Olivier, P. *An evaluation of the self-scoring flexilevel testing model.* Unpublished doctoral dissertation, Florida State University, 1974.

Owen, R. J. *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton, New Jersey: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70*, 351–356.

Samejima, F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 1977, *42*, 193–198.

Urry, V. W. *A Monte Carlo investigation of logistic test models.* Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 1977, *14*, 181–196.

Weiss, D. J. *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974.

Weiss, D. J. Computerized adaptive ability measurement. *Naval Research Reviews*, 1975, *28*, 1–18.

Weiss, D. J., & Betz, N. E. *Ability measurement: Conventional or adaptive?* (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973.

Wood, R. L. Response-contingent testing. *Review of Educational Research*, 1973, *43*, 529–544.

12

A Procedure for Decision Making Using Tailored Testing

MARK D. RECKASE

There are many applications of testing technology that require decisions about whether a person is above or below a criterion score. Criterion-referenced testing and its special case, mastery testing, are examples of such a decision. In the criterion-referenced testing application, it would be especially useful if decisions could be made quickly and conveniently for each student in an individualized instruction program. The recently developed technology of tailored–adaptive testing has the potential to fulfill the requirements of such a testing system. There is no generally accepted procedure for making classification decisions using tailored testing, however, probably because these testing techniques are still relatively new. The few procedures that do exist are either based on randomly sampling items (Epstein, 1978; Sixtl, 1974), which does not take advantage of the power of tailored testing, or on heuristic techniques (e.g., see Chapter 13), which do not have a sound theoretical base. The purpose of this chapter is to present a decision procedure that operates sequentially and can easily be applied to tailored testing without loss of any of the elegance and mathematical sophistication of the examination procedures.

TAILORED TESTING PROCEDURES

Numerous tailored (i.e., adaptive, response contingent, sequential) testing procedures now exist in the research literature, ranging from simple two-stage procedures (Betz & Weiss, 1973) to complex Bayesian procedures (Owen, 1969; see Weiss, 1974, for a good review of the tailored testing procedures that were developed prior to 1974). Although many procedures exist, only tailored testing procedures using item response theory (IRT) and maximum likelihood ability estimation will be considered in this chapter. It will also be assumed that the tests are administered to the examinees on a computer terminal and that the items are selected to maximize the value of the information function at the previous ability estimate. Despite the narrow definition of tailored testing used in this chapter, the results should generalize to any procedure based on IRT.

In applying the decision procedure discussed in this chapter, two specific IRT models will be used: the one- and three-parameter logistic models. Although any other IRT model could just as easily have been used, these models were selected because of their frequent appearance in the research literature and because of the existence of readily available calibration programs (e.g., LOGIST and BICAL) and tailored testing programs (Reckase, 1974).

SEQUENTIAL DECISION PROCEDURES

A cursory review of the statistical literature indicates that much has been written about sequential estimation and classification procedures. Although somewhat more obscure than ANOVA and regression procedures, most intermediate-level mathematical statistics books include at least one chapter on sequential analysis (for example, see Brunk, 1965, chap. 16). In a review of the extensive literature on this topic, it has been found that most procedures fall into one of three categories: (*a*) sequential probability ratio tests (SPRT; Wald, 1947); (*b*) Bayesian sequential procedures (e.g., DeGroot, 1970); and (*c*) curtailed single sampling plans (Dodge & Romig, 1929). Of these procedures, only the SI RT is narrowly specified; the other two refer to families of procedures rather than a single technique.

Although these statistical procedures are widely applied for quality control, little use has been made of them in the area of mental testing, probably because operable sequential testing procedures did not exist until recently. To date, most references in the testing literature to sequential decisions have used the SPRT (Epstein, 1978; Reckase, 1978; Sixtl, 1974). The SPRT will therefore be described and studied here. Since the Bayesian procedures have not been fully developed for practical implementation, and the curtailed sampling plans cannot readily be applied to the commonly used tailored testing procedures, they will not be discussed in this chapter.

THE SEQUENTIAL PROBABILITY RATIO TEST

The sequential probability ratio test (SPRT) was initially developed by Wald (1947) as a quality control device for use by the Armed Forces during World War II. In addition to Wald's (1947) excellent book on the subject, Epstein (1978) also clearly described this procedure. The procedure will therefore be only briefly described here in order to generalize it so that it will more directly apply to tailored testing.

Application to Mastery Decisions

Wald originally developed the SPRT as a statistical test to decide which of two simple hypotheses is more correct. For example, it may be interesting to determine whether a student can answer correctly 60% or 80% of the items in an item pool. The basic philosophy behind the procedure used to decide between these two alternatives is to determine the likelihood of an observed response to an item under the two alternative hypotheses. If the likelihood is sufficiently larger for one hypothesis than for the other, that hypothesis is accepted. If the two likelihoods are similar, another observation is taken. Wald (1947) has shown that one hypothesis will always be selected over another after the administration of a finite set of items.

To demonstrate this procedure, suppose an item is randomly selected from an item pool and administered to a student. If a correct response were obtained, the likelihood under H_1 (80% knowledge) would be .80, and the likelihood under H_0 (60% knowledge) would be .60. To evaluate these likelihoods, Wald takes the ratio of the two,

$$\frac{L(x = 1|H_1)}{L(x = 1|H_0)} = \frac{.80}{.60} = 1.67. \qquad (12.1)$$

If the ratio is sufficiently large, H_1 is accepted; if it is sufficiently small, H_0 is accepted; and if it is near 1.0, another observation is taken. The values of this ratio that are considered sufficiently large or small depend on what is considered acceptable for the two possible decision errors: (*a*) accepting H_1 when H_0 is true (α error) and (*b*) accepting H_0 when H_1 is true (β error).

