DISCUSSION: SESSION 1

BRIAN WATERS
AIR UNIVERSITY

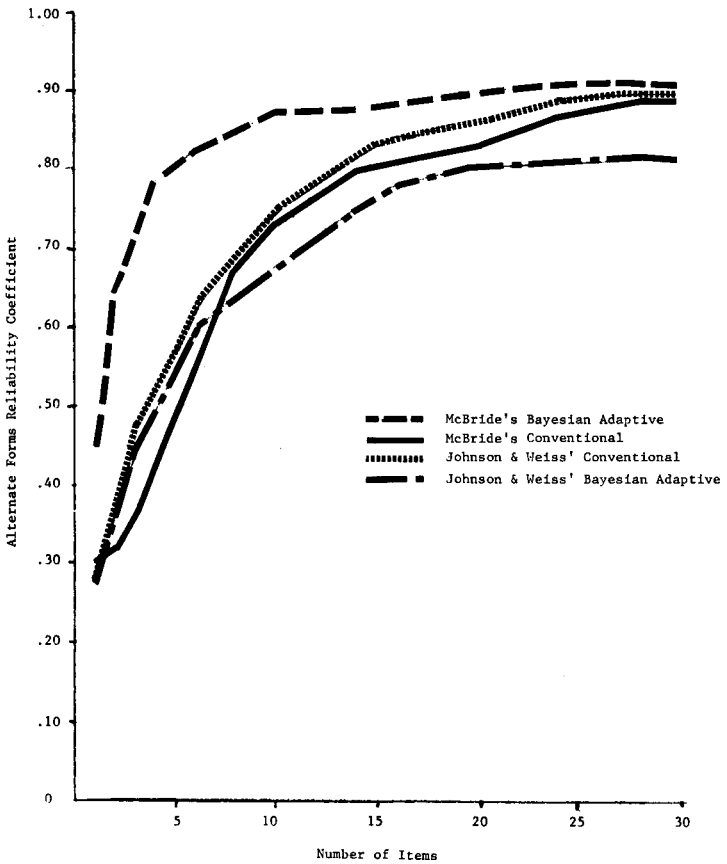
The Department of Defense enlists, classifies, and assigns hundreds of thousands of men and women annually, with test scores a major determinant of these decisions. The testing function must be performed more efficiently, accurately, and equitably; and computerized adaptive testing (CAT) provides the promise of greatly improved large-scale testing efficiencies. Work such as that reported by this session's authors on various adaptive testing strategies is therefore important.

These papers represent two lines of needed research--basic and more applied research on CAT. We still have many theoretical questions best addressed by simulation studies such as Gorman's, as well as myriad practical problems, which are best investigated with live data empirical studies such as McBride's and Johnson and Weiss's. I enjoyed reading each of these papers, particularly the mental exercise of analyzing the contradictory results of the latter two studies.

The primary result from the Johnson and Weiss paper and the McBride paper that caught my attention were the opposite results obtained on McBride's Figure 1 and Johnson and Weiss's Figure 8. These two analyses of Bayesian adaptive testing versus conventional testing both examined parallel forms reliability as a function of test length. McBride's results were consistent with the bulk of similar work done in the past, but the Johnson and Weiss results were startlingly different. The latter paper showed the conventional test yielding consistently higher reliabilities after about 10 items. In an effort to explain this difference in two similarly designed studies, Tom Warm of the Coast Guard Institute, Jim McBride, Marilyn Johnson, Brad Sympson, and I tried to determine what could have led to the conflicting results. Figure 1 shows a plot of the results from the two studies on comparable data. My tentative conclusions attribute the differences to either the parameterization process, the test difficulty, or the examinee characteristics differences. My best "guess" is that the former is the major cause of the contradictory results.

McBride designed an item pool that was extraordinary by any standards. In effect, he followed Urry's guidelines for selection of item characteristics for an adaptive test. All *a* parameters were more than .80 and all *c* parameters were less than .30. His average *a* values for the conventional and adaptive tests were 1.40 and 1.20, respectively. In addition, McBride's items were parameterized on a group of 4,000 examinees from a directly comparable population and produced a nearly rectangular distribution of information.

Figure 1
Alternate Forms Reliability Coefficients
from the McBride and Johnson and Weiss Studies



Johnson and Weiss had test item a parameters as low as .65, with a mean of 1.05 and a range of .65 to 2.25 on the conventional test. The adaptive test a parameter range, however, was .04 to 3.00, with a mean of .76. Particularly in the extreme ranges of ability, some of the items were adding practically no information to the adaptive test. The items were parameterized on far fewer examinees (82 to 1,861 with a median of about 300), and the item distribution was much more peaked. As McBride (paraphrasing Urry, 1970) stated in his paper, "a good tailored test design is superior [to conventional testing], provided that highly discriminatory test items are available." From a purely psychometric viewpoint, I would expect McBride's items to be more effective in an adaptive test as compared to a conventional test and to have more stable item parameter estimates than Johnson and Weiss's.

