Discussion: Session 3

Melvin Novick
University of Iowa

I shall discuss some general methodological issues that bear on the papers by Reckase, Kalisch, and Kingsbury and Weiss and also on previous papers presented, integrating into the discussion relevant points that have been made by Lord, Wainer, Samejima, Lumsden, and others.

The results that have been obtained in these papers are contradictory. There seems to be difficulty deciding whether or not adaptive testing is worthwhile with a Bayesian approach--which is related to the kinds of models that have been adopted and the kinds of statistical analysis that are being performed. Lord made an important comment about the metric in which a least squares analysis is performed; and although the suggestion he made in that context was very good, it opens up the question, which is the correct metric? Wainer's comments about robustness are also very important; indeed, some of the problems that we have had have resulted from allowing a few outliers to mar the analyses. An important part of my discussion will also bear on his comment, "Let's look at the ends, because it doesn't matter what's going on in the middle." Samejima's comments about dimensionality are crucial; and Lumsden's comment about the importance of choosing the statistical analysis for the particular decision at hand is central to my discussion.

I am absolutely delighted to see that everyone is using Bayesian methods: It is a dream come true. The realization of the dream, however, remains imprecise. Although Bayesian procedures are being used, the analyses are not all Bayesian which is part of the problem I hope to correct.

A brief discussion is in order about the development of pre-Bayes statistical theory and its application in a Bayesian decision theoretic context. First was Gauss's work on least squares, which led to a certain mean value as an estimate; this was followed by the Gauss-Markov theorem, which tied least squares with the normal distribution. At about the same time, La Place was working with absolute error loss, which is typically a better loss function than squared error, and in my judgment La Place deserves more credit than he has been given. Once the question, how to obtain an estimator, has been posed and considered in terms of the appropriate loss function, a whole new set of problems arises.

Even though absolute error loss may be better than squared-error loss, in some of the applications this places too much weight on those large discrepancies. In terms of mastery testing, for example, it does not matter if the person is three standard deviations from the criterion or four. Certainly, as Wainer has said, we do not want the analysis to be affected very much by that, particularly when it is recognized that the distributions are not normal but that there are all kinds of outliers and unusual data values. This is not a minor point. It affects all the analyses that are being done. A very careful look must be taken at the loss function in deciding whether the decision rule or even an estimator is any good. In my judgment, none of the loss functions that have been talked about at this conference are acceptable.

I did use threshold loss in papers published several years ago; but at that time, Bayesian methods with threshold loss were better than classical methods. Now there are better methods, and recent papers discussing more general loss or utility functions provide much more acceptable methods. In these papers the normal ogive is used as a utility function. This is a clear improvement over threshold utility. However, there is a Stage 3 in which a cumulative data distribution may be used--perhaps some other ogival forms--as a utility function.

One of the techniques that was used in a paper in an earlier session was to ascertain how adaptive testing improves reliability and squared-error loss. The difference between looking at a reliability and looking at a squared-error loss is that reliability forgets about any kind of bias. However, squared-error is actually irrelevant to a context in which a mastery decision or a selection is being made. This does not mean that I repudiate either classical test theory, which is built largely on mean-squared-error, or latent trait theory. Those methods are useful in certain contexts, e.g., when developing a test that is going to be used for a wide range of purposes, when interest is in discrimination across the whole range of ability and some overall measure is needed, and for the SAT and ACT tests. These methods are much less useful in the context in which there is a question of mastery or selection and one has a fair idea where that selection is going to be. Then, it is desirable to peak the tests at roughly that point, but it is also desirable to use a loss function, or better yet, a utility function that focuses on that point. Therefore, looking at questions of reliability and squared-error loss does not really address the question of the efficacy of the procedure in any real way.

On a related issue, there has been discussion on using Bayesian modal estimates or maximum likelihood estimates, which are, of course, also Bayesian modal estimates assuming a uniform prior distribution on a particular parameterization. These are appropriate only in terms of a zero-one loss function, a most unrealistic loss function in this context. Therefore, the analyses based on maximum likelihood or a Bayesian modal estimator may be unrealistic.