Although Wald (1947) developed a procedure for determining the exact values of these decision points, the procedure is very complex and is seldom used. Instead, good approximations can be determined using the following formulas:

$$\text{lower decision point} = B = \beta/(1-\alpha), \tag{12.2}$$

$$\text{upper decision point} = A = (1-\beta)/\alpha. \tag{12.3}$$

Thus, if the likelihood ratio is less than or equal to B, H_0 is accepted with error probability approximately β. If the likelihood ratio is greater than or equal to A, H_1 is accepted with error probability approximately α. If the ratio is between B and A, another item should be randomly sampled and administered and the decision rule implemented again. If $\alpha = .05$ and $\beta = .10$, for example, the decision points would be at $B = .105$ and $A = 18$. Since the likelihood ratio (1.67) in the above example is between these two values, no decision would be made, and another item would be selected and administered.

Since the responses to the items follow a binomial distribution in this example, a general expression for the likelihood ratio can be developed for the administration of n items:

$$\frac{L(x_1, x_2, \ldots, x_n \mid H_1)}{L(x_1, x_2, \ldots, x_n \mid H_0)} = \frac{p_1^{\Sigma x_i}(1-p_1)^{n-\Sigma x_i}}{p_0^{\Sigma x_i}(1-p_0)^{n-\Sigma x_i}} = \left(\frac{p_1}{p_0}\right)^{\Sigma x_i}\left(\frac{1-p_1}{1-p_0}\right)^{n-\Sigma x_i}, \tag{12.4}$$

where x_i is the score on item i (0 or 1), p_1 is the proportion of items in the item pool known by the student under H_1, and p_0 is the proportion in the item pool known under H_0. If

$$\frac{L(x_1, \ldots, x_n \mid H_1)}{L(x_1, \ldots, x_n \mid H_0)} \geq A, \tag{12.5}$$

accept H_1. If

$$\frac{L(x_1, \ldots, x_n \mid H_1)}{L(x_1, \ldots, x_n \mid H_0)} \leq B, \tag{12.6}$$

accept H_0. Otherwise, continue administering items.

This procedure was originally developed to test simple hypotheses, but Wald (1947) has shown that the procedure operates in the same way for composite hypotheses. For example, suppose it is desired to know whether a student knew more than some proportion p of the items in an item pool. In order to use the SPRT to make this decision, a region for which it does not matter which decision is made must first be selected around p, say, $p_0 < p < p_1$. If p_0 is close to p_1, a very precise decision is required. If p_0 and p_1 define a wide indifference region around p, a rather gross decision rule is all that is needed. The SPRT is then carried out in the same fashion as above, using p_0 and p_1 as the values for hypotheses H_0 and H_1, respectively. When the decision points A and B are computed as above, the error rates α and β hold for true values of p at p_0 and p_1. For true values of p more extreme than p_0 or p_1, the error rates are lower.

Evaluating Outcomes

In order to evaluate the properties of the SPRT, two functions have been derived: the operating characteristic (OC) function and the average sample number (ASN) function. The OC function is defined as the probability of accepting hypothesis H_0 as a function of the true proportion of the item pool known by the student. Although the derivation of the OC function is somewhat complex, the function can be approximated by the following two formulas:

$$p = \frac{1 - [(1-p_1)/(1-p_0)]^h}{(p_1/p_0)^h - [(1-p_1)/(1-p_0)]^h} \tag{12.7}$$

and

$$L(p) = \frac{[(1-\beta)/\alpha]^h - 1}{[(1-\beta)/\alpha]^h - [\beta/(1-\alpha)]^h}. \tag{12.8}$$

These equations are used by substituting various arbitrary values of h and solving for p and $L(p)$. $L(p)$, the probability of accepting H_0, is then plotted against p to describe the OC function. Figure 12.1 shows an OC function for $\alpha = .05$, $\beta = .10$, $p_0 = .6$, and $p_1 = .8$. Note that at $p = p_0$ the height of the curve is equal to $1 - \alpha$, and at $p = p_1$, the height of the curve is equal to β. Note that the OC function is dependent only on α, β, p_0, and p_1. Also, the steeper the curve, the more accurate the SPRT decision rule.

The ASN function is defined as the expected number of items required to make a decision at the various values of the true proportion of known items, $E(n \mid p)$. The formula for the ASN function for the binomial case described above is

$$E(n \mid p) = \frac{L(p) \ln B + [1 - L(p)] \ln A}{p \ln(p_1/p_0) + (1-p) \ln[(1-p_1)/(1-p_0)]}, \tag{12.9}$$

