DOCUMENT RESUME

ED 181 627 TM 800 109

AUTHOR	Reckase, Mark D.
TITLE	Some Decision Procedures for Use with Tailored Testing.
PUB DATE	Jun 79
NOTE	40p.: Paper presented at the Meeting of Computer Assisted Testing (Minneapolis, MN, June, 1979)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Bayesian Statistics: *Criterion Referenced Tests: Decision Making: Guessing (Tests): *Mastery Tests: Measurement Techniques: Probability: *Psychometrics: Test Interpretation
IDENTIFIERS	Computer Assisted Testing: *Tailored Testing

ABSTRACT

This paper describes two procedures for making binary classification decisions using tailored testing: the sequential probability ratio test (SPRT) and a Bayesian decision procedure. The first procedure described, the SPRT, was developed by Wald for quality control work. It has not been widely applied for 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. The results of the application of the SPRT with a simulated procedure are described. The second decision procedure, the Bayesian procedure, includes a prior distribution of student achievement, a loss function for incorrect decisions, and the cost of observations in the development of the decision rule. The basic philosophy of this procedure is to administer items until the expected loss incurred in making a decision is less than the expected loss after the next item is administered plus the cost of administration. This procedure is not yet operational for making decisions under tailored testing because appropriate loss functions for educational decisions have not been determined. (Author/BW)

Some Decision Procedures for Use with Tailored Testing

by

Mark D. Reckase
University of Missouri-Columbia

There are many applications of testing technology that require that decisions be made as to whether a person is above or below a criterion score. Accepting a candidate into a program is an example of such a decision. Criterion-referenced testing and its special case, mastery testing, are other areas that require similar classifications. In the criterion-referenced testing application, it would be especially useful if the decisions could be made quickly and conveniently for each student in an individualized instruction program. The recently developed technology of tailored testing (Lord, 1970) has the potential to fulfill the requirements of such a testing system. However, no generally accepted procedure exists for making classification decisions using tailored testing, 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 (Weiss, 1978) that do not have a sound theoretical base. The purpose of this paper is to present some decision procedures that operate sequentially that can easily be applied to tailored testing without losing any of the elegance and mathematical sophistication of the examination procedures.

Paper presented at the meeting of Computer Assisted Testing '79, Minneapolis, June, 1979. This research was supported by Contract Number N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

Tailored Testing Procedures

Numerous tailored (adaptive, response contingent, sequential, etc.) testing procedures now exist in the research literature ranging from simple two-stage procedures (Betz and Weiss, 1973) to complex Bayesian procedures (Owen, 1969). Weiss (1974) has written a good review of the tailored testing procedures that have been developed up until 1974. Although many procedures exist, for the purposes of this paper only tailored testing procedures using item characteristic curve (ICC) theory and maximum-likelihood ability estimation will be considered. It is also assumed that the tests are administered to the examinees by computer using some type of computer terminal, and that items are selected to maximize the value of the information function at the previous ability estimate. Despite the narrow definition of tailored testing used for this paper, the results should generalize to any procedure based upon item characteristic curve theory.

In applying the decision procedures discussed in this paper, two specific ICC models will be used; the one- and the three-parameter logistic models. These models were selected because of their frequent appearance in the research literature and because of the existence of readily available calibration (LOGIST, CALFIT) and tailored testing programs (Reckase, 1974). Any other ICC model could just as easily be used.

Sequential Decision Procedures

A cursory review of the statistical literature quickly indicates that much has been written about sequential estimation and classification procedures. While somewhat more obscure than ANOVA and regression procedures, most intermediate level mathematical statistics books include at least one chapter on sequential analysis (see Brunk, 1965: Chapter 16 for example). In an ongoing review of the extensive literature that exists on this topic (which has accumulated over 200 references), it has been found that most procedures fall into one of three categories: sequential probability ratio tests (SPRT) (Wald, 1947), Bayesian sequential procedures (eg. DeGroot, 1970), and curtailed single sampling plans (Dodge and Romig; 1929). Of these procedures, only the SPRT 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. Since all references in the testing literature to sequential decisions discovered to date have used the SPRT (Sixtl, 1974; Epstein, 1978; Reckase, 1978), that procedure will be described first, followed by the Bayesian procedure. The curtailed sampling plans will not be discussed in this paper because they cannot be readily applied to the commonly used tailored testing procedures.

The Sequential Probability Ratio Test (SPRT)

The sequential probability ratio test was initially developed by Abraham Wald as a quality control device for use by the Armed Forces during World War II. Since he has written an excellent book on the subject (Wald, 1947) and since this procedure was clearly described at the last meeting of this conference (Epstein, 1978), the procedure will be only briefly described here. It is not the purpose of this paper to duplicate the efforts of Epstein, but rather to generalize the procedure so that it will more directly apply to tailored testing.

Wald originally developed the SPRT as a statistical test to decide which of two simple hypotheses is more correct. For example, it might 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 decida between these two alternatives was to determine the likelihood of an observed response to an item under the two alternative hypotheses. If the likelihood were sufficiently larger for one hypothesis than the other, that hypothesis would be accepted. If the two likelihoods were similar, another observation would be taken. Wald (1947) has shown that one hypothesis will always be selected over another using 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

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

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 upon 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) has 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:

$$\text{lower decision point} = B = \frac{\beta}{1 - \alpha} \qquad (2)$$

$$\text{upper decision point} = A = \frac{1 - \beta}{\alpha} \qquad (3)$$

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) 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:

$$\frac{L(x_1, x_2, \ldots, x_n|H_1)}{L(x_1, x_2, \ldots, x_n|H_1)} = \frac{p_1^{\Sigma x_i}(1-p_1)^{n-\Sigma x_i}}{p_0^{\Sigma x_i}(1-p_0)^{n-\Sigma x_i}} \qquad (4)$$

$$= \left(\frac{p_1}{p_0}\right)^{\Sigma x_i} \left(\frac{1-p_1}{1-p_0}\right)^{n-\Sigma x_i}$$

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

$$\frac{L(x_1, \ldots, x_n|H_1)}{L(x_1, \ldots, x_n|H_o)} \geq A, \text{ accept } H_1. \qquad (5)$$

If

$$\frac{L(x_1, \ldots, x_n|H_1)}{L(x_1, \ldots, x_n|H_o)} \leq B, \text{ accept } H_o. \qquad (6)$$

Otherwise, continue administering items.

Although this procedure was originally developed to test simple hypotheses, Wald (1947) has shown that the procedure operates in the same way for composite hypotheses. For example, suppose it was desirable to know whether a student knew more than some proportion, p_1, of the items in an item pool. In order to use the SPRT to make this decision, a region must first be selected around p for which it does not matter which decision is made--say $p_o<p<p_1$. If p_o is close to p_1, a very precise decision is required. If p_o 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 exactly the same fashion as above, using p_o and p_1 as the values for hypotheses H_o 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_o and p_1. For true values of p more extreme than p_o or p_1, the error rates are lower.

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_o 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.

$$p = \frac{1 - \left(\frac{1-p_1}{1-p_o}\right)^h}{\left(\frac{p_1}{p_o}\right)^h - \left(\frac{1-p_1}{1-p_o}\right)^h} \qquad (7)$$

$$L(p) = \frac{\left(\frac{1-\beta}{\alpha}\right)^h - 1}{\left(\frac{1-\beta}{\alpha}\right)^h - \left(\frac{\beta}{1-\alpha}\right)^h} \qquad (8)$$

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

Insert Figure 1 about here

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, $E(n|p)$. The formula for the ASN function for the binomial case described above is

$$E(n|p) = \frac{L(p) \ln B + (1-L(p)) \ln A}{p \cdot \ln \frac{p_1}{p_0} + (1-p) \ln \frac{1-p_1}{1-p_0}} \quad (9)$$

where all of the symbols are as described above and the logrithms are to the base e. Figure 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 the closer together the values of p_0 anc p_1 are, the higher the curve in that region. In general, the lower the ASN curve, the more efficient the decision rule.

