DOCUMENT RESUME

ED 269 462 TM 860 316

AUTHOR Frick, Theodore W.
TITLE An Investigation of the Validity of the Sequential Probability Ratio Test for Mastery Decisions in Criterion-Referenced Testing.
PUB DATE 16 Apr 86
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Adaptive Testing; Classification; Comparative Analysis; *Computer Assisted Testing; Computer Science Education; Criterion Referenced Tests; Cutting Score ; *Decision Making; Higher Education; Hypothesis Testing; Item Banks; Latent Trait Theory; *Mastery Tests; Mathematical Models; *Predictive Validity; Ratios (Mathematics); Statistical Studies; Test Length
IDENTIFIERS *Adaptive Mastery Testing; Neyman Pearson Hypothesis Testing; *Sequential Probability Ratio Test (Wald)

ABSTRACT

The sequential probability ratio test (SPRT), developed by Abraham Wald, is one statistical model available for making mastery decisions during computer-based criterion referenced tests. The predictive validity of the SPRT was empirically investigated with two different and relatively large item pools with heterogeneous item parameters. Graduate students in a course on computer assisted instruction were administered tests on the Dimension Authoring Language and the COM Test, which measured knowledge of how computers functionally work. It was contended that, if the SPRT were used conservatively, it would remain robust as a decision model. Overall agreement coefficients ranged from .84 to .98, depending on the method of determining mastery status on the total test. The Neyman-Pearson classical approach to hypothesis testing was also included. The expected agreement was .95. An average of about 20 test items were required to reach SPRT mastery decisions, reducing testing time by 75 to 80 percent. (GDC)

AN INVESTIGATION OF THE VALIDITY OF THE SEQUENTIAL PROBABILITY RATIO TEST FOR MASTERY DECISIONS IN CRITERION-REFERENCED TESTING

Theodore W. Frick

Department of Instructional Systems Technology
School of Education
Indiana University
Bloomington

Presented at the Annual Meeting of
the American Educational Research Association
San Francisco
April 16, 1986

ABSTRACT

A variety of statistical models are available for making mastery decisions during computer-based criterion-referenced tests. Some of these decision models serve to shorten the length of a test, depending on the response pattern of an examinee during a test. The sequential probability ratio test (SPRT), developed by Abraham Wald, is one such model. In this study, the predictive validity of the SPRT was empirically investigated with two different and relatively large item pools with heterogeneous item parameters. It was contended that, if the SPRT is used conservatively, it remains robust as a decision model. Overall agreement coefficients ranged from .84 to .98, depending on the method of determining mastery status on the total test, when expected agreement was .95. About 20 items were required on the average to reach SPRT mastery decisions, a 75 to 80 percent reduction in test administration time for the item pools used in this study.

INTRODUCTION

Criterion-referenced achievement testing has gained increasing acceptance over the last twenty-five years, particularly in mastery learning contexts. Since computers have become less expensive and more prevalent in schools and universities, tests administered interactively to individuals by computers are becoming more practicable. Computer-based mastery tests can be adapted and shortened, depending on an examinee's response pattern during the test. One of the major advantages of adaptive testing is reduction of administration time necessary for mastery classifications.

Adaptive Mastery Testing

One of the more promising approaches to adaptive mastery testing (AMT) is based on item response theory (Weiss & Kingsbury, 1984). In this approach a one-, two-, or three-parameter logistic ogive is assumed to describe the functional relationship between an achievement continuum and the probability of observing a correct response to any of the items on the test. Information available in any test item is considered to be a function of the item's difficulty, discriminatory power, and lower asymptote (i.e., the "guessing" parameter). As a test is administered in the AMT approach, the item selected next is that which provides the most information about student achievement at that point in the test. After scoring a response to an item, a student's achievement level is estimated by a test characteristic curve (TCC), which is a mathematical function that describes the relationship between an achievement continuum and the expected proportion of correct responses that a person at any achievement level would attain had all the items on the test been administered. If a Bayesian confidence interval surrounding a student's predicted achievement level does not include the cut-off point used for decision making and lies above that point, then a mastery decision is rendered; or if below, nonmastery. Otherwise, if the confidence interval includes the cut-off point, the test is continued by selecting the item in the remaining pool which is predicted to provide the most information about that student's achievement level. In other words, a test is adapted to an individual's achievement level and ends as soon as a mastery or nonmastery decision can be reached, given a priori classification error rates.

Comparison of Adaptive, Sequential, and Conventional Mastery Tests

In a computer-based Monte Carlo simulation, Kingsbury and Weiss (1983) compared the AMT approach to the sequential probability ratio test (SPRT--developed by Wald, 1947), and to conventional tests of various fixed lengths. The SPRT is described in detail below (pp. 7 - 14). Conventional mastery tests are those in which an examinee is given a fixed set of items, and the proportion of correct answers is compared to a predetermined cut-off for mastery decisions. While the SPRT was the most efficient method when items were of equal difficulty levels, the AMT was found to be superior under test conditions where item parameters were varied. Although the AMT almost always required more items than the SPRT to reach a mastery/nonmastery decision, the AMT yielded fewer classification errors when item parameters were varied. Thus, it would appear from this simulation that the AMT is, overall, a better approach than either the SPRT or conventional fixed length tests.

It is not surprising that the SPRT resulted in more classification errors than the AMT, since shorter tests tend to be less reliable than longer ones. One might wonder if the SPRT would have predicted more accurately had it been used more conservatively (i.e., with smaller alpha's and beta's). One might also wonder if the comparisons were truly equitable, since the SPRT compares two simple hypotheses rather than two composite hypotheses in determining a person's mastery status. For example, what if a narrower zone of indifference (the gap between the two hypotheses) had been used with the SPRT? It is clear from the SPRT model that narrower zones of indifference will tend to increase the average sample number required to choose one of the hypotheses. It should be noted that Kingsbury and Weiss (1983) did recognize these difficulties in comparing the AMT and SPRT.

It should be also noted that the SPRT assumes random sampling from an item pool in order to predict the decision that would be reached had the entire pool been administered to an individual, whereas the AMT assumes nonrandom sampling based on factors described above. In this sense, the comparison with the SPRT is somewhat questionable, since the SPRT is, at least as originally formulated, not an adaptive methodology--though see Reckase's (1983) modification of the SPRT for tailored testing.