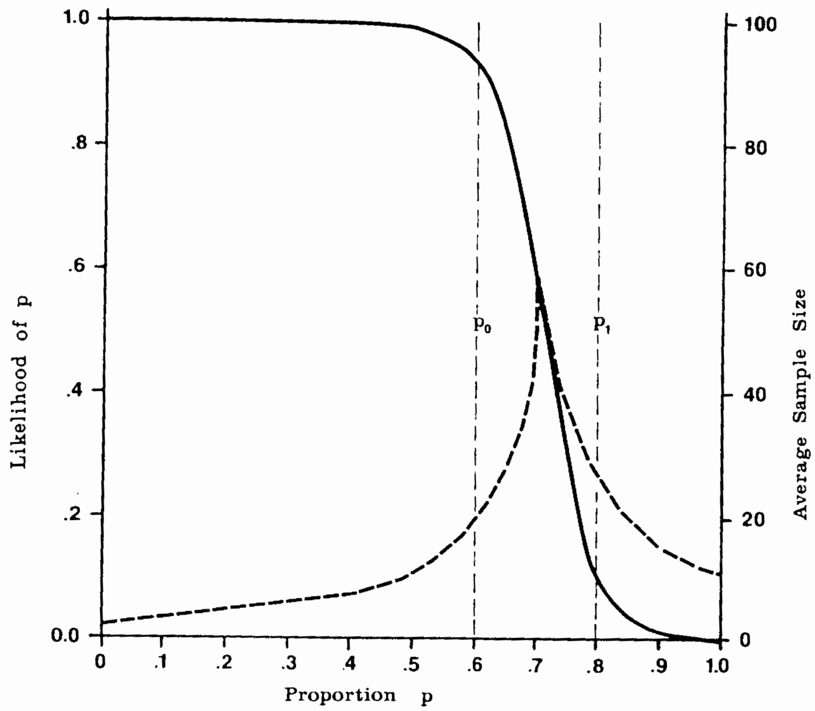

FIGURE 12.1. Example of the OC (solid line) and ASN (dashed line) functions.

where all of the symbols are as described above and the logarithms are to the base e. Figure 12.1 also shows the ASN function for the example presented above. Note that the ASN function is highest between the points p_0 and p_1 and that the closer together the values of p_0 and p_1 are, the higher the curve in that region. In general, the lower the ASN curve, the more efficient the decision rule since fewer observations are required, on the average, to make a decision.

Application to Tailored Testing

Although the SPRT as defined above is a valuable technique for decision making in many situations, it makes an implicit assumption that limits its usefulness for tailored testing. The technique as presented assumes that the probability of a correct response is the same for all items in the pool. This assumption is reasonable if items are randomly selected and p is the proportion of the items that a student can answer correctly, but it is not reasonable if items are selected to maximize information at an ability level. Under the tailored testing procedure assumed in this chapter, the probability of a correct response changes with each item, requiring a modification of the SPRT technique.

Fortunately, a detailed analysis of Wald's (1947) work indicates that the sequential random sampling assumption is not necessary for the application of the SPRT but is needed only for the derivation of the OC and ASN functions. The SPRT can be directly applied to tailored testing, but the OC and ASN functions must be determined in a different manner. One approach to determining these functions will be presented below.

To demonstrate the application of the SPRT to tailored testing as defined by this chapter, suppose that a tailored test is being used to determine whether a student has exceeded the criterion score specified for a criterion-referenced test. Although the method for selecting this criterion score is currently not well specified, assume that a value θ_c has been determined and that students above this value on the latent achievement scale pass the unit, while those below θ_c are given more instruction.

In order to use the SPRT, a region must be specified around θ_c for which it does not matter whether a pass or a fail decision is made. If high accuracy is desired for the decision rule, a narrow indifference region must be specified, but more items will be required to make the decision. As the region gets wider, the decision accuracy declines, but fewer items are required. Values on the ability scale, θ_0 and θ_1, mark the boundaries of this indifference region ($\theta_0 < \theta_c < \theta_1$). Once these values have been selected, the likelihood ratio can be defined as

$$\frac{L(x_1, \ldots, x_n|\theta_1)}{L(x_1, \ldots, x_n|\theta_0)} = \left(\prod_{i=1}^{n} P_i(\theta_1)^{x_i} Q_i(\theta_1)^{1-x_i}\right) \Big/ \left(\prod_{i=1}^{n} P_i(\theta_0)^{x_i} Q_i(\theta_0)^{1-x_i}\right), \quad (12.10)$$

where $L(x_1, \ldots, x_n|\theta_k)$, $k = 0, 1$, is the likelihood of the student's response string of n items administered so far; x_i is the 0, 1 score on item i; $P_i(\theta_k)$ is the probability of a correct response to item i assuming ability θ_k and the appropriate latent trait model; and $Q_i(\theta_k) = 1 - P_i(\theta_k)$.

If the one-parameter logistic model is used as a basis for the tailored testing procedure, Eq. (12.10) becomes

$$\frac{L(x_1, \ldots, x_n|\theta_1)}{L(x_1, \ldots, x_n|\theta_0)} = \left(\prod_{i=1}^{n} \frac{\exp[x_i(\theta_1 - b_i)]}{1 + \exp(\theta_1 - b_i)}\right) \Big/ \left(\prod_{i=1}^{n} \frac{\exp[x_i(\theta_0 - b_i)]}{1 + \exp(\theta_0 - b_i)}\right), \quad (12.11)$$

where b_i is the difficulty parameter for item i. Equation (12.11) can be simplified to

$$\frac{L(x_1, \ldots, x_n | \theta_1)}{L(x_1, \ldots, x_n | \theta_0)} = \left(\exp \sum_{i=1}^{n} x_i(\theta_1 - \theta_0)\right) \prod_{i=1}^{n} \frac{1 + \exp(\theta_0 - b_i)}{1 + \exp(\theta_1 - b_i)}. \quad (12.12)$$

The values of this likelihood ratio can then be used to test whether the student is above or below θ_c using the same method presented earlier. If the ratio is greater than $A = (1 - \beta)/\alpha$, the student is classified as being above θ_c; if it is below $B = \beta/(1 - \alpha)$, the student is classified below the criterion score; otherwise, another item is administered. If the three-parameter logistic model is the basis for the tailored testing procedure, the SPRT procedure is applied in the same manner as above, except that

$$P_i(\theta_k) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta_k - b_i)]}{1 + \exp[Da_i(\theta_k - b_i)]} \quad (12.13)$$

is used in Eq. (12.10) instead of the simple logistic form.