Although the SPRT as defined above is a valuable procedure for decision making in many situations, it makes an implicit assumption that limits its usefulness for tailored testing. The model 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 model assumed by this paper, the probability of a correct response changes with each item, requiring a modification of the model.

Fortunately, a detailed analysis of Wald's (1947) work indicates that the sequential random sample 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 then 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 later.

To demonstrate the application of the SPRT to tailored testing as defined by this paper, suppose that a tailored test is being used to determine whether a student has exceeded the criterion specified for a criterion-referenced test. Although the method for selecting this criterion 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 of θ, θ_o and θ_1, mark the boundaries of this indifference region ($\theta_o<\theta_c<\theta_1$). Once these values have been selected, the likelihood ratio can be defined as

$$\frac{L(x_1, \ldots, x_n|\theta_1)}{L(x_1, \ldots, x_n|\theta_o)} = \frac{\prod_{i=1}^{n} P_i(\theta_1)^{x_i} Q_i(\theta_1)^{1-x_i}}{\prod_{i=1}^{n} P_i(\theta_o)^{x_i} Q_i(\theta_o)^{1-x_i}} \qquad (10)$$

where $L(x_1, \ldots, x_n|\theta_k)$, $k = 0, 1$, is the likelihood of the student's response string for the 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 determined from the appropriate ICC 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, Equation 10 becomes

$$\frac{L(x_1, \ldots, x_n|\theta_1)}{L(x_1, \ldots, x_n|\theta_o)} = \frac{\prod_{i=1}^{n} \frac{e^{x_i(\theta_1-b_i)}}{1 + e^{(\theta_1-b_i)}}}{\prod_{i=1}^{n} \frac{e^{x_i(\theta_o-b_i)}}{1 + e^{(\theta_o-b_i)}}} \tag{11}$$

where b_i is the difficulty parameter for Item i. Equation 11 can be simplified to

$$\frac{L(x_1, \ldots, x_n|\theta_1)}{L(x_1, \ldots, x_n|\theta_o)} = e^{\sum_{i=1}^{n} x_i(\theta_1-\theta_o)} \prod_{i=1}^{n} \frac{1 + e^{(\theta_o-b_i)}}{1 + e^{(\theta_1-b_i)}} . \tag{12}$$

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 $A = \frac{(1-\beta)}{\alpha}$, the student is classified as being above θ_c; if it is below $B = \frac{\beta}{(1-\alpha)}$, the student is classified below the criterion; 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 exactly the same manner as above, except

$$P_i(\theta_k) = c_i + (1-c_i) \frac{e^{Da_i(\theta_k-b_i)}}{1 + e^{Da_i(\theta_k-b_i)}} \tag{13}$$

is used in Equation 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 due to 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 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. Some OC and ASN functions for tailored tests based on the one- and three-parameter logistic models will be presented later in this paper. 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 β at θ_1, assuming that the item parameters are known. Since in all cases except simulations the item parameters are only estimated, in reality these two points are not known either.

Bayesian Sequential Decision Procedure

The Bayesian decision procedure is an alternative to the SPRT for deciding whether or not a student has exceeded the criterion, θ_c. Although this procedure is much more complicated than the SPRT, it has the capability of using additional information in making the decision. This added information may improve the decision process. In order to describe this procedure, some basic concepts will first be defined.

Initially, it is assumed that a population of students exists such that each student has some definable achievement level, θ. Individual achievement levels are labeled θ_i. Each person is to be tested and a decision is to be made concerning placement above or below the criterion. The decision to place above the criterion score is labelled d_1, and the decision to place below the criterion score is d_2.

In order to decide upon a decision rule using Bayesian methodology, three pieces of information are required in advance. These are (a) a prior distribution of θ, (b) a loss function relating the achievement levels to the decisions, and (c) the cost of each observation. Using these three types of information, a decision rule (technique for selecting a decision) and a stopping rule (technique for deciding when a decision should be made) can be determined.

The basic concept used in choosing a decision rule is the concept of risk. Risk is defined as the expected loss given a decision. Obviously, the decision that minimizes the risk is the desired one. When a Bayesian prior is used, this minimum risk is called the Bayes risk.

The stopping rule used with the Bayesian sequential decision procedure is also based upon the Bayes risk concept. If the expected risk after taking another observation plus the cost of the observation is less than the risk before the observation is taken, the sampling should go on. However, if the expected risk plus cost of a new observation is greater than the risk without the observation, then sampling should cease. In some cases, it is best not to take any observations at all because the expected risk plus the cost of an observation is greater than the initial risk of a guess based on the prior distribution of achievement.

Based on this framework, theorems have been proven that show that an optimal procedure exists, and that the optimal procedure will reach a decision after some finite number of observations (DeGroot, 1977). If the risk decreases with each observation, the procedure is called a regular sequential decision procedure. Only regular procedures will be considered here since it is assumed that each item administered yields some positive information rather than providing some misinformation.

In order to make the description of this procedure easier to follow, a simplified example will now be presented. Although this example is not realistic, it demonstrates the basic concepts without requiring complicated mathematical expressions. The extension of the procedure to realistic situations is direct, but the mathematics is cumbersome. Suppose that two types of individuals exist in the population of interest, those with $\theta_i = -.8$ and those with $\theta_i = +.8$ on a latent achievement dimension. A tailored test is to be used to classify the individuals into two groups--those above and those below the criterion score 0.0. Thus, two decisions are possible: classify as d_1, above the criterion; and d_2, below the criterion.

If persons with ability -.8 are classified above the criterion, a loss of 25 is incurred in each case. If they are classified below the criterion, there is no loss. If persons with ability .8 is classified above the criterion, there is no loss, while a loss of 15 is incurred for each person classified below the criterion. This loss function is summarized below. It should be noted that these loss function values are totally arbitrary.

Loss Function

	d_1	d_2
.8	0	15
-.8	25	0

Suppose that the prior belief that a randomly selected person has ability .8 is .6 and that he/she has ability -.8 is .4. Then the first step in using a Bayesian sequential decision process is to determine the risk associated with d_1 and d_2 when no observations are taken. The expected loss (risk) if decision d_1 is picked is

$$\begin{aligned} E(\text{loss}|d_1) &= P(\theta_1)\ell(d_1|\theta_1) + P(\theta_2)\ell(d_1|\theta_2) \\ &= .4 \times 25 + .6 \times 0 \\ &= 10, \end{aligned}$$

where $P(\theta_i)$ is the prior probability of θ_i anc $\ell(d_j|\theta_i)$ is the loss from picking decision d_j when θ_i is true. The expected loss (risk) if d_2 is picked is

$$\begin{aligned} E(\text{loss}|d_2) &= P(\theta_1)\ell(d_2|\theta_1) + P(\theta_2)\ell(d_2|\theta_2) \\ &= .4 \times 0 + .6 \times 15 \\ &= 9. \end{aligned}$$

Thus the Bayes decision when no observation is taken is d_2, and the Bayes risk is 9. The decision d_2 is obviously chosen because it has the lower risk.

Although the proper decision has been determined for the case when no observations have been taken, it has not been determined whether or not an observation should be taken. To do that, the expected risk after one observation plus cost must be compared to the Bayes risk without an observation. Determining the expected risk after an observation requires several steps, the first of which is determining the posterior distribution of ability after an observation.