Limitations of Adaptive Mastery Testing

While the item response theory (IRT) on which the AMT approach is based has some distinct advantages over classical test theory (c.f., Lord & Novick, 1968; Hambleton & Cook, 1977), IRT does have some limitations: 1) Its validity depends on the adequacy of the posited test characteristic curve for modeling an achievement continuum. If the functional form of the mathematical model does not correspond to a true achievement continuum for a test (i.e., it is not an ogive, or perhaps not a continuous function at all), then decisions based on students' predicted achievement levels would be based on an incorrect model and hence lack validity. 2) In order to use IRT for making decisions about test results, it is first necessary to estimate item characteristic curves (ICCs) and a test characteristic curve (TCC). To obtain good estimates of item parameters, administration of test items to a fairly large number of individuals is required. It has been suggested that an n of at least 200 is needed for reasonably accurate estimates of item parameters (Hambleton & Cook, 1983--though see Lord's (1983) discussion of the one parameter model).

The first limitation is more serious. To the extent the chosen mathematical model is incorrect, test decisions are not valid. The second limitation is a practical one for typical classroom testing situations. Many teachers who design their own tests will not have the luxury of waiting until 200 students have taken a given test in order to estimate item parameters, let alone have access to the computing power and software necessary to calculate ICCs and TCCs, or possess the expertise to implement it correctly. Moreover, developers of computer-assisted instruction (CAI) programs, where embedded mastery tests are used, will probably find such a complex procedure unwieldy for many practical applications.

While IRT appears promising for standardized or large-scale testing situations, where test developers are more likely to have the resources and expertise to implement it, the practicality of this approach for most classroom testing situations and CAI embedded mastery tests can be seriously questioned at present.

Further Examination of the SPRT

One of the attractive features of he SPRT is that it is not very difficult for a competent programmer to implement on a computer--roughly 15 to 25 lines of code in most high-level languages--and could be incorporated in a fairly straightforward way into computer-based testing systems and CAI programs as an alternative decision model to conventional testing. Moreover, the SPRT does not require advanced estimates of item parameters and could be used immediately for mastery test decisions.

Why, then, has the SPRT seldom been used as a decision model for mastery testing? The most frequent criticism is that if item parameters vary widely, probability estimates in the SPRT will be incorrect--i.e., a major assumption of the SPRT model is violated. This criticism will be addressed in considerable detail below. The second difficulty with the SPRT is that it requires two "cut-off" levels rather than a traditional single cut-off used in criterion-referenced testing to which most practitioners are accustomed. The second problem is no different in principle, however, than the problem of classification of test scores near a single cut-off point when measurement error is considered, and so is of lesser concern here--though not everyone may share this view.

The author has developed a computer simulation of the SPRT in order to observe the number of test items required to reach mastery or nonmastery decisions with different response patterns when mastery, nonmastery, alpha and beta levels are systematically varied. Generally, fewer test items are required to reach decisions when the zone of indifference (the gap between mastery and nonmastery levels) is greater or when alpha and beta decision error rates are higher. The converse is true as well. These results should not be surprising given the formulation of the SPRT. Also, nonmastery decisions tend to be reached more quickly than mastery decisions when a pattern of mostly incorrect responses is given, compared to a pattern of mostly correct ones, using typical mastery and nonmastery levels.

The SPRT was then pilot tested in a computer-based instructional program that taught a programming concept that few students had previously learned. A test item pool of 20 items was developed and used for both pretesting and posttesting. The items were fairly uniform and all

required constructed responses. In 45 out of 46 cases students agreed that the decision reached by the SPRT was valid at both pre- and posttest occasions. This was independently cross-checked by informal observation of student performance. Typically, 3 to 5 items were required to reach pretest nonmastery decisions, and 8 to 14 for posttest mastery decisions (using a mastery level of .85, nonmastery level of .50, alpha = .05, and beta = .10).

Thus, pilot test results suggested that the SPRT was promising as a decision methodology when items were mostly uniform. These results were consistent with those in the Kingsbury and Weiss Monte Carlo simulation. However, will SPRT decisions be valid with heterogeneous item pools? The Kingsbury & Weiss simulation suggested that the SPRT will predict less well under these conditions. On the other hand, if used conservatively, the SPRT might nonetheless predict well enough to be satisfactory in many mastery learning contexts, though not as precise as the AMT approach.

In short, despite an apparent violation of an assumption of the SPRT model, it might still remain robust as a decision model if used conservatively (similar to ANOVA, for example, when the normality assumption is violated to some extent). The predictive validity of the SPRT with heterogeneous item pools is the major focus of the present study. Before discussion of methodology and results, a brief review of the classical hypothesis testing procedures on which the SPRT is modeled and a description of the SPRT itself are presented for those who are unfamiliar with these models.

BACKGROUND

The Neyman-Pearson Classical Approach

This example of classical hypothesis testing in the Neyman-Pearson framework is provided in order to contrast it subsequently with the sequential probability ratio test.

Suppose a quality control inspector were faced with the task of deciding whether or not to reject a large batch of mass-produced integrated circuits (ICs). When the production system is working normally, 85 percent or more ICs meet expected standards and 15 percent

or less do not; buyers of large quantities of these ICs are willing to accept this failure rate and simply discard bad chips when encountered. When the production system is not working properly, 60 percent or less are good, as determined from past experience, and a 40 percent or higher failure rate is clearly unacceptable to buyers.

There would be two hypotheses in the Neyman-Pearson approach:

H_0: p(good IC) = .60 H_1: p(good IC) = .85

If by randomly sampling ICs from the lot either H_0 or H_1 can be chosen with a fairly high degree of confidence, then it will be unnecessary to test the entire lot, which would be prohibitively expensive. Suppose that 40 ICs are sampled randomly without replacement from the lot, and after testing, 31 are found to be good. Which of the two hypotheses is more likely to be true?

