Discussion

Mark D. Reckase
American College Testing Program

It is usually the role of a discussant to take a set of papers and show how they contribute to a larger coherent body of work. Unfortunately, in this case the contents of the two papers I am to discuss are so different that it is difficult to deal with them together. While they both deal with very important topics in psychometrics--the effects of multidimensionality on test results, and how to make decisions using test results--each paper is best considered separately from the other.

Weiss and Suhadolnik

The Weiss and Suhadolnik paper addresses the issue of the robustness of adaptive testing to the violation of the unidimensionality assumption required by most item response theory (IRT) models. The approach that they take is to generate simulated test data according to a multidimensional model, use that data in a unidimensional simulation of adaptive testing, and then compare the obtained θ estimates to the known θ values from the first dimension.

For studies of this type to be meaningful, the procedure used to generate the simulated item response data must be as similar to real item response data as possible. If it is not, the results of the study cannot be generalized to real testing situations. Of course, simulated test data never totally capture the richness of real test data, but we are usually willing to give up the richness in sources of variation in exchange for knowledge of the true ability of the examinees (simulees in Weiss and Suhadolnik's terminology). Whether the procedure used in this paper generates data that are "close enough" to real item response data is critical to interpreting the results of this paper. Therefore, I will concentrate most of my remarks on the data generation procedure.

Before addressing the data generation procedure directly, a definition of dimensionality is needed. To me, the dimensionality of a set of item response data is the result of a complex interaction between the characteristics of the examinees and the characteristics of the items. Examinees have many different abilities, the sum total of which define the complete latent space (Lord & Novick, 1968). These abilities may be related to each other in complex ways. The items also have many characteristics. They have different reading levels, different numbers of symbols, different lengths, and address different concepts. The response to an item is a function of the person's and item's characteristics:

$$x_{ij} = f(\underset{\sim}{\alpha}_i, \underset{\sim}{\theta}_j) \qquad [1]$$

where x_{ij} is the response to item i by person j,

$\underset{\sim}{\varsigma}_i$ is the vector of item characteristics, and

$\underset{\sim}{\theta}_j$ is the vector of person characteristics.

When simulating item response data, both $\underset{\sim}{\varsigma}_i$ and $\underset{\sim}{\theta}_j$ must be specified as well as the function that relates them to the response.

In this study, $\underset{\sim}{\theta}$ was generated so that the θs on each dimension were uniformly distributed between -3.2 and +3.2 and the dimensions were unrelated. In addition, for the first dimension, only 17 uniformly spaced values of θ were used, with 100 simulees at each value. This was done to allow for the computation of several indices of the quality of the adaptive testing procedure.

The initial question that comes to mind when considering this θ structure is, "Is it reasonable?" Certainly the θ distribution for actual groups of examinees is probably not uniform on any dimension, and most likely the dimensions should be somewhat related. However, for the sake of a uniform evaluation of the adaptive testing procedure, this part of the simulation is probably justified. However, the heavy "tails" of the distributions should be kept in mind when interpreting the results. The high number of high and low θs will probably result in more perfect and zero raw scores than usual.

The specification of the item characteristics for this study is much more complicated. The initial characteristics of the test items were borrowed from the factor structure of the General Science subtest of the ASVAB, presumably to ensure that they matched real items. In reviewing the structure (see their Table 1), I noticed that the loadings of the first factor tended to be inversely related to the item number. I have seen this type of pattern before when analyzing multiple-choice test items that were arranged in order of increasing difficulty. The reduction in factor loadings is due to the relative increase in guessing for the more difficult items (see Reckase, 1981, for examples using simulated test data). This is an important factor, because in order to develop the full item pool the factor loadings were reproduced six times and, I assume, they were not reordered. This means that the relationship between the factor loadings and the item difficulty was probably not maintained in the full 150-item adaptive testing item pool.

Using the 150 × 4 factor loading matrix derived by reproducing the matrix from the 25 ASVAB items six times, 45 different item pool structures were developed and item responses were generated. I will not discuss this process for all of the item pool structures, but a detailed analysis of several may prove informative.

Structure 1 is defined by the factor loadings from the first factor for the 150 simulated items. In order to generate dichotomous data to correspond to this structure, parameters were generated for each item for the 3-parameter logistic model. The a parameters were obtained from the following formula which was derived from the normal ogive model (Lord & Novick, 1968, p. 378):

$$a_{gj} = F_{gj}/(1 - F^2_{gj})^{\frac{1}{2}} \qquad [2]$$

where a_{gj} is the a parameter estimate for item g and factor j, and F_{gj} is the factor loading for item g on factor j. The b parameters were randomly sampled from a uniform distribution from -3.2 to 3.2, and the c parameters were randomly sampled from a normal distribution with mean equal to .20 and standard deviation equal to .02.

The three parameters for each item were used to compute the probability of a correct response for the first dimension. This probability was compared to a uniform random number in the 0-1 range to determine the dichotomous response. If the random number was greater than the probability, an incorrect response was generated; if it was less than the probability, a correct response was generated.

There are two possible problems with generating the data in this way. First, the difficulty and guessing parameters do not correspond to the a parameters, and second the a parameters were based on a normal ogive model, but were used in a logistic model. Since the normal ogive as are on a different scale than the logistic as, this will produce aberrant results.

When a two-dimensional dataset was produced, the loadings on the second factor were a multiple of the factor loadings on the first factor (except in one case when the actual ASVAB factor was used). For Dataset 2, the multiplier was selected so that the second factor accounted for 1/8 the variance of the first factor. The IRT item parameters were computed in exactly the same way for the second factor as for the first factor.