Suppose that an item of 0.0 difficulty is administered to a person with ability .8 or -.8. Depending upon whether the response is correct or incorrect, a Bayesian posterior can be determined using Bayes theorem.

$$P(\theta_i|x) = \frac{P(x|\theta_i)P(\theta_i)}{\sum\limits_{i=1}^{2} P(x|\theta_i)P(\theta_i)} \qquad (16)$$

If a correct response is obtained to the item, the posterior probability of a .8 ability is given by

$$P(.8|x = 1) = \frac{P(1|.8)P(.8)}{P(1|.8)P(.8) + P(1|-.8)P(-.8)} \qquad (17)$$

The probabilities of an ability of .8 or -.8 were given in the prior distribution as .6 and .4 respectively. The probability of a correct response, given the known ability, can be determined from the appropriate ICC model. For example, using the one-parameter logistic model

$$P(1|.8) = \frac{e^{(.8-0)}}{1 + e^{(.8-0)}} = .69 \qquad (18)$$

while $P(1|-.8) = .31$. The posterior probability of .8 is then $P(.8|1) = .77$. Similarly, the posterior probability of -.8 is $P(-.8|1) = .23$. The posterior probability of the .8 and -.8 abilities given an incorrect response can likewise be determined using Equation 16. The posterior probabilities given an incorrect response are $P(.8|0) = .37$ and $P(-.8|0) = .63$.

The next step is to determine the risk using the posterior distributions just computed. If a correct response is obtained, the expected loss for d_1 is .23 x 25 + .77 x 0 = 5.75. The expected loss for d_2 is .77 x 15 + .23 x 0 = 11.55. Thus if a correct response is obtained, the Bayes decision is d_1 with a Bayes risk of 5.75. If an incorrect response is obtained, the expected loss for d_1 is .63 x 25 + .37 x 0 = 15.75, while the expected loss for d_2 is .37 x 15 + .63 x 0 = 5.55. Thus, after an incorrect response, d_2 is the Bayes decision with a Bayes risk of 5.55.

Since it is not known whether a correct or incorrect response will be given, the expected risk regardless of the response must be computed. To compute the overall expected risk, the probability of a correct and an incorrect response is needed. The probability can be obtained using the following formula:

$$\begin{aligned} P(1) &= P(1|.8)P(.8) + P(1|-.8)P(-.8) \\ &= .69 \times .6 + .31 \times 4 \\ &= .538 \\ P(0) &= 1 - P(1) = .462. \end{aligned}$$

The expected risk after a response can now be determined from

$$\begin{aligned} E(\text{risk}|\text{response}) &= E(\text{loss}|1)P(1) + E(\text{loss}|0)P(0) \\ &= 5.75 \times .538 + 5.55 \times .462 \\ &= 5.66. \end{aligned}$$

At this point, whether or not another observation should be taken can be determined. If the expected loss after an observation plus cost is greater than the risk before an observation, than administration of items should cease. If the risk before an observation is taken is greater, than another item should be administered. In the example given here, assume the cost of a response is 1 unit. The expected loss after a response plus cost is then 5.66 + 1 = 6.66. Since the Bayes risk with no items administered was 9, another item should be administered. Depending on the response to the item, decision d_1 or d_2 could be selected. After the item is administered, the apporpriate posterior becomes the new prior and the process continues as above. A flowchart of the entire decision process is presented in Figure 2 so that a more global picture of the steps involved can be obtained.

Insert Figure 2 about here

Although there are many postitive factors in the use of the Bayesian procedure, the very information that makes the control of the testing situation more precise also makes it difficult to initially implement. For example, specifying reasonable loss functions on the same metric as the cost of an observation is difficult for most educational applications. What is the cost of misclassifying persons below the criterion score when they really should be classified above it? Some attempts have been made by this author to specify loss functions for tailored testing applications, but no satisfactory results have been obtained so far.

A second difficulty in the application of this procedure is in specifying the prior distribution of achievement for a group. This is not as serious a problem as determining loss functions since performance data are usually available from previous groups. But of course, the more accurate the prior distribution, the more accurate the decision based on the procedure.

It should be realized that the procedure presented here is a simplification of a procedure that would be used for actual tailored testing applications. Achievement levels are usually continuous, rather than discrete as presented here, and the loss due to an incorrect decision is a function of the person's distance from the criterion score rather than a constant value. The procedure can also be modified by changing the cost of observations with increasing test length to allow for fatigue effects. Unfortunately, the Bayesian decision procedure as described here has not yet been implemented in conjunction with an operating tailored testing procedure. However, plans are being developed to evaluate an operational version at the Tailored Testing Research Laboratory at the University of Missouri.

Some Simulation Results for the SPRT

Before implementing the SPRT procedure described earlier in this paper, information was desired on how the procedure functioned when items were not randomly sampled from the item pool. Also, some experience was needed in selecting the bounds of the indifference region, θ_0 and θ_1. The effects of guessing on the accuracy of classification when the one-parameter logistic model was used was another area of interest.

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. These procedures have been described in detail previously (Koch and Reckase, 1978) so 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 previous 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 on 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.

Research Design

Tailored tests were simulated twenty-five times at each true ability 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 tailored test procedure and the probability ratio calculations (Equation 11) were based using the one-parameter model, but the item responses were determined using the three-parameter model. This was done to determine the effects of guessing on correct classification using the one-parameter logistic model.

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 twenty items had been administered (an arbitrary upper limit on test length), 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.

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 were 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 on difficulty was used. These simulated items had constant discrimination parameters of .588 (this value yields a 1.0 when multipled by D = 1.7) and psuedo-guessing 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 procedure used only the b-values from the pool.

Results

The results of the simulation studies will be presented in three parts; first the one-parameter SPRT, then the three-parameter SPRT, and finally the results of the combined simulations. Plots of the OC and ASN functions are presented to summarize the results of the SPRT for these models.

One-parameter model

Figure 3 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 reasonably similar regardless of the indifference region. The similarity indicates that in all three cases the classification accuracy is nearly the same.

Insert Figure 3 about here

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 effects of restrictive stopping rules 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.

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 4. The curves plotted correspond to the ASN functions using indifference regions for ±.3, ±.8, and ±1. It can immediately be seen from the graph that there is a substantial difference in the average number of items needed to reach a decision, with the greatest number required when the indifference region is narrowest. It can also be seen that the largest expected number of items is near the criterion score of 0.0 and that the average number drops 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 to 5 items was needed for classification for the wider regions, while 6 to 11 were needed for the ±.3 indifference region. Ncte that the ±.3 curve is approaching the arbitrary twenty item limit for the tailored tests.

Insert Figure 4 about here

Figure 5 shows the theoretical curves for the ASN and OC functions based on the ±.3 indifference region for comparison purposes. 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 3 and 4 with Figure 5 shows that the OC curve for the theoretical function is steeper at the cutting point than the simulated curves, and 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.

Insert Figure 5 about here

Three-parameter model

The results of the simulation of the three-parameter logistic tailored test are given in Figures 6 and 7. Figure 5 presents the OC functions for the three-parameter model, again using the indifference regions of ±.3, ±.8, and ±1. Notice 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 ±1.0 indifference range curve near the +1 and -1 points, indicating a decline in decision precision for that region. At the -.3 value for the ±.3 indifference range, 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 6) also shows similar results to those obtained from the one-parameter model. The ±.3 indifference region required the greatest number of items, while ±.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.

Insert Figures 6 and 7 about here

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 for classifying low to compensate for the non-zero 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 5, 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 ten times more items were specified by the theoretical ASN curve when unlimited identical items were assumed. However, it should be noted that the theoretical curves are based on the one-parameter model.

Effect of guessing on the one-parameter model

Figure 8 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 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 using this procedure than when guessing does not enter into the decision. The effective criterion has been shifted down to -1.5 instead of being at zero. Clearly the values of the OC function at the limits of the indifference region are entirely different from the theoretical values.

Insert Figure 8 about here

The ASN functions for the three indifference regions, ±.3, ±.8, and ±1, are shown in Figure 9. The difference between these graphs and those presented in Figure 4 are that the curves are higher (more items are required) and the highest point of the curve is shifted over to the steepest part of the OC curve. The relationship 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.