The theoretical sampling distributions for the two hypotheses are illustrated in Figure 1. There are two types of decision errors that could be made. If H_1 is chosen when H_0 is really true, we have made a Type I error (alpha). Conversely, if H_0 is chosen when H_1 is actually true. we have made a Type II error (beta). Typically, an alpha level and sample size are determined in advance, and these choices determine beta, given the hypotheses in question. (We could, however, set alpha and beta in advance, which would determine the sample size; or instead set beta and the sample size, which would determine alpha.) If we set alpha = .05 for a random sample of 40, then a critical region of the H_0 sampling distribution is established. H_0 will be rejected if the obtained number of good ICs falls within the critical region. In this example, the critical region determined from the H_0 sampling distribution is 30 or higher with alpha = .05 and $\underline{n}$ = 40; beta is therefore approximately .03.

Since the obtained number of good ICs (31) in our random sample of 40 lies within the critical region, we reject H_0 and accept the alternative, H_1. The probability of a sample with 31 successes out of 40 occurring in the H_1 distribution is about .0682, whereas it is about .0095 in the H_0 distribution. In other words the odds are about 7 to 1 in favor of the sample occurring in the H_1 vs. the H_0 distribution. Notice that the obtained number of good ICs in the sample was not equal to 34; but it is 7 times more likely that such a sample would be drawn from a theoretical binomial distribution with an expected value of 34 vs. 24 ($\underline{n}$ = 40).

Figure 1. Theoretical Sampling Distributions for N = 40 (Null Hypothesis: p = .60)

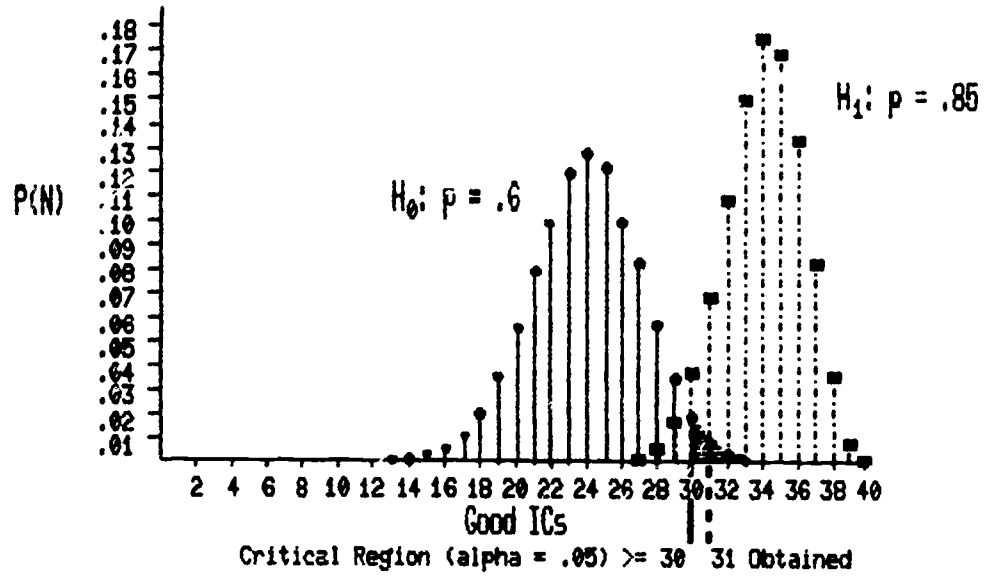
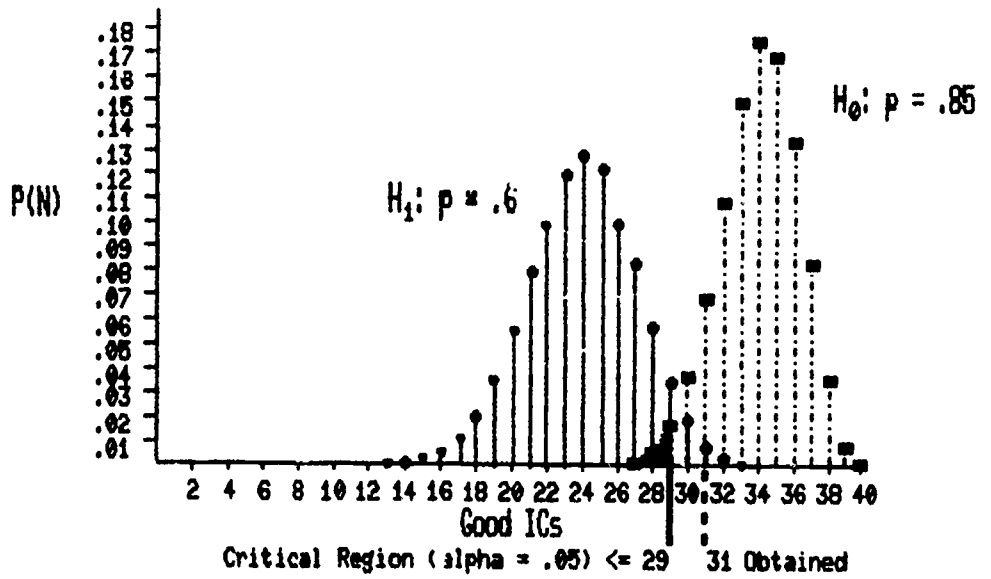


Figure 2. Theoretical Sampling Distributions for N = 40 (Null Hypothesis: p = .85)



Notice also that we would have reached the same conclusion for H_0's with p's less than .60 and H_1's with p's greater than .85, and it can be shown that alpha's and beta's would be no greater than their levels set for the original hypotheses.

One might wonder why the null hypothesis was chosen to be $p \leq .60$. What if the null and alternative hypotheses were switched? If the null hypothesis is taken to be $p \geq .85$, will the decision be the "same" with the obtained sample? In this case the critical region is 29 or less good ICs for an alpha = .05, with n = 40, and beta = .034. See Figure 2. In this example with an obtained sample of 31 good ICs, the decision is not to reject the null hypothesis that $p \geq .85$, which is parallel to the earlier decision. However, this will not always be the case. For example, if the obtained sample were 29 or 30 good ICs, the decision will depend on which hypothesis is treated as null—though it should be noted that the alpha's and beta's are not exactly equivalent here, since the sampling distributions are discrete. Normally, the null hypothesis is the one to be rejected--i.e., there must be compelling evidence that it is not true before we change our minds about it. In this quality control example, if the expectation is that the production system is working normally, then it would probably be more appropriate to take that as the null hypothesis (Figure 2). If the sequential probability ratio test is used for the statistical decision, as discussed below, it does not matter which hypothesis is taken to be null.