The evaluation of the OC and ASN functions cannot be performed as easily as for the simple binomial model because of the presence of the item parameters in the formula for computing the probability of a correct response. Since the item parameters for the next item to be administered are dependent on the item pool used and on the responses to the previous items, the derivation of these functions depends on a complex string of conditional expectations. The conditional probabilities involved make the derivation of these functions, for all practical purposes, impossible. Therefore, the OC and ASN functions can only be approximated using simulation techniques, but these approximations should be adequate for most purposes. Note, however, that although the full OC function cannot be derived, the value of the function is equal to $1 - \alpha$ at θ_0 and to β at θ_1, assuming that the item parameters are known. In reality, these two values of the function are not known either, since in all cases except simulations the item parameters are only estimated.

RESEARCH DESIGN

The purposes of this research were (*a*) to obtain information on how the SPRT procedure functioned when items were not randomly sampled from the item pool; (*b*) to gain experience in selecting the bounds of the indifference region, θ_0 and θ_1; and (*c*) to obtain information on the effects of guessing on the accuracy of classification when the one-parameter logistic model was used.

Tailored Testing Procedure

To determine the effects of these variables, the computation of the SPRT was programmed into both the one- and three-parameter logistic tailored testing procedures that were operational at the University of Missouri—Columbia. Since these procedures have been described in detail previously (Koch & Reckase, 1978), they will be merely summarized here. The programs implementing both models used a fixed stepsize method for branching through an item pool until both a correct and an incorrect response had been given. After that point, all ability estimates were obtained using an empirical maximum likelihood estimation procedure. Items were selected for both models to maximize the item information at the most recent ability estimate.

To evaluate the decision-making power of the SPRT, subjects with known ability were needed. Therefore, a simulation routine was built into the tailored testing program in place of the responding live examinee. At the beginning of each simulation run, the true ability of the simulated examinee was input into the program. This value was used to determine the true probability of a correct response to the administered items based on the model used (one- or three-parameter logistic) and the estimated item parameters. A number was then randomly selected from a uniform distribution in the range from 0 to 1. If the randomly selected number was less than or equal to the probability of a correct response, the item was scored as correct. If the randomly selected number was greater than the probability of a correct response, the item was scored as incorrect. This procedure continued for each item in the tailored test.

Tailored tests were simulated 25 times at each true ability level using different seed numbers for the random number generator. True abilities from -3 to $+3$ at .25 intervals were used for both the one- and three-parameter models to evaluate the performance of the SPRT. In addition, simulations were run on a composite procedure in which the tailored testing procedure and the probability ratio calculations [Eq. (12.11)] were based on the one-parameter model, but the item responses were determined by using the three-parameter model. This was done to determine the effects of guessing on correct classification using the one-parameter logistic model.

Criterion Values

In computing the probability ratios, three sets of limits of the indifference regions were used: ±.3, ±.8, ±1. A criterion of $\theta_c = 0$ was assumed in all cases. The ratios were computed after each item was administered, and the results were compared to an A value of 45 and a B value of .102. These were determined based on α = .02 and β = .10. A classification was made the first time these limits were exceeded. If the limits were not exceeded before 20 items had been administered (an arbitrary upper limit on test length), the values above 1.0 were classified as above θ_c and the values below 1.0 were classified as below θ_c. This is called a truncated SPRT. At each true ability used for the simulation, the proportion of the 25 administrations classified below θ_c and the average number of items administered were computed. Plots of these values against the true abilities approximate the OC and ASN functions, respectively. These plots were made for each combination of indifference region and tailored testing method, yielding nine plots of the OC and ASN functions.

Item Pools

Two different item pools were used for this study. For the analyses using just the one-parameter or the three-parameter model, an existing pool of 72 vocabulary items was used. This item pool had an approximately normal distribution of difficulty parameters. For the one-parameter tailored test using three-parameter responses, an item pool with 181 items, rectangularly distributed between −3 and +3 in difficulty, was used. These simulated items had constant discrimination parameters of .588 (this value yields a discrimination of 1.0 when multiplied by D = 1.7 in the exponent of Eq. (12.13)) and a pseudoguessing parameter of .12. This simulated item pool was selected over the real vocabulary pool to have better control over the guessing parameters. The one-parameter tailored testing procedure used only the b values from the pool, whereas the item responses were determined on the basis of all three parameters.

RESULTS

One-Parameter Model

Figure 12.2 shows the OC functions for the one-parameter logistic model based on the vocabulary item pool. The figure shows three graphs, one for each of the ±.3, ±.8, and ±1 indifference regions. Note that the curves are similar regardless of the indifference region. The data indicate that in all three cases the classification accuracy was nearly the same.

The values of the curves at the limits of the indifference region give further evaluative information. At the lower point the OC function should pass through 1 − α. At the −.3 value the curve is in fact .85 when it should be .98, showing the degrading effect of the restrictive stopping rule used by the tailored testing procedure. At the −.8 and −1 points for the corresponding curves, the results are about as expected, being .94 and 1.00 rather than .98.