Insert Figure 9 about here

Summary and Conclusions

The purpose of this paper has been to describe two procedures for making binary classification decisions using tailored testing, the sequential probability ratio test (SPRT) and a Bayesian decision procedure, and to present some simulation data showing the characteristics of the operation of the SPRT for two item characteristic curve models. The first procedure described, the SPRT, was developed by Wald for quality control work.

It has not been widely applied for 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 only be met for testing applications by randomly sampling items for administration, the procedure has not been used with tailored testing. In this paper, 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 then used to estimate these functions.

The SPRT procedure described is operational at the Tailored Testing Research Laboratory of the University of Missouri-Columbia in two forms: a live tailored testing procedure and a simulated procedure. The results of the application of the simulation procedure to three studies were described in this paper. 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 of the study showed that the average number of items needed for classification was quite low when the true ability of a simulated person was not too close to the criterion socre, 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 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 guessing in effect 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 criterion greatly increased the error rates in classification. The effect is strong enough to preclude the use of the one-parameter model for classification decisions when guessing is a factor.

The second decision procedure described in this paper allows the use of a greater amount of information in making a decision than the SPRT. The Bayesian procedure includes a prior distribution of student achievement, a loss function for incorrect decisions, and the cost of observations in the development of the decision rule. The basic philosophy of this procedure is to administer items until the expected loss incurred in making a decision is less than the expected loss after the next item is administered plus the cost of administration. At that point a decision is made that minimizes the expected loss. The Bayesian procedure is described in detail and a simple example is given of its use. The Bayesian procedure is not yet operational for making decisions under tailored testing because appropriate loss functions for educational decisions have not been determined. However, simulation studies of the procedure will commence in the near future.

Both of the decision procedures described in this paper show promise for use in tailored testing. Both also require substantial research effort before they can be applied with confidence. It is hoped that this paper will help to stimulate that research.

References

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

Brunk, H. D. An introducation to mathematical statistics (2nd Ed.). New York, Blaisdell, 1965.

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

Dodge, H. F. and 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. Proceedings of the 1977 Computerized Adaptive Testing Conference. University of Minnesota, Minneapolis, Minn.: July, 1978.

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

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance, New York: Harper and Row, 1970.

Owen, R. J. A Bayesian approach to tailored testing. Princeton, N. J.: Educational Testing Service, Research Bulletin RB-69-92, 1969.

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.

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.

Sixtl, F. Statistical foundations for a fully automated examiner. Zeitshrift für Entwicklungspsychologie und Pägagogische Psychologie. 1974, 6(1), 28-38.

Wald, A. Sequential Analysis, New York: Wiley, 1947.

Weiss, D. J. Presentation at the ONR Contractors meeting. University of Missouri, Columbia, MO: September, 1978.