The Sequential Probability Ratio Test (SPRT)

Abraham Wald (1947) originally developed the SPRT as a statistical decision procedure to solve problems of inference similar to the one above concerning quality control. Wald indicated that the SPRT will require, on the average, about half the sample size required by a classical Neyman-Pearson test of the same hypotheses using the same alpha and beta levels. How can this be?

One difference between the two procedures is that in the classical approach the statistical test of the hypotheses does not occur until a sample of n observations is obtained and evaluated, where the outcome of of each observation is characterized dichotomously (e.g., good/bad,

success/failure). In the SPRT, a test of the hypotheses is made after each observation. If one of the hypotheses can be chosen, given the sequence of observations thus far and established alpha and beta levels, sampling terminates; otherwise another object is randomly chosen and the SPRT is applied again. If there is a clear trend favoring one hypothesis over the other early in the sequence of observations, it is likely that the same conclusion would have been reached by a classical Neyman-Pearson test with the same alpha and beta levels. Moreover, the average sample number (ASN) for the SPRT would be about half the n required for an equivalent classical test (Wald, 1947, p. 57).

Normally both approaches require that observations are independent and that sampling is random without replacement. Wald (1947) claimed that the SPRT is also valid when observations are dependent (p. 44).

The SPRT relies on three inequalities:

Reject H_0 (accept H_1) if:	$p_{1m}/p_{0m} \geq A$	[1]
Do not reject H_0 if:	$p_{1m}/p_{0m} \leq B$	[2]
Continue sampling if:	$B < p_{1m}/p_{0m} < A$	[3]

It is assumed here that the p for H_1 is greater than that for H_0; $B < A$; p_{1m} is the probability of the observed sequence when H_1 is true; and p_{0m} is the probability of the observed sequence when H_0 is true. Wald demonstrated that the constant A is approximated conservatively by [(1 - beta)/alpha], and B by [beta/(1 - alpha)]. Formulas for determining p_{1m} and p_{0m} depend on whether or not observations are assumed to be independent.

Inequality [1] can be interpreted: If the odds of the observed sequence of observations, when H_1 is true vs. H_0, are equal to or greater than the odds of rejecting H_0, when H_1 is true vs. when H_0 is true, then stop sampling and reject H_0.

Inequality [2] can be interpreted: If the odds of the observed sequence of observations, when H_1 is true vs. H_0, are less than or equal to the odds of accepting H_0, when H_1 is true vs. H_0, then stop sampling and do not reject H_0.

As an example using the same hypotheses and alpha and beta levels as above for the Neyman-Pearson test, we begin randomly sampling from the lot of ICs. The first one is good. The SPRT is applied. Inequality [3] is true, so we sample another, and so on, until we just happen to have

found 19 good ones and 4 bad ones so far. At this point, inequality [3] is still true (with H_0: p = .60; H_1: p = .85; alpha = .05; beta = .03). We sample another IC and it is a good one (20 good, 4 bad so far). We apply the SPRT and inequality [1] is now true. We therefore reject H_0, and accept the hypothesis H_1 that the lot is an acceptable one (where p(good IC) $\geq$ .85). The total sample size this particular time was 24, substantially less than the 40 required by the Neyman-Pearson test. If we were to begin sampling again from this same lot, the SPRT sample size would probably be different from before, but the same decision will be reached in accordance with the a priori alpha and beta error rates. Occasionally, wrong decisions will be made via the SPRT, due to sampling error, but no more often than would occur in a large number of samples using the Neyman-Pearson approach with equivalent alpha and beta levels (Wald, 1947).

Use of the SPRT in Mastery Testing

Although the SPRT has been used widely as a decision methodology in manufacturing quality control settings, few references to the SPRT have been found in the educational and psychological testing literature. Ferguson (1969) used the SPRT for making mastery decisions in an individually prescribed instruction (IPI) framework. Reckase (1979, 1981, 1983), McArthur and Chou (1984), and Kingsbury and Weiss (1983) have explored the use of the SPRT in criterion-referenced testing, particularly for computer-based tests.

The major criticism of the SPRT is that it does not account for variability in item parameters, which in turn might result in invalid probability estimates in inequalities [1] to [3] (c.f., Kingsbury and Weiss, 1983; Reckase, 1979; McArthur & Chou, 1984). A second criticism of the SPRT for use in mastery test decisions is that it requires in effect two cut-off levels, rather than the traditional single cut-off level. Typically, a cut-off score is established (e.g., .85) and examinees who score at or above the cut-off are classified as masters, and those who score below as nonmasters.

The second criticism is somewhat misleading. It is known that misclassifications are likely to occur when examinees score near the cut-off score (c.f., Novick & Lewis, 1974). Given the reliability of a

mastery test, it is possible to construct a confidence interval around each obtained score, based on the standard error of measurement. If that confidence interval does not include the cut-off score, then fewer classification errors would be expected. However, when a confidence interval includes the cut-off score, we cannot be as sure. Due to error of measurement and possibly other factors, an examinee who happened to score just above the cut-off this time might score below if the test (or an equivalent one) were taken again. An alternative way of viewing the situation would be to establish a confidence interval around the cut-off score and require that obtained scores lie outside that interval for classification, whereas scores falling inside the interval would not be classified as either mastery or nonmastery. For example, suppose that a cut-off of .80 were established, and the 95 percent confidence interval was determined to be .80 $\pm$ .07. Thus, scores falling in the .73 to .87 range would be classified as no decision, those below .73 as nonmasters, and those above .87 as masters.

Though not the same, the latter procedure and the SPRT are very similar. The SPRT requires two hypotheses. Following Wald (1947, p. 29), the zone of indifference should be established by answering two questions:

1) What is the highest proportion of correct responses on the test above which we would not want to classify someone as a NONMASTER?

2) What is the lowest proportion of correct responses on the test below which we would not want to classify someone as a MASTER?

These two proportions then determine the zone of indifference and the hypotheses tested by the SPRT. For example, in a mastery learning situation we might decide that we would not want to classify someone who scored at least .85 on the test as a nonmaster. Similarly, we might decide that we would not want to classify someone who scored .60 or lower as a master. How these levels are chosen will depend on the nature of the situation and the consequences of incorrect decisions.