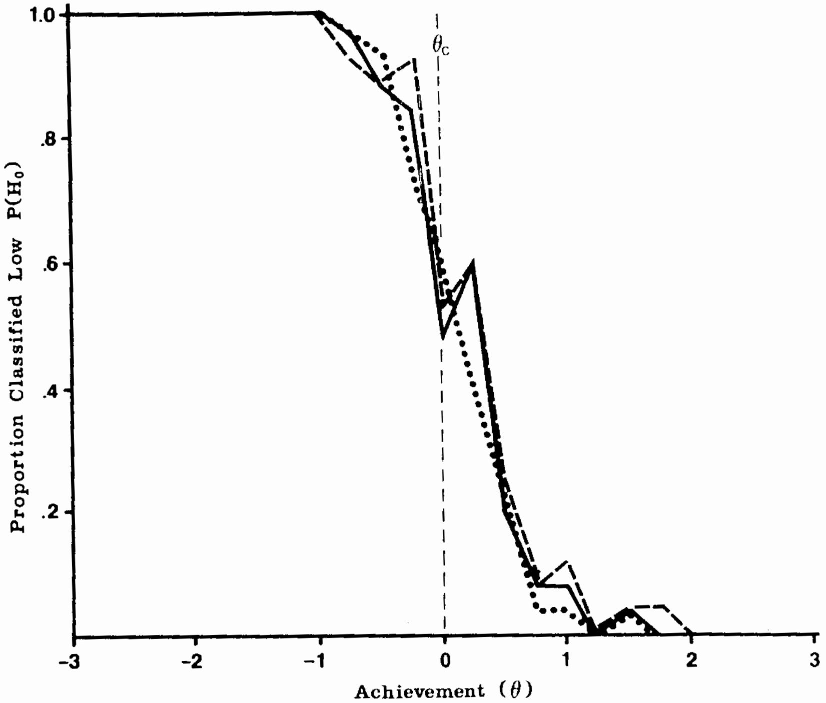

FIGURE 12.2. One-parameter OC functions for three indifference regions: solid line, ±.3; dashed line, ±.8; dotted line, ±1.0.

At the upper limit of the indifference region, the OC function should have a value of .1. For the +.3 case it is in fact .5 rather than .1, again showing the effects of truncating the procedure. At the values of +.8 and +1 the values of the OC function were near or better than what they should have been based on the theoretically expected results.

The ASN functions for the one-parameter model are given in Figure 12.3. The curves plotted correspond to the ASN functions using indifference regions of ±.3, ±.8, and ±1. It can immediately be seen that there was substantial variability in the average number of items needed to reach a decision, with the greatest number required when the indifference region was narrowest. It can also be seen that the largest expected number of

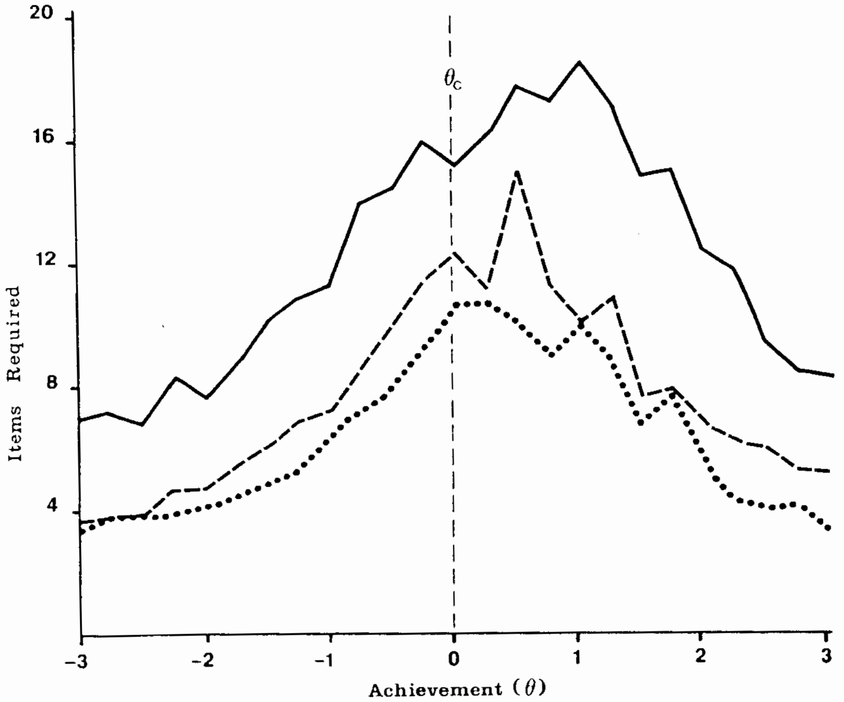

FIGURE 12.3. One-parameter ASN functions for three indifference regions: solid line, ±.3; dashed line, ±.8; dotted line, ±1.0.

items was near the criterion score 0.0 and that the average number dropped off at the extreme abilities. The slight lack of symmetry in the curves is due to the fact that α was not equal to β. For abilities beyond ±1, an average of only about 3–5 items was needed for classification for the wider regions, but 6–11 items were needed for the ±.3 indifference region. Note that the ±.3 curve approached the arbitrary 20-item limit for the tailored tests.

Figure 12.4 shows, for comparison purposes, the theoretical curves for the ASN and OC functions based on the ±.3 indifference region. An infinite number of items with difficulty 0.0 was assumed for the theoretical functions, and the tests were assumed to have no upper limit on the number of items administered. A comparison of Figures 12.2 and 12.3 with Figure 12.4 shows that the OC curve for the theoretical function is steeper at the cutting point than the simulated curves, and that the ASN function is substantially higher. The difference in the theoretical and simulated OC curves shows the effect of the 20-item stopping rule and the selection of items of differing difficulty.