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

FIGURE I

EXAMPLE OF THE OC AND ASN FUNCTIONS

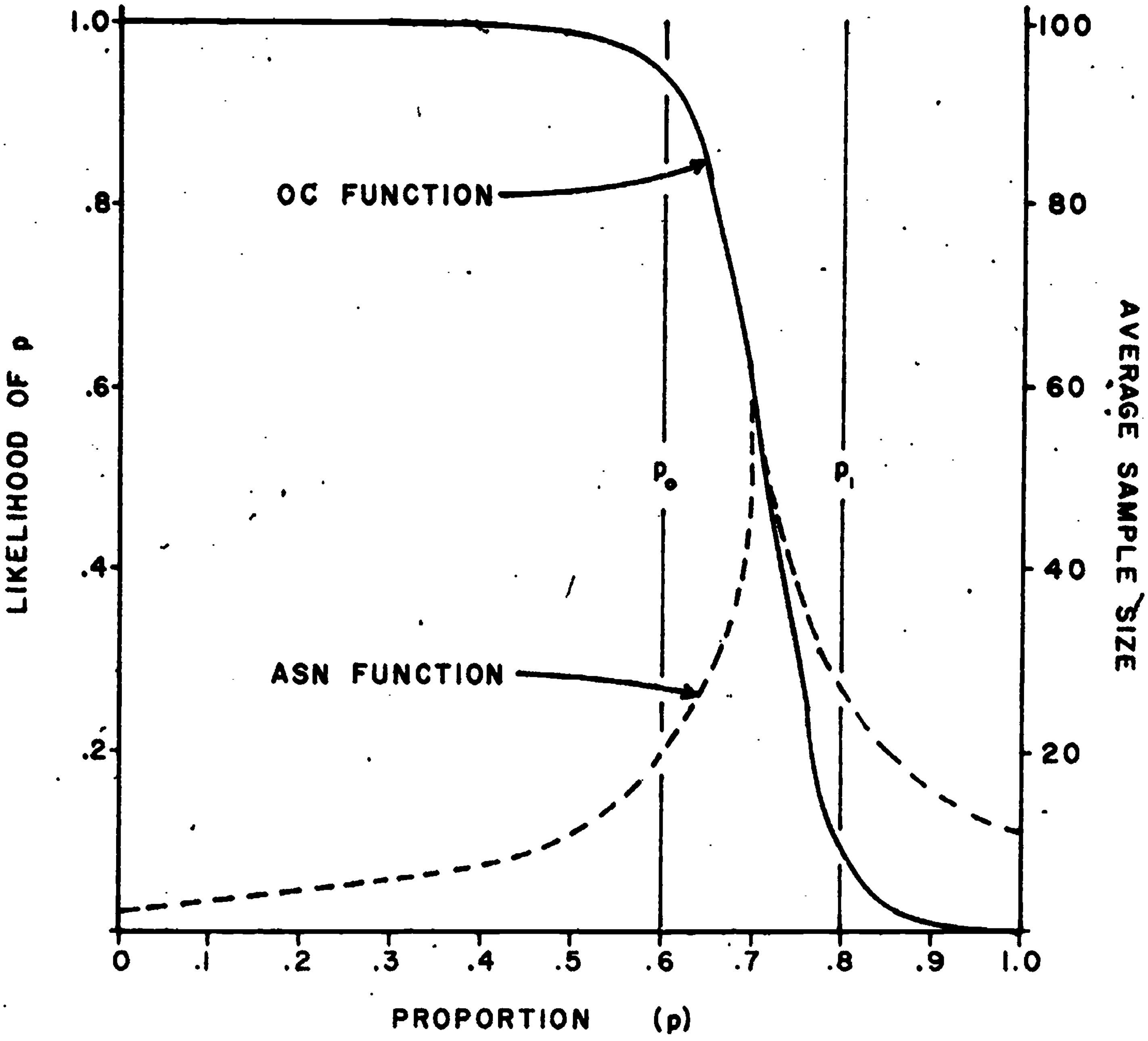


FIGURE 2
FLOWCHART OF BAYESIAN DECISION PROCESS

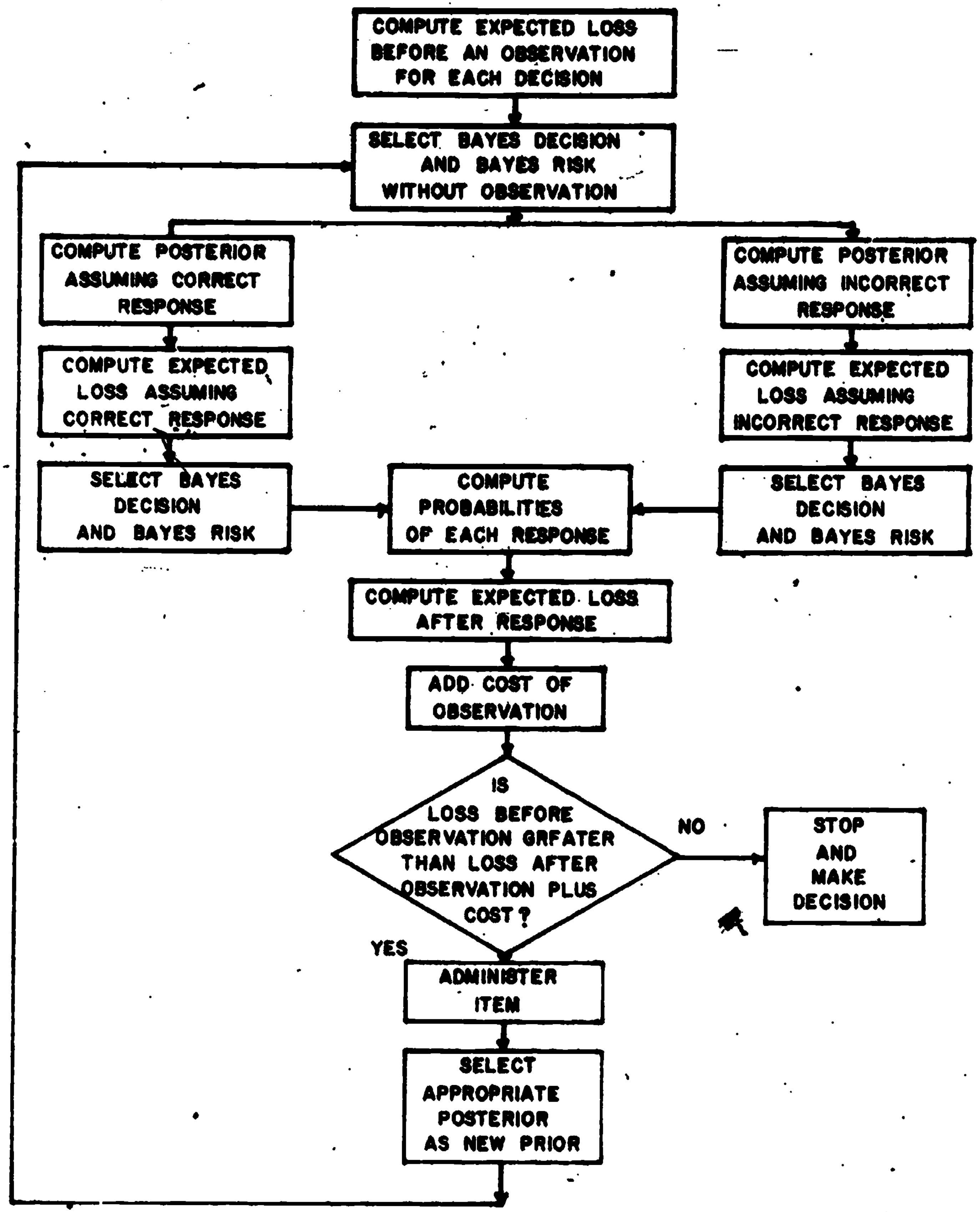

COMPUTE EXPECTED LOSS BEFORE AN OBSERVATION FOR EACH DECISION
SELECT BAYES DECISION AND BAYES RISK WITHOUT OBSERVATION
COMPUTE POSTERIOR ASSUMING CORRECT RESPONSE
COMPUTE POSTERIOR ASSUMING INCORRECT RESPONSE
COMPUTE EXPECTED LOSS ASSUMING CORRECT RESPONSE
COMPUTE EXPECTED LOSS ASSUMING INCORRECT RESPONSE
SELECT BAYES DECISION AND BAYES RISK
COMPUTE PROBABILITIES OF EACH RESPONSE
SELECT BAYES DECISION AND BAYES RISK
COMPUTE EXPECTED LOSS AFTER RESPONSE
ADD COST OF OBSERVATION
IS LOSS BEFORE OBSERVATION GRFATER THAN LOSS AFTER OBSERVATION PLUS COST ?
NO
STOP AND MAKE DECISION
YES
ADMINISTER ITEM
SELECT APPROPRIATE POSTERIOR AS NEW PRIOR