The dimensionality issue is crucial. Something like the reliability-validity paradox may, in fact, be occurring here, as Samejima suggests. It would not surprise me at all if we are dealing with a test that is multidimensional and a criterion that is almost certanly multidimensional. If this is true, and if a dominant trait is focused on, we may be building up reliability and not measuring the other traits that are essential in prediction. Thus, validity will suffer. The answer to this is probably to study the predictor and the criterion carefully and to define the factors or traits and see that each one is measured carefully.

Next, if least squares is to be used, which I do not really advocate, there is the question of what metric to do it in. Should it be done in a latent variable metric? This causes problems because computations sometimes do not converge. Should it be done in the true score metric, which is tighter? Although I do not know the answer to that question at present, the question should not be ignored.

The questions that need to be considered are (1) How much efficiency is being obtained? (2) Where is the efficiency being sought? and (3) What is the appropriate measure of efficiency? If a procedure is being designed to assist in selection near the top of the distribution or at some other criterion point, it really does not matter whether or not better estimation is being obtained away from that point. It is totally irrelevant to state that there is only a 5% increase in efficiency overall. A 50% increase could be obtained where needed, still averaging out to 5% overall. That would not be bad. This is a question of how the gains are computed, which, again, may be related to the question of robustness. We may be doing terrible things with some outlier; but if a testee is completely off the scale, perhaps it does not really matter, because a large error will not affect the decision.

The Owen procedure is a good Bayesian procedure: It does make some assumptions. Even though some of the assumptions that it makes may not be terribly well satisfied for the first one-half dozen items, improvements are possible, but that is not important. If any reasonable Bayesian procedure is used, a great deal will be gained from the Bayesian allocation. If a person is seated at a terminal, it may not be very significant whether he/she takes 5 items or 6 items. Thus, I am not so sure that the emphasis on variable stopping is important. Some rules could probably be worked out that, by and large, would provide good results if all testees were given a Bayesian allocated test of specified length and the decision were made at that point. The advantage would be that most of the inaccuracies in the approximations of the Owen procedure would be eliminated. If, indeed, the saving of one item, on the average, has a high pay off, then presumably someone would be willing to make a large investment to obtain the needed refinements.

Now, I should like to treat some specifics of the Reckase and Kingsbury and Weiss papers. In each paper there is an emphasis on the Wald Sequential Probability Ratio Test (SPRT). The original application of this method was that there was a production process in control with a certain error rate that was tolerable. The concern was that something had happened that seriously degraded production quality and it was desirable to identify the problem very quickly. Therefore, it was very reasonable to take a certain point hypothesis, a 3% error rate, with the recognition that if the process was not in proper working order, that error rate was going to go up to 10% and therefore the alternative hypothesis of 10% should be used. That paradigm is not correct in the context of adaptive testing. What the SPRT formulation gives is utility functions with three levels corresponding to the false positive, false negative, and indifference zones. In fact, an appropriate utility function would be continuous and not abrupt in change of magnitude of the first derivative (see Novick & Lindley, 1978). This is very important because it has a very substantial effect on the analysis, both in terms of the number of observations needed and in terms of the decision rule to be adopted.

A minor technical point is that one simply cannot look ahead one step, computing the cost of taking an observation and comparing this with the expected gain, and then stopping when it is discovered that the expected gain does not exceed the cost of the observation. In fact, all possible sample sizes would have to be investigated to make sure that none yielded an expected gain. I am not, however, arguing for this complication; indeed, I am arguing for a simplification to fixed sample sizes.

Finally, although there are a half a dozen other examples within an epsilon of the one I selected to discuss, Kingsbury and Weiss's paper presents the most simple and striking example of doing Bayesian analysis without a saturation of understanding Bayesian theory. The idea of looking at the Bayesian confidence interval, or as I would prefer to call it, the Bayesian credibility interval, and then stopping when that Bayesian interval no longer included a particular point is perfectly reasonable. In the context of mastery decision making, however, I cannot understand why a two-sided interval was computed. It makes no sense at all from any kind of Bayesian logic. Any consideration of a concept of utility or loss must lead to a one-sided interval. That struck me as being the most glaring failure to bring decision theory to bear on what is being done. If pressed, however, I could find a half a dozen more examples; and that, I think, is discouraging.

REFERENCES

Novick, M. R., & Lindley, D. The use of more realistic utility functions in educational applications. Journal of Educational Statistics, 1978, 15, 81-91.