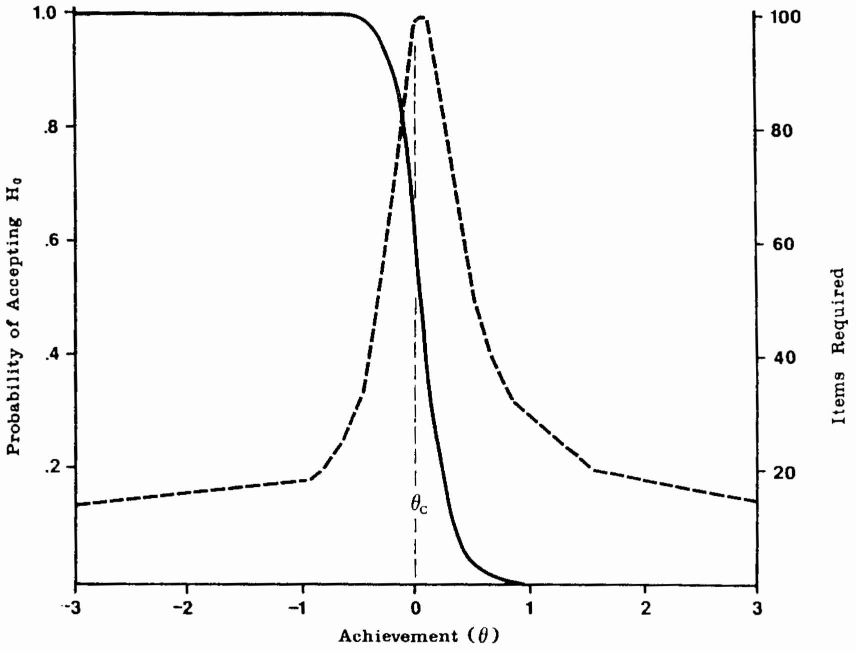

FIGURE 12.4. Theoretical OC (solid line) and ASN (dashed line) functions.

Three-Parameter Model

The results of the simulation of the three-parameter logistic tailored test are given in Figures 12.5 and 12.6. Figure 12.5 presents the OC functions for the three-parameter model, again using the indifference regions of ±.3, ±.8, and ±1. Note that, as with the one-parameter model, the OC curves are fairly similar for the three indifference regions throughout most of the range of ability. However, there are discrepancies for the ±1 indifference region curve near the +1 and −1 points, indicating a decline in decision precision for that region. At the −.3 value for the ±.3 indifference region, the value of the curve is .96, fairly close to the .98 theoretical value. At the upper end (+.3), however, the value is .2 instead of the .1 value that it should be. This may show the effects of guessing on the decision process. The ±.8 and ±1 indifference regions again yield better error probabilities than would be expected from the theory.

The ASN function for the three-parameter model (Figure 12.6) also shows similar results to those obtained from the one-parameter model. The ±.3 indifference region required the greatest number of items, whereas ±.8 and ±1.0 required about the same number. As before, the largest number was required near the criterion score. However, with the three-parameter model far fewer items, on the average, were required to make a decision than for the one-parameter model. Of special note is the ASN value of about 1.0 in the −1 to −3 range on the ability scale. Decisions seem to be possible with very few items in that range.

Because of the guessing component of the three-parameter logistic model, the ASN function tended to yield more asymmetric results than the one-parameter model. More items were required when classifying high than when classifying low to compensate for the nonzero probability of a correct response. Also, the ASN curve for the ±.3 indifference region was much more peaked than its one-parameter counterpart. If the simulated curves for the three-parameter model are compared to the theoretical curves presented in Figure 12.4, the OC functions can be seen to match the theoretical functions fairly closely, while the ASN functions show that substantially fewer items were required. Over much of the ability range, as many as 10 times more items were specified by the theoretical ASN curve when unlimited identical items were assumed. It should be noted, however, that the theoretical curves are based on the one-parameter model.

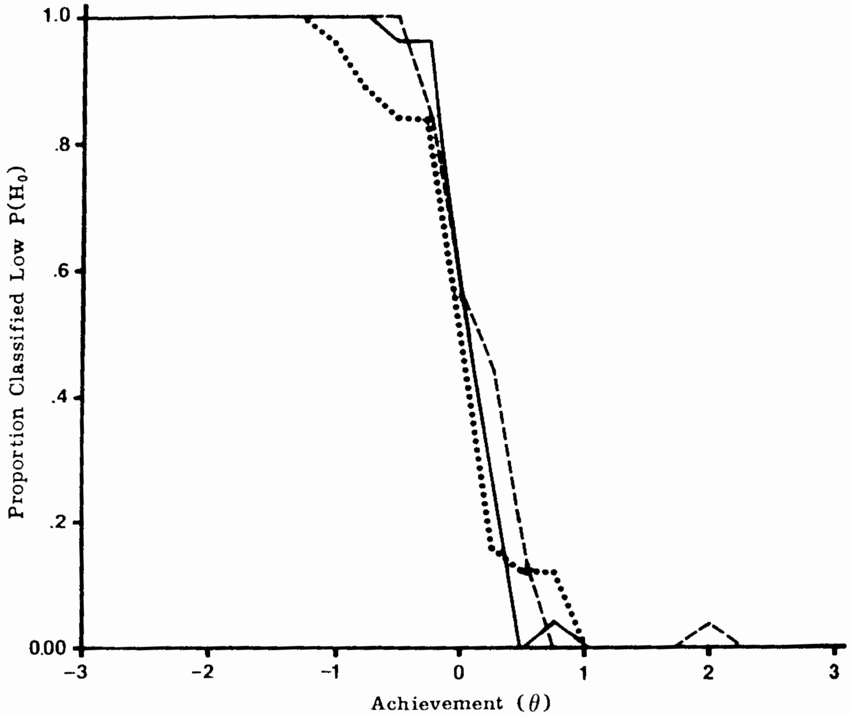

FIGURE 12.5. Three-parameter OC functions for three indifference regions: solid line, ±.3; dashed line, ±.8; dotted line, ±1.0.