FIGURE 3
ONE-PARAMETER OC FUNCTIONS
FOR THREE INDIFFERENCE REGIONS

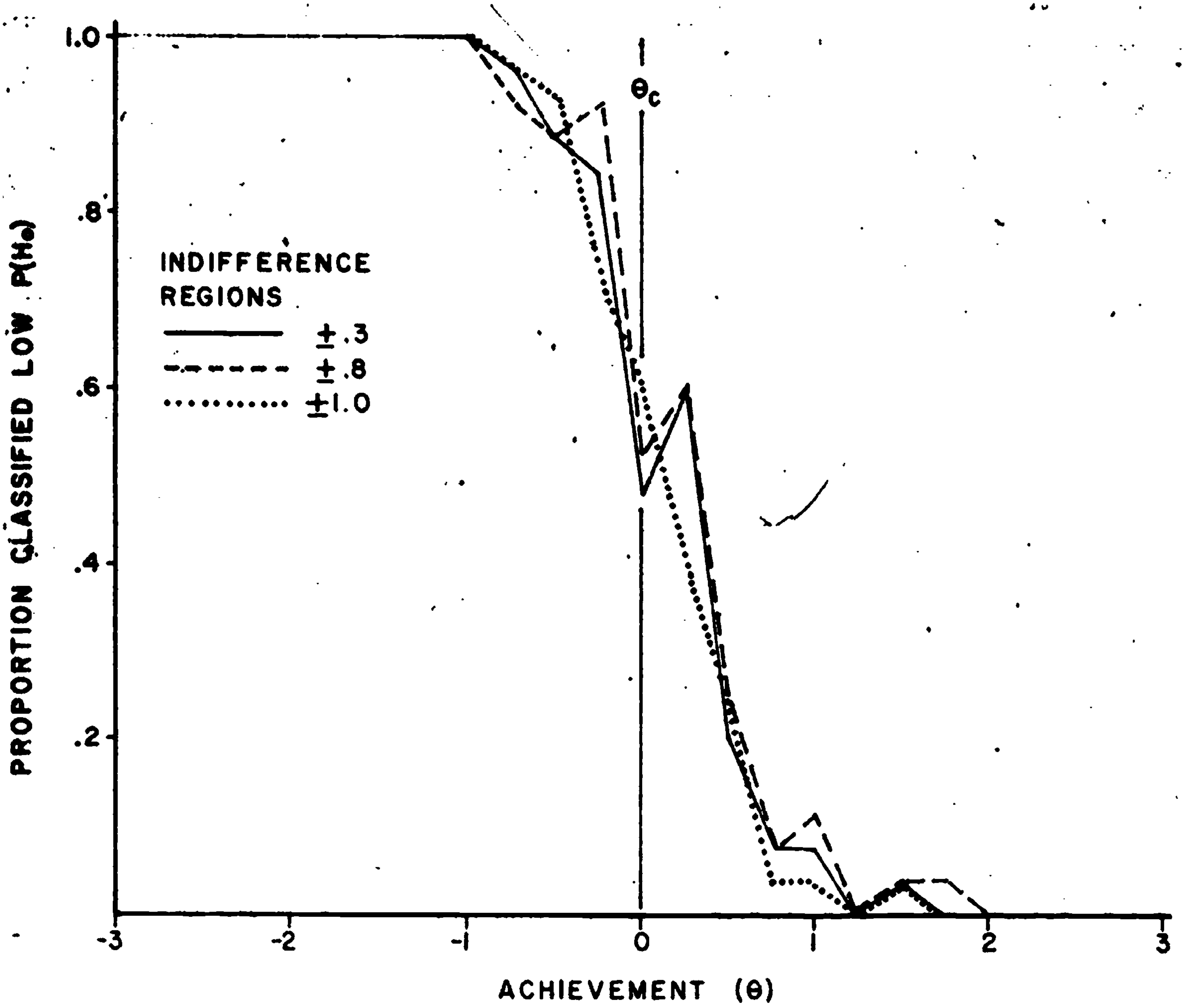


FIGURE 4
ONE-PARAMETER ASN FUNCTIONS
FOR THREE INDIFFERENCE REGIONS

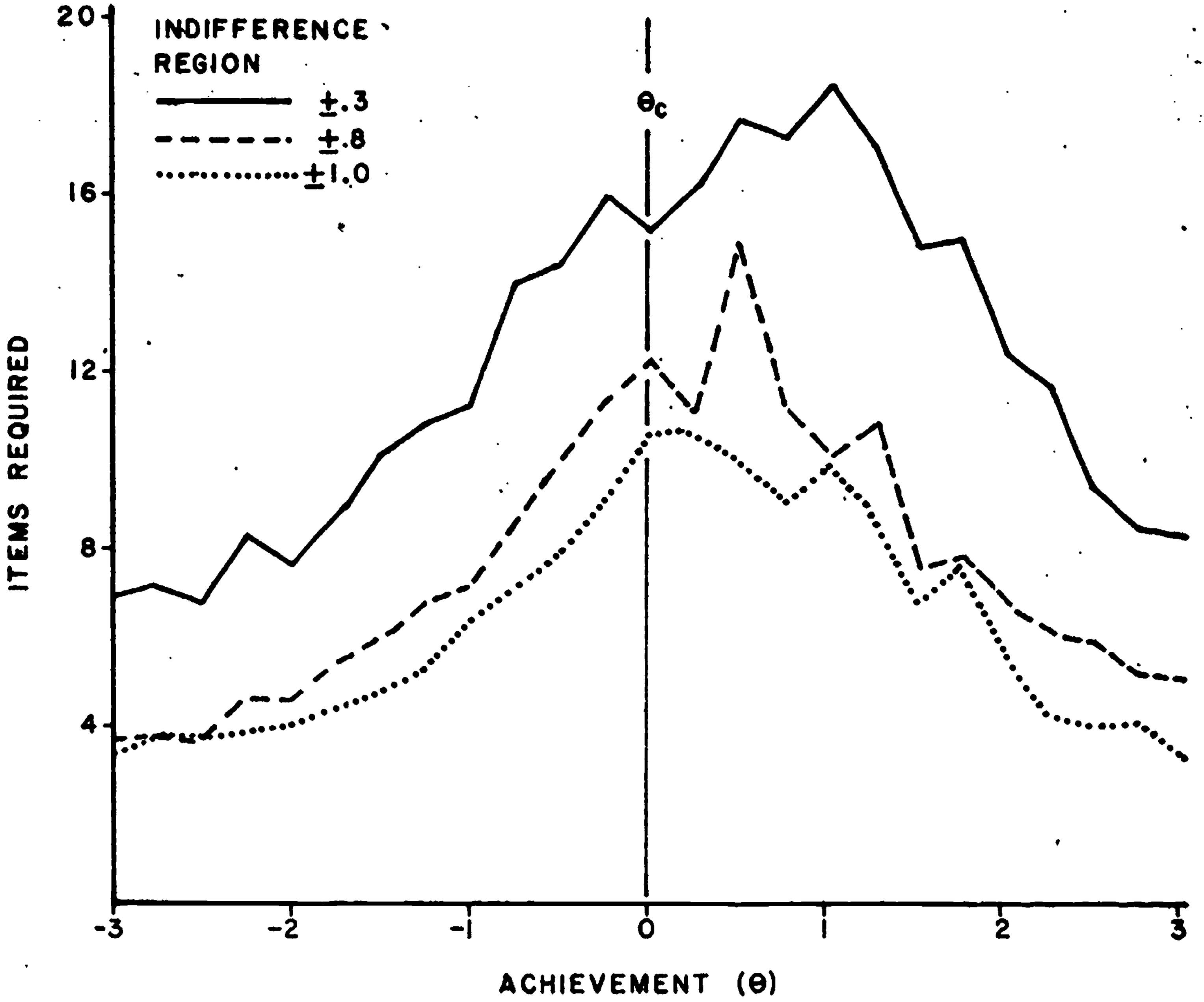


FIGURE 5
THEORETICAL OC AND ASN FUNCTIONS