One might ask, "But what do we do about students who score in the zone of indifference?" The answer may be a little surprising. If the item pool is large enough, one of the hypotheses will eventually be chosen by the SPRT. Why is that? Recall in the earlier quality control

example of the sample of 31 good ICs (see Figure 1). The alternative hypothesis was that there are at least .85 good ICs in the population (the lot in question). If the alternative hypothesis is true, we would expect 34 good ICs in a sample of 40, but due to sampling error the number of ICs will not be exactly 34 most of the time. Although a sample of 34 good ICs in 40 would be expected most often under the alternative hypothesis, the probability of obtaining exactly 34 good parts in 40 is about .17. In other words, about 83 percent of the samples of 40 would be expected to yield a number of good parts other than 34.

A student's obtained score may lie in the zone of indifference, or it may be at or below the nonmastery level, or at or above the mastery level. The SPRT simply indicates which of the two hypotheses is most likely to be true, given a priori alpha and beta decision error rates. For example, a student may have answered 78 percent of the items correctly thus far in a test. Sampling would end, with a mastery decision, if it is true that the odds of a sample of this size with 78 percent correct, when the mastery vs. nonmastery hypothesis is true, are equal to or greater than the odds of a correct vs. an incorrect mastery decision. See inequality [1].

Before discussing the issue of variability in item parameters, such as difficulty level and discriminatory power, terminology and formulas related to use of the SPRT in mastery testing are addressed next.

Mastery hypothesis (H_m: $p \geq P_m$) This is the hypothesis that the examinee is a master of some educational objective, as indicated by responses to test items which match the objective, where items are scored dichotomously (i.e., right or wrong). The P_m for the mastery hypothesis is established by answering the question, "What is the highest proportion of correct responses on the whole test above which we would not want to classify someone as a nonmaster?"

Nonmastery hypothesis (H_{nm}: $p \leq P_{nm}$) This is the hypothesis that the examinee has not mastered some educational objective, as indicated by responses to test items which match the the objective, where items are scored dichotomously. The P_{nm} for the nonmastery hypothesis is established by answering the question, "What is the lowest proportion of correct responses on the whole test below which we would not want to classify someone as a master?" It is further assumed that $P_{nm} < P_m$.

Incorrect mastery decision (alpha) This is the probability of concluding mastery when the examinee is actually a nonmaster, and should indicate our tolerance for making decision errors of this type. Note that (1 - alpha) is the probability of a correct nonmastery decision.

Incorrect nonmastery decision (beta) This is the probability of concluding nonmastery when the examinee is actually a master. Note that (1 - beta) is the probability of a correct mastery decision.

P_m, P_{nm}, alpha and beta are established by the decision maker prior to administration of the mastery test. Their values will depend on the purpose of testing and the relative consequences of incorrect decisions.

The final two pieces of information needed by the SPRT are the number of right (R) and wrong (W) answers observed thus far in a test.

The decision formulas are as follows:

CHOOSE H_m IF:

$$\frac{(P_m)^R(1 - P_m)^W}{(P_{nm})^R(1 - P_{nm})^W} \geq \frac{(1 - \text{beta})}{\text{alpha}} . \qquad [1']$$

Another way of expressing this is:

$$\frac{P(\text{sequence}|H_m)}{P(\text{sequence}|H_{nm})} \geq \frac{P(\text{Mastery decision}|\text{Master})}{P(\text{Mastery decision}|\text{Non-master})} .$$

CHOOSE H_{nm} IF:

$$\frac{(P_m)^R(1 - P_m)^W}{(P_{nm})^R(1 - P_{nm})^W} \leq \frac{\text{beta}}{(1 - \text{alpha})} . \qquad [2']$$

Another way of expressing this is:

$$\frac{P(\text{sequence}|H_m)}{P(\text{sequence}|H_{nm})} \leq \frac{P(\text{Nonmastery decision}|\text{Master})}{P(\text{Nonmastery decision}|\text{Nonmaster})} .$$

OTHERWISE, MAKE NO DECISION, AND CONTINUE TESTING.

It should be noted that when dealing with finite populations which are rather small the above formulas for calculating the probabilities of the sequence of observations under the two hypotheses should be modified (see Wald, 1947, p. 44).

In order to calculate the probabilities of the observed sequence of responses to test items under H_m and H_{nm}, respectively, it appears necessary to assume that observations are independent and that the probability of a correct response to any given test item is invariant, though not the same, under each hypothesis (using the above formulas).

Translated into practical terms, the first assumption implies that the probability of a correct response on any given test item for a given examinee should not differ depending on which items may have been answered previously. If items are randomly selected and no feedback is given during the test, this assumption should generally be met--at least in principle, though it could be empirically tested.

The second assumption is apparently the troublesome one. For example, suppose an examinee were taking a test where items varied widely in terms of their difficulty level. It could happen, just by chance, that very easy items were sampled early in the test, resulting in a SPRT mastery decision; yet, had the whole test been taken, a nonmastery decision would have been reached. Conversely, it could likewise happen that very hard items were sampled early in the test, resulting in a SPRT nonmastery decision that would disagree with a total test mastery decision. This problem is similar to that which might occur in a quality control setting if the sample were not representative enough. If an inspector happened to take a sample from one area of the lot where there were many bad ICs, the lot would most likely be rejected although it might have been perfectly acceptable had a larger and more representative sample been taken.

P_m and P_{nm} have often been interpreted as the probabilities of a correct response to any item on a test under the two hypotheses (c.f., Ferguson, 1969; Kingsbury & Weiss, 1983; Reckase, 1983, McArthur & Chou, 1984). It is argued that since the probability of a correct response to a test item will depend on the difficulty of the test item, the ability of the examinee, and other factors, the SPRT is therefore an inappropriate model--particularly if items are selected to maximize information at various ability levels, as is done in tailored or adaptive testing.

On the other hand, if items are selected randomly, and p is the proportion of items a student can correctly answer, this SPRT assumption

would not appear to be violated. That is, the SPRT is merely trying to predict the decision that would be reached had the entire universe of test items been taken by a particular examinee at this particular time. In other words, given a smaller sample of responses to test items which have been selected at random from a larger sample of test items (which in turn have been selected from the universe of test items), the SPRT is simply predicting the decision that would be reached had all the items in the larger sample been administered to this particular examinee on a particular testing occasion (c.f., Lord & Novick, 1968, Chapter 11).