These contradictory results concern me in another way. Johnson and Weiss's data come from a much more "real world" situation. McBride's careful item selection, parameterization, and design are to be highly commended; however, in many applications, the "ideal" item pool he used is simply just not obtainable. Unfortunately, most of us will be faced with a pool more like that of Johnson

and Weiss. If, in fact, their results become typical, the practical application of adaptive testing is threatened. The Johnson and Weiss study thus needs replication.

McBride's study was exceptionally well done. It is nice to see data from the real world rather than from just "Psychology 101" students. I would have liked to have seen test statistics, including reliability, reported for the 50-item criterion test used in the validity analyses. McBride's results of a large increase in reliability with no significant change in validity is not atypical. More information on the criterion measure would have helped the reader conjecture why the validities did not increase with the reliability coefficients. My feeling is that it is related to the fact that the correlation coefficient only uses mean values and that the criterion measure was a conventional test score. If, as the errors of measurement suggest, the adaptive scores had less error variance and more true variance in them, then I would expect less correlation between adaptive and conventional scores than between two conventional scores. The additional true variance would be unique to the adaptive scores, whereas some of the error variance would be common, by chance, to the conventional test scores.

In a recent conversation with McBride, I discovered that since the conference he has acquired another criterion score on the examinees from this study. He reports that the validity coefficients on the adaptive tests were consistently higher (up to .19) than the conventional test validities, with the largest gain at shorter test lengths.

Before leaving these two papers, I would like to comment briefly on McBride's conclusion that fixed test length was as reliable as variable test length. I have a difficult time conceptually accepting this result, if for no other reason than that I believe that individual differences must make a difference. Practically, fixed length is certainly logistically and legally more realistic, which are perfectly valid reasons for using this testing strategy. Theoretically, however, I feel that potential efficiencies must exist with variable length. As Richard Anderson of the University of Illiniois has said, "You can't let bad data ruin a good theory."

Gorman's paper really consisted of two independent monte carlo simulation studies that followed up work suggested by McBride and Weiss (1976) and Urry (1977). It focused on the relative merits of two Bayesian models--the Owen algorithm and Samejima's Bayes modal procedure--and conventional rights-only scoring. Gorman's first study evaluated the efficiency of the two Bayesian models and conventional scoring on static (i.e., nonadaptive, or conventional) tests using three measures of efficiency: (1) average bias, (2) average accuracy, and (3) test score precision. He generated 2,000 simulated examinees (sims) from a normal distribution (mean 0, variance 1) and 80 item scores for each sim for both Bayesian and conventional sim group members. He then used ANCILLES to analyze the data.

Gorman's first study results showed considerably less bias of estimation for the two Bayesian procedures than for the conventional scoring at all points

on θ except at $\theta = -.5$ to $+.5$, with the Owen scoring generally better than the Bayes modal scoring at the lower θ levels and vice versa at the higher θ levels.

On his second measure of efficiency, conditional accuracy, again the conventional scoring yielded less accurate parameter estimation than the two Bayesian methods. Little accuracy differences between the latter two methods evolved, although the Owen procedure did show slightly more error than the Bayes modal model for most of the ability continuum.

Gorman's conditional test score precision measure showed substantial gains for both latent trait scoring models over conventional scoring, with nearly identical results between the two mathematical models. He also found statistically significant, though relatively small (.02), gains in fidelity coefficients in favor of the latent trait models. He concluded from his first study that measurement improvements can be realized through the use of the latent-trait-theory-based models to score static tests.

Gorman's second study was a follow-up of Urry's (1977) suggestion on McBride and Weiss's (1976) study results, which documented the regression to the mean effect using Owen's procedure. Urry suggested dividing the Bayesian regressed ability estimate by the test reliability (the Bayesian posterior variance squared). Gorman followed this procedure in a monte carlo simulation using ANCILLES, the revision of OGIVIA3, for evaluating the efficiency of the Bayes modal and the Owen models with the correction for regression applied.

Gorman's study results showed the Owen procedure to be generally preferable to the Bayes modal procedure in terms of conditional bias, conditional accuracy, and conditional precision when the correction for regression was used.

Considering the work performed on differences between the various computer program ability estimates, such as Bejar and Weiss (1979) showed for different maximum likelihood and Bayesian procedures, I am glad to see studies such as Gorman's being done. Somehow, we need to settle the arguments of the advantages and disadvantages of the various models whereby the results of each study are questioned by the proponents of other models. Algorithm comparisons with known parameters are an effective way to address this research question.

As a final observation on the subject of this session, I was very pleased to see two empirical live-data studies done. Although basic research is important, many of our funding agencies respond more to data from real people as opposed to simulees. I would suggest that future empirical studies include cost data in their battery of dependent variables. There has been a dearth of these data, and they have substantial impact on a funding agency's decisions. I recommend that proposals for future empirical adaptive testing studies should all include cost variables. In the competition for limited research dollars, this information could well be the difference between obtaining funding and not; but more importantly, the information is important for us as adaptive testing researchers.

REFERENCES

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1979. (NTIS No. AD A067752)

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Urry, V. W. A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Tailored Testing: A spectacular success for latent trait theory. Springfield, VA: National Technical Information Service, 1979.