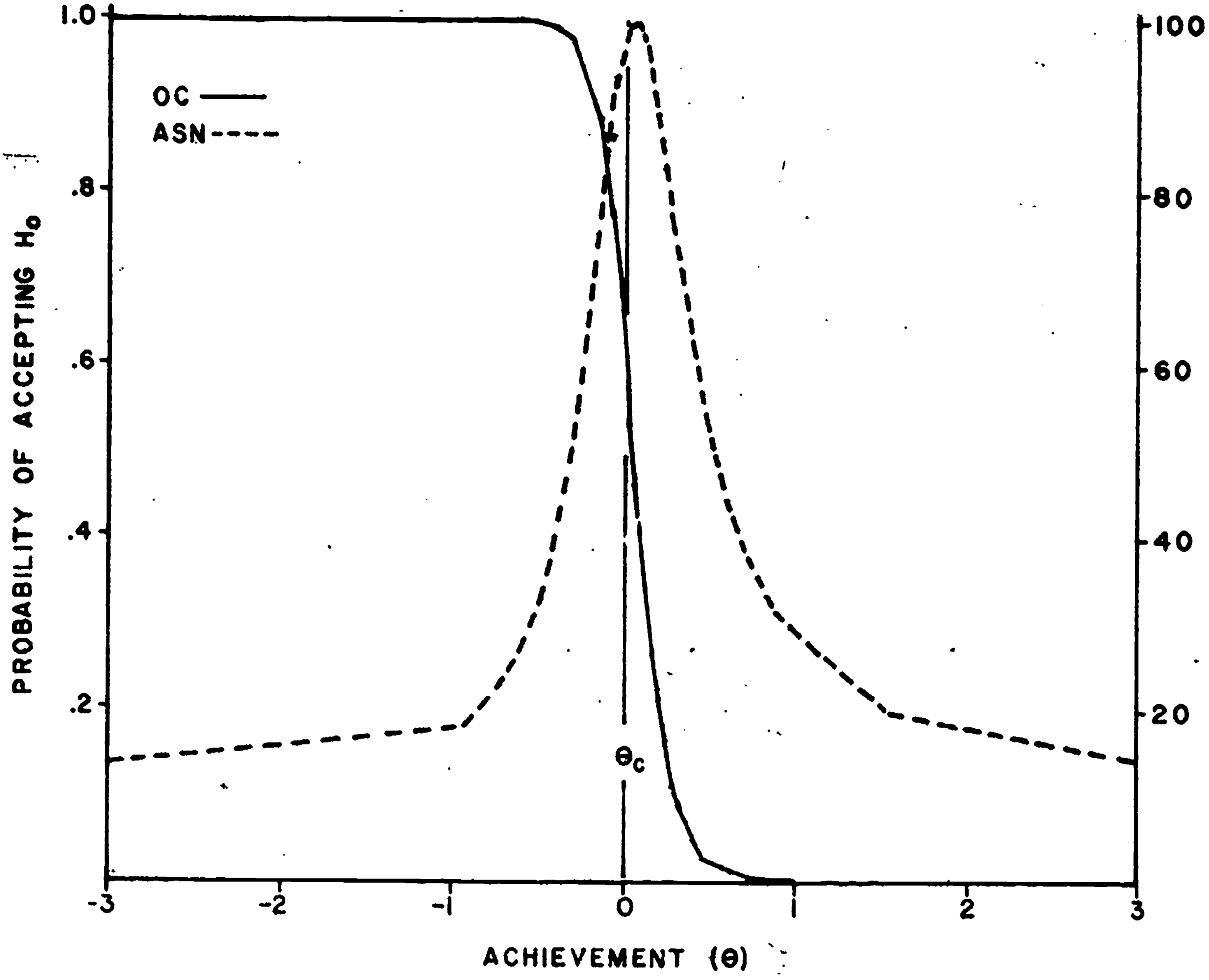


FIGURE 8

THREE-PARAMETER OC FUNCTIONS FOR THREE INDIFFERENCE REGIONS

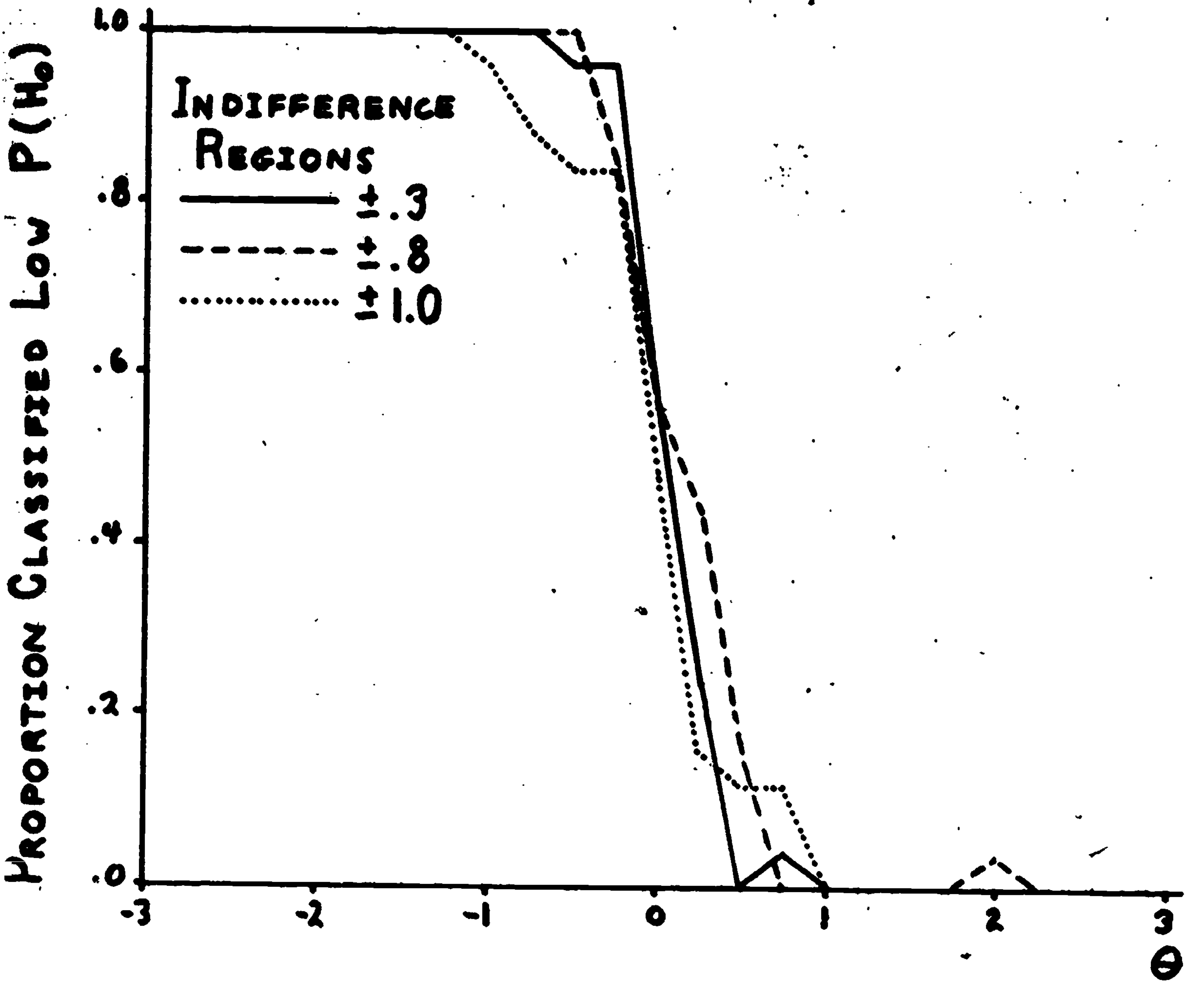


FIGURE 7

THREE-PARAMETER ASN FUNCTIONS FOR THREE INDIFFERENCE REGIONS

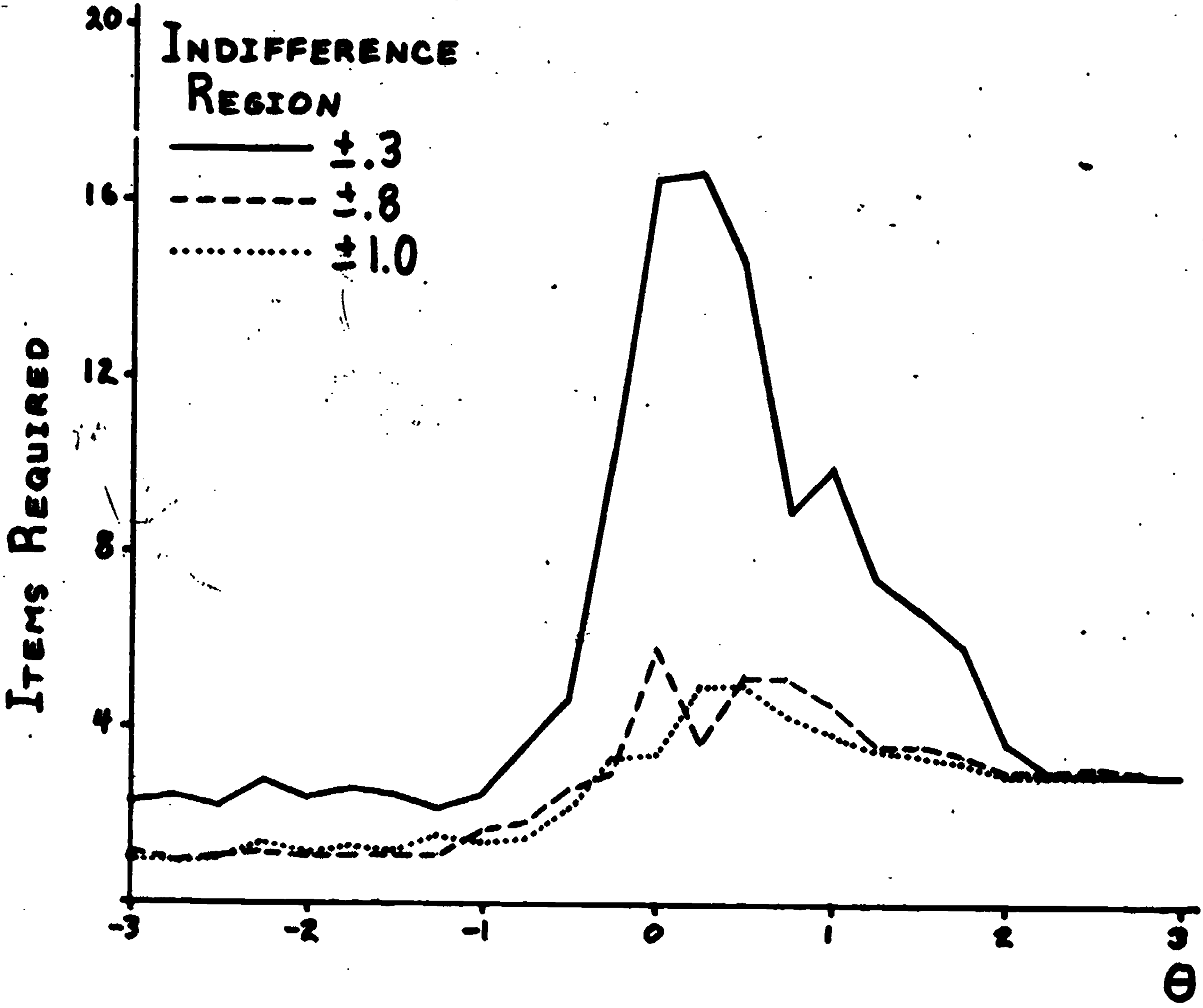