Furthermore, it can be argued that the probability of a correct response to a particular test item on a particular test by a particular examinee on a particular occasion is either zero or one--i.e., a person either gets that item right or wrong on a particular administration of the test (assuming dichotomous scoring). As an analogy, suppose an urn contained 100 balls of various sizes and shapes, 70 of which were colored red (R) and 30 white (W). If we select a particular ball, it is either R or W--the probability that it is R is either zero or one, and likewise for W. However, assuming the balls have been mixed up, none has been selected so far, and we sample randomly, we would say the probability of selecting a red ball is .70.

Thus, the danger in using the SPRT is not that the probability of selecting a test question that an examinee would answer correctly will change according to item difficulty, when the universe of generalization is a particular examinee's mastery status, inferred from his or her total test score at that time. The danger in using the SPRT is terminating the test too quickly, before obtaining a sample of items representative enough of the whole pool. Therefore, if a test is suspected or known to have widely varying item parameters, then the SPRT should be used conservatively to insure that enough items are administered which are representative of the entire item pool, which in turn are assumed to be representative of the universe of test items for measuring mastery of some instructional objective. In other words, alpha and beta (particularly beta), should be kept very small when test item parameters vary widely. In addition, narrower zones of indifference will tend to increase the ASN in the SPRT model.

METHOD

Tests

Computer-based tests were constructed on: 1) the structure and syntax of the Dimension Authoring Language (DAL test), and 2) knowledge of how computers functionally work (COM test). Test items representative of these content domains, respectively, were constructed so that difficulty levels would be expected to vary. About half of e items on each test were multiple choice, one fourth binary choice, and one fourth constructed short answer. Subsequent item analyses indicated that items did vary considerably in difficulty and discriminatory power (see Appendix A).

The DAL test consisted of 97 items, and the COM test 85 items. Coefficient alpha was .977 and .943 for the two respective tests, based on results from the two groups described below. The DAL test was perceived by examinees as a very hard test. The mean score was 63.2 (66 percent correct) with a standard deviation of 24.6 (n = 53). The COM test was easier on the whole, with a mean score of 67.3 (79 percent, S.D. = 13.6, n = 105).

Tests were individually administered by the STEEL Computer-based Criterion-referenced Testing System (Frick, 1985). As an examinee sat at a computer terminal, items were selected at random without replacement from the total item pool until all items were administered. (Due to an oversight, only 96 items were administered on the DAL test.) Students were not allowed to go back and change previous answers to items, nor was feedback given during the test. When the test was finished, complete data records were stored in a database, including the actual sequence in which items were randomly administered, response time, literal response to each item, and the response judgment (right or wrong). Students were also informed of their total test scores at the end of the test. The COM test typically took 30 to 45 minutes to complete, whereas the DAL test usually took between 60 and 90 minutes.

Examinees

The examinees who took the DAL test were mostly either current or former graduate students in a course on computer-assisted instruction (CAI) taught by the author. Currently enrolled students took the DAL test twice, once about mid-way through the course when they had some knowledge of the Dimension Authoring Language (which they were required to learn in order to develop CAI programs), and once near the end of the course when they were expected to be fairly proficient in DAL. The remainder of the examinees took the DAL test once, and had never taken the test before. Since the test was long and known to be difficult, no one was asked to take the test who did not have some knowledge of DAL or other CAI authoring languages.

About two-thirds of the students who took the COM test were current or former graduate students in two sections of an introductory course on using computers in education taught by the author. Current students took the test as a pre- and posttest. The remaining one-third were undergraduate education students taking a beginning course in instructional compuung and took the test once, as well as did former students who had never taken the test before.

Though students were not chosen randomly, the timing of testing and other prior indications of their knowledge in these two content areas helped insure that there were fairly wide ranges of scores on both tests. The total number of administrations of the COM test was 105, and 53 for the DAL test.

Almost all examinees had some first-hand experience with computers prior to testing and, with few exceptions, did not appear to be intimidated by using a computer terminal or appear to be especially nervous about taking a computer-based test. Many indicated that they would have liked to go back and change some previous answers to questions, but were not allowed to do so by the testing system.

Method of Determining SPRT Outcomes

The SPRT was applied retroactively, since each student was originally given all the items in a pool. This was accomplished by a computer program which retrieved test results for each examinee from a database in which results were stored in the order the randomly selected

items were administered. P_m was set a priori to .85, P_{nm} to .60, and alpha and beta to .025. The SPRT wa applied after each item, as it would have been used during the actual testing, until a mastery or nonmastery decision was reached or the item pool was exhausted. The SPRT outcome, number of right and wrong answers required to reach a decision by the SPRT, and the total test results were written to a separate data file for further analysis.

The mean number of items required for SPRT mastery decisions on the DAL test was 19.1 (S.D. = 12.9) and for nonmastery decisions it was 17.4 (S.D. = 16.3). For the COM test the mean was 21.6 (S.D. = 12.6) for mastery decisions and 18.6 (S.D. = 14.7) for nonmastery decisions. Only once was the item pool exhausted without reaching an SPRT decision on either test.

Methods of Determining Mastery Status for the Total Item Pool

At first glance, a method of determining mastery status based on results from administration of the entire item pool to an examinee may appear to be straightforward and simple. One approach would be to classify any person who scored at or above P_m as a master; at or below P_{nm} as a nonmaster; and anywhere in between P_m and P_{nm} as neither (no decision). This approach would appear appropriate if: 1) measurement error is zero; and 2) the test item pool is considered to be the universe of test items that could be used to assess attainment of some instructional objective. If this approach were adopted, then calculations of probabilities in [1'] and [2'] should be altered to reflect sampling from a finite population (Wald, 1947). For example, if the nonmastery level is set for 60 or less out of 100 questions answered correctly, and someone has already missed 40 during sampling, then the test should be obviously terminated with a nonmastery decision. The probability that someone is a nonmaster is one in this example using this approach.

However, this approach is not considered suitable here, since measurement is not perfect and the total test item pool for a given instructional objective is considered to be a representative sample of the universe of test items that could be used to test mastery.