To generate the multidimensional data, probabilities of a correct response were computed separately for each dimension. These were then combined using a weighted average procedure, weighting the probabilities by the squared factor loadings. This process results in a compensatory model, but the compensation is on the probability metric, not on the θ metric as it is in some other models (McKinley & Reckase, 1983).

The characteristics of the data generated by this procedure are very difficult to predict. The smaller a values for the second dimension will tend to keep the probabilities computed for that dimension close to .5, but weighting the probabilities by the squared factor loading will tend to reduce the influence of the second factor. In effect, the impact of the second factor is being reduced twice, first from the reduced a value, and second from the weighting by the factor loading.

An important point to be made about this procedure is that the magnitude of effects of the factors can no longer be described in terms of proportion of variance accounted for. Such a description is only appropriate when dimensions are combined linearly on the factor score metric. In this case, the combination of dimensions is being done on the probability metric, a nonlinear transformation of the factor score metric.

The point of describing the data generation process in detail was to demonstrate how difficult it is to determine the characteristics of the data that were produced. The process has so many complexities that it is even difficult

to determine whether the resulting data are really multidimensional. It would have been informative if the authors had factor analyzed the simulated data so that the characteristics of the data could be determined. Until confidence can be gained in the data generation procedure, it is difficult to have confidence in the results of the study.

Weitzman

The Weitzman paper addresses a second important problem in the area of psychometrics: how to make decisions using test scores. In addressing this problem, he proposes a very intriguing variation of Wald's (1947) sequential probability ratio test (SPRT). Before discussing Weitzman's variation, I will first briefly describe Wald's procedure.

When it is desired to decide whether a person is above or below a particular cutting score, the SPRT requires that an indifference region be defined. This is an area on the score scale where it is a matter of indifference whether a person is classified as above or below the cutting score. The region is typically defined by its upper and lower boundaries, θ_1 and θ_0 respectively. The SPRT has error rates of α and β at the limits of the indifference region. The error rates are higher within the region and lower outside the region. The actual test statistic is the ratio of the likelihood of the observed responses $x_1, x_2, \ldots, x_n$ at the upper limit of the indifference region to the likelihood of the observed responses at the lower limit of the indifference region,

$$\frac{L(x_1, x_2, \ldots, x_n \mid \theta_1)}{L(x_1, x_2, \ldots, x_n \mid \theta_0)} . \qquad [3]$$

Weitzman proposes to do away with the indifference region by changing the form of the likelihood functions used. His test statistic can be formulated as

$$\frac{L(x_1, x_2, \ldots, x_n \mid \theta > \theta_c)}{L(x_1, x_2, \ldots, x_n \mid \theta < \theta_c)} \qquad [4]$$

where θ_c is the cutting score. His Equation 1 is a form of this statistic using quantiles to approximate the continuous functions involved.

At first glance, Expression 4 would seem to reach Weitzman's goal of eliminating the indifference region. However, the expression can be reformulated to show that it is equivalent to a classical SPRT with an indifference region.

The value of $L(x_1, x_2, \ldots, x_n \mid \theta > \theta_c)$ is essentially a weighted average of the $L(x_1, x_2, \ldots, x_n \mid \theta)$ for $\theta > \theta_c$ where the weights are the density of the distribution of θ at each θ value. Therefore, the value of the likelihood will be between $L(x_1, x_2, \ldots, x_n \mid \theta_c)$ and $L(x_1, x_2, \ldots, x_n \mid \theta \max)$. Thus, for some value of θ, θ', such that $\theta_c < \theta' < \theta \max$,

$$L(x_1, x_2, \ldots, x_n \mid \theta') = L(x_1, x_2, \ldots, x_n \mid \theta > \theta_c). \qquad [5]$$

Likewise a value of θ, θ'', can be found such that

$$L(x_1, x_2, \ldots, x_n \mid \theta'') = L(x_1, x_2, \ldots, x_n \mid \theta < \theta_c). \quad [6]$$

Expression 2 can then be rewritten as

$$\frac{L(x_1, x_2, \ldots, x_n \mid \theta')}{L(x_1, x_2, \ldots, x_n \mid \theta'')} \quad [7]$$

where θ' and θ'' are the limits of the implied indifference region. It is at these two points that the specified error rates will hold. Thus, Weitzman's procedure is not really any different than Wald's SPRT. The only difference is that he does not know the extent of his indifference region, while Wald shows how to specify the region. Also, according to the procedure specified, the region will shift after each item is administered rather than remaining constant throughout the decision-making process. It would probably be better to have control over the limits of the indifference region rather than allow them to float as a function of the items selected.

Despite the problems with the procedure, Weitzman's study does show the value of SPRT types of procedures in increasing the efficiency of decision making using test data. He has also demonstrated an item selection procedure based on classical test theory that is equivalent to the IRT approach I proposed earlier (Reckase, 1983). Both of these results are a valuable contribution to the research in this area.

REFERENCES

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

McKinley, R. L. and Reckase, M. D. An application of a multidimensional extension of the two-parameter logistic latent trait model (Research Report ONR83-3). Iowa City, IA: The American College Testing Program, Resident Programs Department, 1983.

Reckase, M. D. A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press, 1983.

Reckase, M. D. The formation of homogeneous item sets when guessing is a factor in item responses (Research Report 81-5). Columbia, MO: The University of Missouri, Educational Psychology Department, 1983.

Wald, A. Sequential analysis. New York: Dover Publications, 1947.