verbal skills. It appears that, in most people's minds, there is going on simultaneously with verbal and logical reasoning based on semantic mediation, a kind of analogical, spatial, non-verbal reasoning capable of solving concrete problems without the aid of the semantic device. The evidence for this is in work that shows the relative specialization of the left and right hemi-

spheres of the brain--the left to semantic processes and the right to spatial or configural processes. Some of this work has been summarized by Jerre Levy in the Proceedings of the 32nd Annual Biology Colloquium.

This work should remind us that if we restrict ourselves to the written or spoken word, we may end up measuring just half the brains of our subjects! Rather than do that, we should demand some type of computer-controlled equipment that is capable of handling visual displays as well as verbal displays. We could imagine stand-alone equipment based on a small mini-computer and some sort of random-access slide file. Slides would be selected under computer control and projected as the item stimuli. We might also want to have auditory display, but I think for most purposes a random access slide file would be sufficient. Prototypes of such equipment already exists, but I do not believe they are available off-the-shelf at non-prohibitive prices.

As education becomes increasingly centered on the individual, it is not unreasonable to assume that a school of any size ought to have some such special equipment for individual evaluation of students. Students could then be sent to the facility for individualized testing under control of the mini-computer at any time that a question arises about their educational progress. If schools had such facilities, the possibility would be open for much more frequent individual testing and monitoring of student progress. That's the direction I would like to see educational testing move in the next decade.