Another obvious method would be to use the SPRT itself on the total test results from an examinee. While tempting, this method should

be avoided because it is likely to be biased. That is, the SPRT sample and total test decisions might agree very well (and they do tend to, by the way), but the decisions may be incorrect.

Since Wald claimed that the SPRT would predict Neyman-Pearson (N-P) decisions, the latter would appear to be a viable method of comparison, as long as measurement error is considered and alpha and beta levels are equivalent respectively in both approaches. For example, if the item pool is very large and if the SPRT alpha is used for the N-P test, the N-P beta will ordinarily be much smaller than the SPRT beta (i.e., the N-P test would be more powerful than the SPRT test). Conversely, if the SPRT beta is used for the N-P test, then the N-P alpha will typically be much smaller.

Double N-P tests. One solution to this problem of non-equivalent alphas and betas would be to perform two Neyman-Pearson tests, where the H_m and H_{nm} are treated, respectively, as null hypotheses and an obtained score is treated as the alternative hypothesis, H.

One test would be:

$[T_1]$ H_m: $p \geq P_m$ vs. H: $p < P_m$

(where the N-P alpha = SPRT beta and N-P beta = SPRT alpha).

The other test would be:

$[T_2]$ H_{nm}: $p \leq P_{nm}$ vs. H: $p > P_{nm}$

(where the N-P alpha = SPRT alpha and N-P beta = SPRT beta).

Unfortunately, the power of these tests of composite hypotheses will vary depending on $\underline{p}$ and could be problematic in rendering valid comparisons of the N-P and SPRT. (However, see below.) If H_m is rejected but $\underline{p}$ is barely in the region of rejection, it is a less powerful test than when $\underline{p}$ is further away from P_m, and similarly for H_{nm}.

Another issue is measurement error. Given the reliability of a test item pool for a group of examinees, a confidence interval can be established around an obtained score (or proportion). For $[T_1]$ to be powerful enough, we should require that the confidence interval around the obtained score lies entirely in the region of rejection of the null hypothesis, H_m, and the confidence interval be established on the N-P beta (e.g., if beta = .025, then use a .95 confidence interval so the right tail of the theoretical sampling distribution for obtained score measurement error for H is beta). Similarly, for $[T_2]$ we should require

that the confidence interval surrounding the obtained score lies entirely in the region of rejection of the null hypothesis, H_{nm}, such that the left tail of the sampling distribution for obtained score measurement error for H is equal to the N-P beta. By requiring the use of the confidence interval around an obtained score, as described here, the power of the statistical test should be thus comparable to that of the SPRT.

There are four possible joint outcomes of $[T_1]$ and $[T_2]$:

		$[T_2]$	
		Reject H_{nm}	Do not reject H_{nm}
$[T_1]$	Reject H_m	NO DECISION	NONMASTERY
	Do not reject H_m	MASTERY	NO DECISION

One of these outcomes may be a little surprising—i.e., when both H_m and H_{nm} are rejected. This will occur when P_m and P_{nm} are far enough apart and the item pool is large enough that the confidence interval for an obtained score somewhere mid-way between P_m and P_{nm} lies in regions of rejection for both $[T_1]$ and $[T_2]$. So we choose neither H_{nm} or H_m.

Mid-point with a confidence interval. As mentioned above, one of the criticisms of the SPRT was that it requires two "cut-off" points, although it was argued that the use of a single cut-off point is prone to misclassifications when obtained scores lie near the cut-off. In other words, when measurement error is considered, the result is a no-decision interval surrounding the single cut-off, which in effect creates an upper and lower bound for mastery and nonmastery decisions in a manner analogous to the SPRT. Therefore, it is intuitively appealing to choose the mid-point between P_{nm} and P_m. Then, if the confidence interval for an obtained score does not include the mid-point and lies above it, a mastery decision would be made; or if below, nonmastery. Otherwise if the confidence interval includes the mid-point, no decision would be rendered.

It should be noted that this method is not as parallel to the SPRT in a statistical sense as is the Neyman-Pearson double test. Nonetheless, the

mid-point has been used in other comparison studies (c.f., Kingsbury & Weiss, 1983) and appears to be consistent with extant conceptions of determining mastery status during criterion-referenced testing.

Mid-point with no confidence interval. This method is similar to the one above, except that no confidence interval is used. Thus, the decision rule is simply to choose which hypothesis an obtained total score is closest to, or make no decision if the obtained score is equal to the mid-point. While the above two methods are preferable to this one, it nonetheless indicates the extent to which SPRT decisions are in the right direction.

Application of the Three Rules for Total Test Decisions

Neyman-Pearson double test. For the DAL test the H_m sampling distribution is 82 out of 96 items correct (for P_m approximately equal to .85). The critical region (left tail) for alpha less than or equal to .025 is 74 or less correct. The standard error of measurement was 3.73; thus, half the .95 confidence interval for an obtained score, assuming a normal distribution of errors, is 1.96 x 3.73 = 7.31. The right tail of this distribution is therefore .025, equal to the SPRT beta chosen a priori. The highest obtained score that has a confidence interval which lies entirely in the rejection of rejection of H_m is 66 ([66 + 7.31] < 74). An alternative method of establishing a confidence interval around an obtained score would be to use the binomial sampling distribution corresponding to that number correct out of 96 and require that .975 of that distribution lie in the region of rejection (c.f., Lord & Novick, 1968, Chapter 11). It turns out that with a relatively large number of items on the test (e.g., 50 or more), obtained scores not near the extremes from a highly reliable test (in the classical sense) will have confidence intervals based on a normal distribution of errors nearly identical to those based on a binomial distribution for that number correct.

For the DAL test the H_{nm} sampling distribution is 58 out of 96 items correct (for P_{nm} approximately equal to .60). The critical region (right tail) for alpha less than or equal to .025 is 67 or more correct. The .95 confidence interval requires a score of 75 or higher so that (75 - 7.31) > 67 and it lies entirely in the region of rejection of H_{nm}.

Therefore, to reject H_{nm} and not reject H_m requires an obtained score of 75 or more to reach a mastery decision; to reject H_m and not reject H_{nm} requires a score of 66 or lower to reach a nonmastery decision; and scores between 67 and 74 inclusively result in no decision.