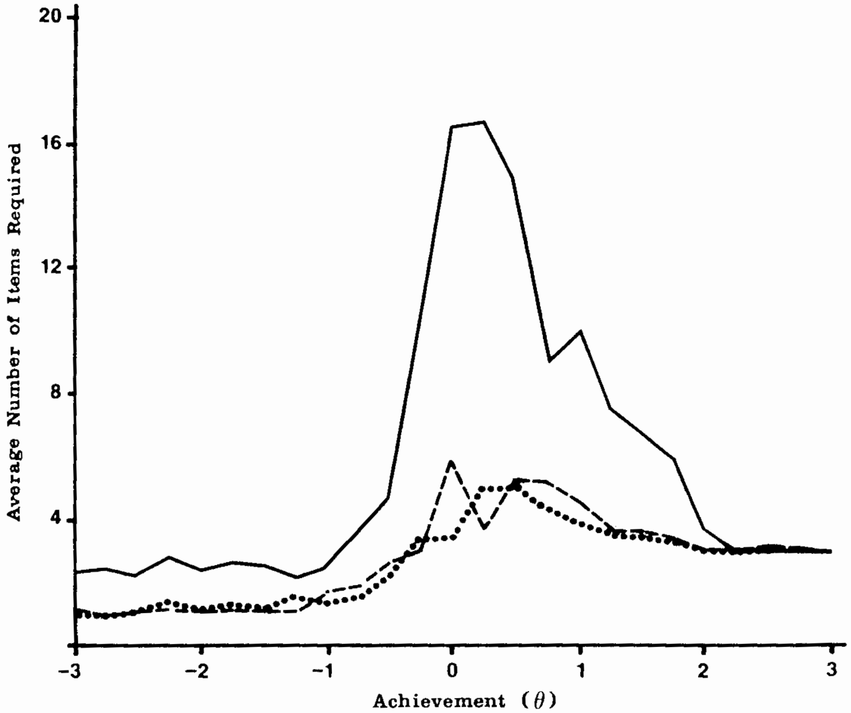

FIGURE 12.6. Three-parameter ASN functions for three indifference regions: solid line, ±.3; dashed line, ±.8; dotted line, ±1.0.

Effect of Guessing on the One-Parameter Model

Figure 12.7 shows the OC functions for the one-parameter model when the three-parameter model was used to determine the responses. The figure shows three graphs, one for each of the ±.3, ±.8, and ±1 indifference regions. Note that the curves are fairly similar regardless of the

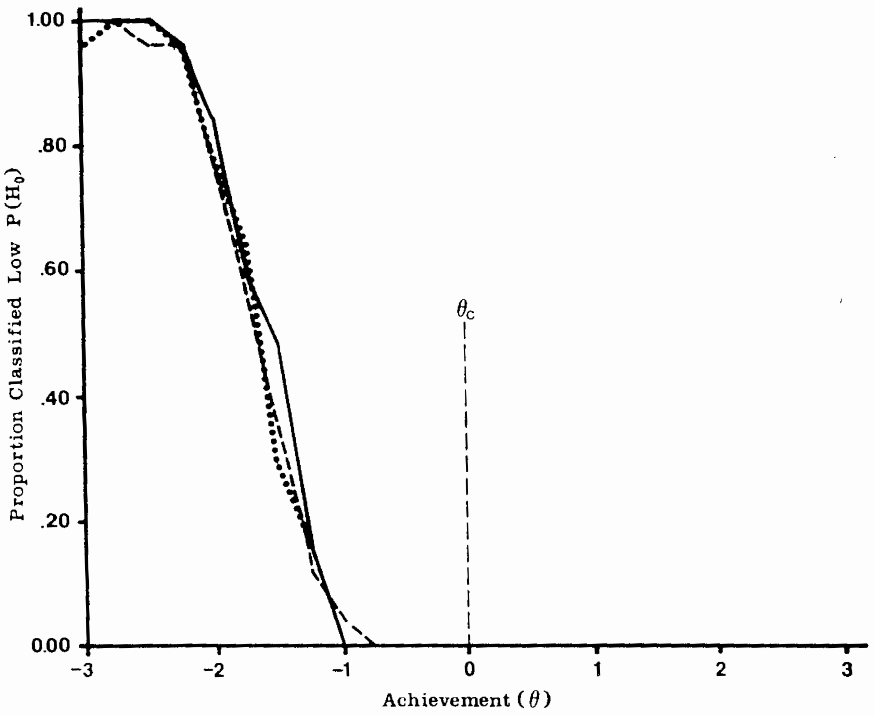

FIGURE 12.7. Composite OC functions for three indifference regions: solid line, ±.3, dashed line, ±.8; dotted line, ±1.0.

indifference region but that they are shifted substantially to the left compared to the previous OC curves. This indicates that the probability of classifying a person below θ_c has dropped off substantially until an ability of about −2 has been reached. In other words, it is much easier to be classified above the criterion score with this procedure than when guessing does not enter into the decision. Instead of being at zero, the effective criterion has been shifted down to −1.5. Clearly, the values of the OC function at the limits of the indifference region are entirely different from the theoretical values.

The ASN functions for the three indifference regions (±.3, ±.8, and ±1) are shown in Figure 12.8. The differences between these graphs and those presented in Figure 12.4 are that the curves are higher (more items were required) and the highest point of the curve is shifted to the steepest part of the OC curve. The relation between the height of the ASN function and the width of the indifference region still holds; however, as the region gets wider, the average number of items decreases.