Figure 8
Composite OC Functions For Three Indifference Regions

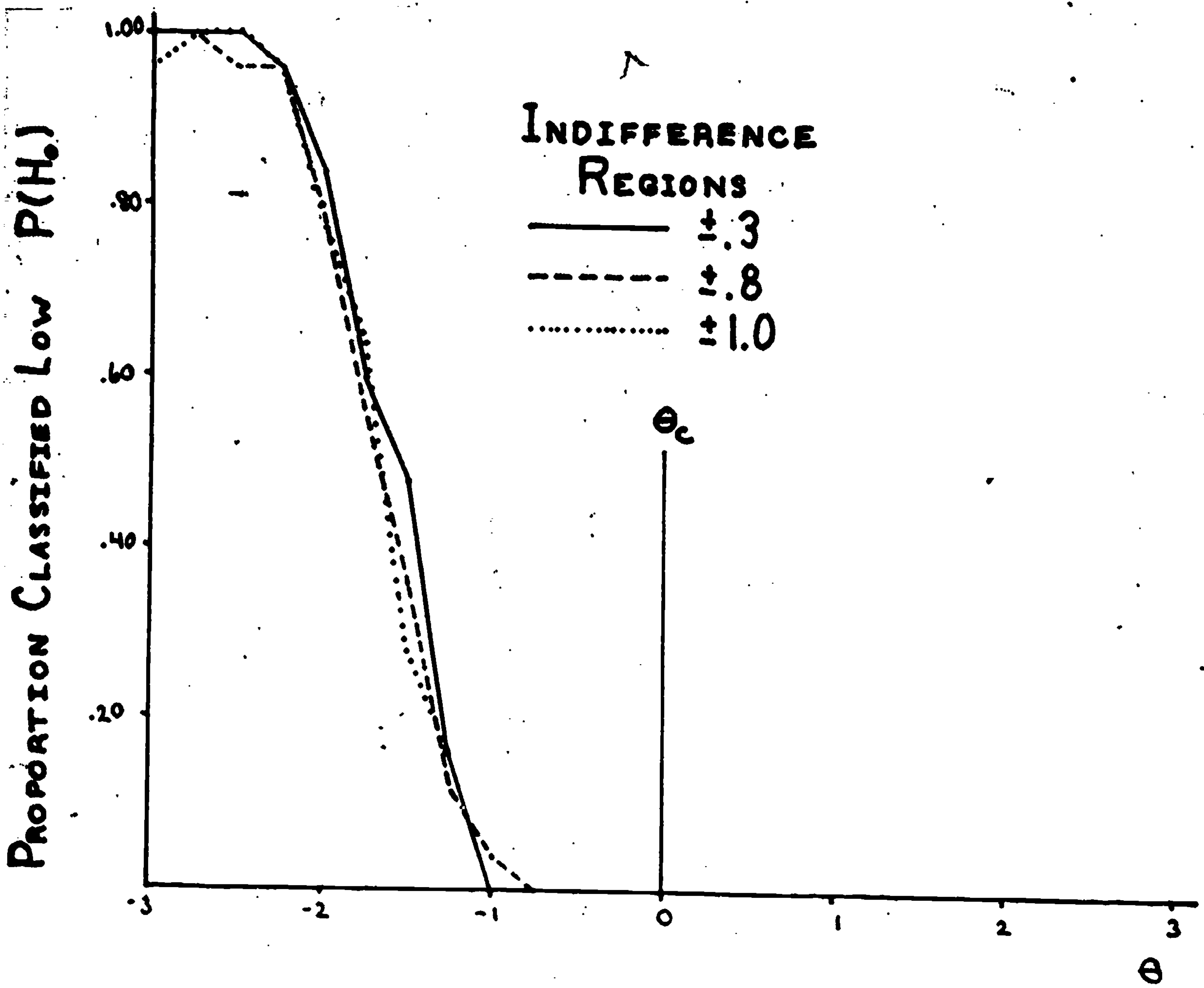
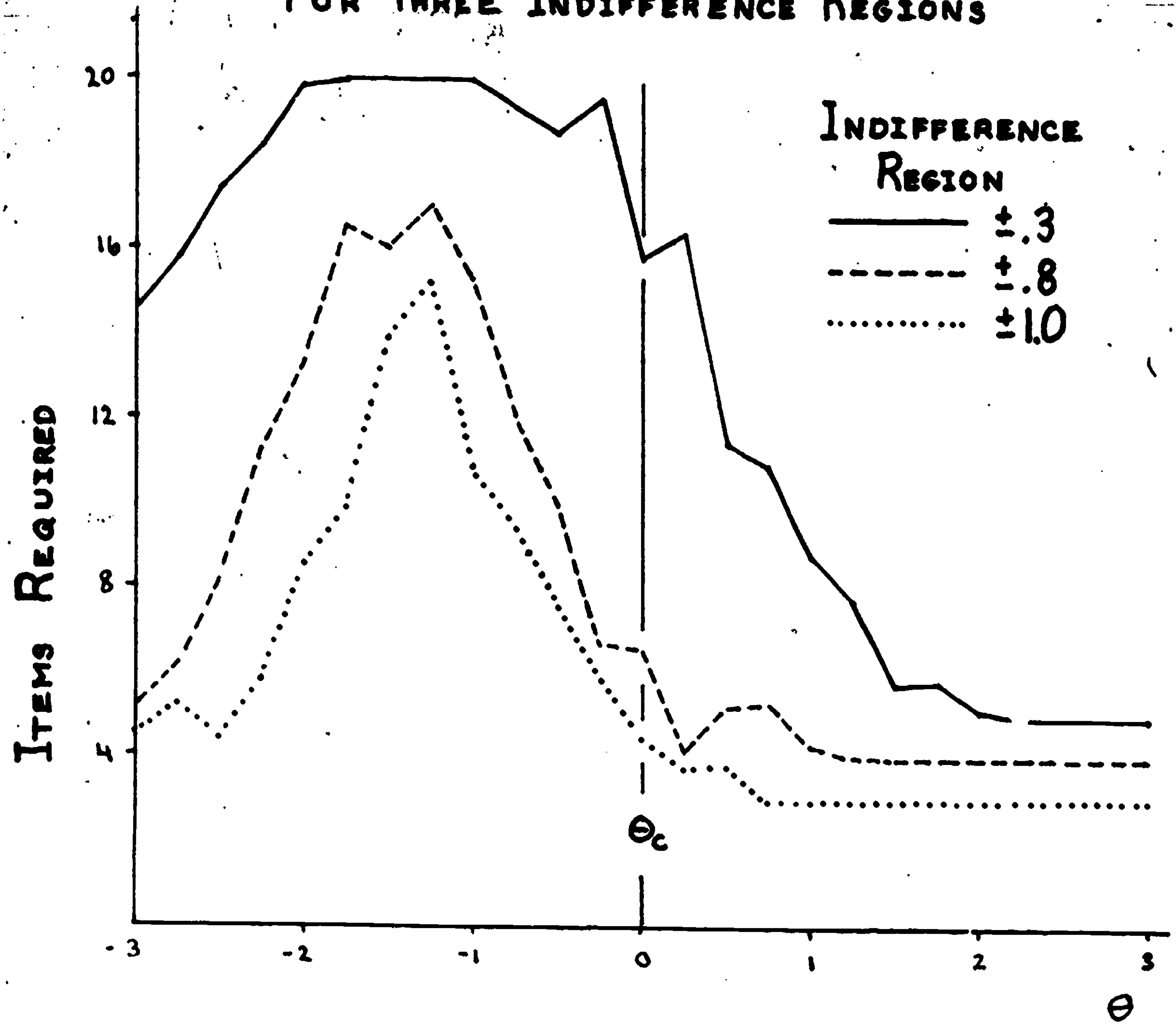


Figure 9
Composite ASN Functions For Three Indifference Regions


Items Required
θ
θ_c
Indifference Region
±.3
±.8
±1.0