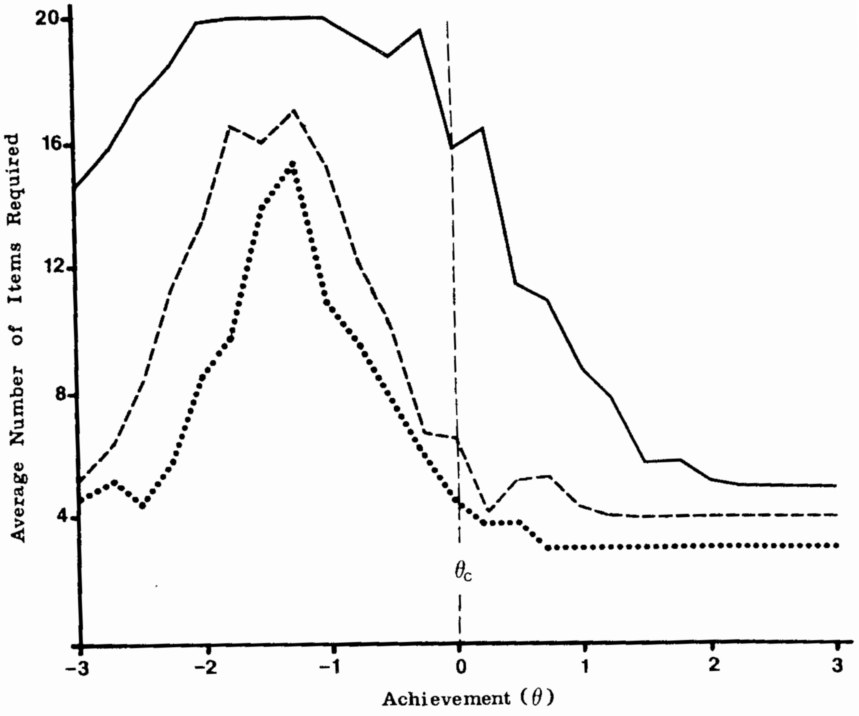

FIGURE 12.8. Composite ASN functions for three indifference regions: solid line, ±.3 dashed line, ±.8; dotted line, ±1.0.

SUMMARY AND CONCLUSIONS

The purpose of this study was to describe a procedure for making binary classification decisions using tailored testing in conjunction with the sequential probability ratio test (SPRT) and to present some simulation data showing the characteristics of the operation of the SPRT for two IRT models. The SPRT, which was developed by Wald for quality control work, has not been applied for tailored testing applications because the assumption of an equal probability of a correct response was made to facilitate the derivation of the operating characteristic (OC) and average sample number (ASN) functions. Since this assumption can be met for testing applications only by randomly sampling items for administration the procedure has not been used with tailored testing. In the present research the probability of a correct response was allowed to vary from item to item, although it made the derivation of the OC and ASN functions impossible. Simulation procedures were used to estimate these functions.

The results of the application of the SPRT procedure in three simulation studies were described. The first study estimated the OC and ASN functions for a one-parameter logistic-based tailored testing procedure in which the size of the indifference region around the criterion score was varied. The results showed that the average number of items needed for classification was quite low when the true ability of a simulee was not too close to the criterion score and that the width of the indifference region did not greatly affect the OC function. The width of the indifference region did have a substantial effect on the ASN function. The accuracy of classification of the simulated tailored test was not quite as good as administering a large number of items with difficulty values equal to the criterion score. This result was explained by the arbitrary 20-item limit imposed on the tailored test and by the variation in the difficulty parameters of the items administered.

The second study estimated the OC and ASN functions for a three-parameter logistic tailored testing procedure, also varying the size of the indifference region. The results were similar to those for the one-parameter model, but even fewer items were generally needed for classification. The results of these first two studies both indicated that the SPRT could be successfully applied to tailored testing.

The third simulation study estimated the OC and ASN functions for the one-parameter model when guessing was allowed to enter into the responses to the items administered. The results showed that, in effect, guessing lowered the criterion score, making it easier to classify an examinee above the criterion and raising the average number of items needed for classification. This spurious shift in the effective criterion score greatly increased the error rates in classification. Although this effect was exaggerated in this study because the difficulty parameter estimates for the one- and three-parameter models were assumed to be equal except for the constant D when in reality there would be some difference in the parameter estimates, the effect was strong enough to raise serious questions concerning the use of the one-parameter model for making classification decisions when guessing is a factor in item responses.

On the basis of the results of these three simulation studies, several conclusions can be drawn. First, the combination of tailored testing and the SPRT results in a practical decision-making procedure. Second, accurate decisions can be made based on the administration of relatively few test items. Third, procedures based on the three-parameter model require fewer items for making decisions than do procedures based on the one-parameter model. Finally, guessing has a seriously detrimental effect on the decision making procedure when it is based on the one-parameter logistic model. Overall, the tailored testing–SPRT combination shows substantial promise for testing applications. Future research should concentrate on verifying the findings reported here in live-testing situations.

ACKNOWLEDGMENT

This research was supported by Contract No. N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

REFERENCES

Betz, N. E., & Weiss, D. J. *An empirical study of computer-administered two-stage ability testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1973.

Brunk, H. D. *An introduction to mathematical statistics* (2nd ed.). New York: Blaisdell 1965.

DeGroot, M. H. *Optimal statistical decisions*. New York: McGraw-Hill, 1970.

Dodge, H. F., & Romig, H. G. A method of sampling inspection. *Bell System Technical Journal*, 1929, *8*, 613–631.

Epstein, K. Applications of sequential testing procedures to performance testing. In D. J Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Koch, W. R., & Reckase, M. D. *A live tailored testing comparison study of the one- and three-parameter logistic models* (Research Report 78-1). Columbia: University of Missouri, Tailored Testing Research Laboratory, June 1978.

Owen, R. J. *A Bayesian approach to tailored testing* (Research Bulletin RB-69-92). Princeton, New Jersey: Educational Testing Service, 1969.

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation*, 1974, *6* 208–212.

Reckase, M. D. *A generalization of sequential analysis to decision making with tailored testing*. Paper presented at the meeting of the Military Testing Association, Oklahoma City, November 1978.

Sixtl, F. Statistical foundations for a fully automated examiner. *Zeitschrift für Entwicklungspsychologie und Padagogische Psychologie*, 1974, *6*, 28–38.

Wald, A. *Sequential analysis*. New York: Wiley, 1947.

Weiss, D. J. *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Method Program, December 1974.