The standard error of measurement for the 85-item COM test was 3.24. Similarly following the above rules, the mastery region was determined to be 67 or higher, nonmastery 57 or lower, and no decision for scores in the range 58 to 66.

Mid-point with confidence interval. For the DAL test the mid-point between the mastery and nonmastery hypotheses is 70 correct. Scores of 78 or higher have .95 confidence intervals which are above and do not include the mid-point (mastery decisions), scores of 62 or lower resulted in nonmastery decisions, and scores in the range 63 to 77 were classified as no decisions.

For the COM test the mid-point was 61.5. Scores of 68 or higher were classified as mastery, 55 or lower as nonmastery, and 56 through 67 as no decision.

Mid-point with no confidence interval. For the DAL test scores of 71 or higher were classified as mastery, 69 or lower as nonmastery, and 70 as no decision. For the COM test, scores of 61 or lower resulted in nonmastery decisions, and 62 or higher in mastery decisions.

When comparing the Neyman-Pearson double test with the .95 confidence interval rule using the mid-point, it can be seen that the latter creates a slightly wider no-decision interval. It should be noted that the no-decision interval for both these approaches is wider than it would have been had the SPRT itself been applied at the end of the total test. Thus, if the SPRT decisions based on the smaller sample of items were to predict perfectly the SPRT decisions for the total test, the predictions would be less than perfect when compared to the Neyman-Pearson double test or .95 confidence interval decisions, since the no-decision interval is greater for the latter two approaches. The no-decision intervals are nonetheless in the same general areas for all these approaches for the test results in this study.

RESULTS

To address the validity of the SPRT in making mastery classifications when items vary in difficulty levels, contingency tables were constructed for the DAL test and COM test which indicate the agreement between SPRT decisions and those reached by the Neyman-Pearson double test, the mid-point with a .95 confidence interval, and the mid-point without a confidence interval. See Table 1. For example, if the SPRT reached a mastery decision for an examinee and a mastery decision was also reached by the Neyman-Pearson double test, then a tally was entered in the top left cell of that contingency table, etc. Frequencies in the main diagonal of each table indicate agreements, whereas off-diagonal cells indicate disagreements. It should be noted that the expected proportion of agreement is .95. That is, in a large number of cases (assuming about half masters and half nonmasters) we would expect to make classification errors about 2.5 percent of the time for mastery decisions and 2.5 percent for nonmastery decisions.

SPRT vs. Neyman-Pearson Double Test

On the DAL test the SPRT predicted very well (.96), about what would be expected from the established alpha and beta error rates. The two misclassifications were when the SPRT predicted nonmastery, but no decision could be reached by the N-P double test. Note that there were no mastery/nonmastery reversals.

On the COM test the SPRT predicted less well (.88) than on the DAL test, somewhat less than expected. The majority of classification errors were when the SPRT predicted mastery or nonmastery, but the N-P double test resulted in no decisions (12 out of 105 cases). Only one mastery/nonmastery reversal was found. If the results from both tests are combined, the overall agreement is .91, compared to an expected agreement of .95. The average test length required to reach an SPRT decision on either test was about 20 items.

Table 1. Agreement of SPRT Mastery Decisions with Total Test Decisions on Two Different Mastery Tests, where Total Test Decisions are Determined by Three Different Methods: Neyman-Pearson Double Test, Mid-point with a .95 Confidence Interval, and Mid-point with No Confidence Interval. [(P_m = .85, P_{nm} = .60, Alpha = Beta = .025, Expected Agreement = (1 - alpha - beta) = .95)]

DAL Test
(96 items, n = 53, r_{xx} = .977)

		Neyman-Pearson Double Test			Mid-Point (.95 c.i.)			Mid-Point (no c.i.)		
		M	NM	ND	M	NM	ND	M	NM	ND
SPRT	Mastery (M)	23	0	0	18	0	5	23	0	0
	Nonmastery (NM)	0	27	2	0	24	5	1	28	0
	No Decision (ND)	0	0	1	0	0	1	1	0	0
	Percent Agreement		.96			.81			.96	
	Coefficient Kappa		.92			.68			.92	

Mean number of items for SPRT mastery decisions = 19.1 (S.D. = 12.9)
Mean number of items for SPRT nonmastery decisions = 17.4 (S.D. = 16.3)

CCM Test
(85 items, n = 105, r_{xx} = .943)

		Neyman-Pearson Double Test			Mid-Point (.95 c.i.)			Mid-Point (no c.i.)		
		M	NM	ND	M	NM	ND	M	NM	ND
SPRT	Mastery (M)	68	0	8	67	0	9	76	0	0
	Nonmastery (NM)	1	24	4	1	22	6	1	28	0
	No Decision (ND)	0	0	0	0	0	0	0	0	0
	Percent Agreement		.88			.85			.99	
	Coefficient Kappa		.74			.68			.98	

Mean number of items for SPRT mastery decisions = 21.6 (S.D. = 12.6)
Mean number of items for SPRT nonmastery decisions = 18.6 (S.D. = 14.7)

Percent Agreement (both tests)	.91	.84	.98
Coefficient Kappa	.83	.71	.96

SPRT vs. Mid-Point with a .95 Confidence Interval

It can be seen from Table 1 that more disagreements were observed for this comparison on both the DAL and COM test, with agreements of .81 and .85, respectively; and only one reversal was found. The disagreements were SPRT mastery or nonmastery decisions when no decision could be reached with the .95 confidence interval method. Overall agreement on both tests was .84.

SPRT vs. Mid-Point with No Confidence Interval

This comparison indicates the extent to which SPRT predictions are in the right direction. It can be seen that across both tests (158 cases) only three disagreements were observed, two of which were reversals. Overall agreement was .98.

Efficiency of the SPRT

On the average between 20 and 25 percent of the total item pool was required to reach a decision in this study, an approximate savings of 75 to 80 percent over the administration time necessary for the whole pools. Only twice in 158 cases was a reversal of mastery status observed. If we were to flip a coin to predict mastery status (ignoring the no-decision outcome), we would be correct about half the time, assuming no prior information and about the same number of masters and nonmasters in the population of examinees of interest. Given the number of observed agreements between the SPRT mastery decisions and the other methods in this study, the SPRT can be said to improve our decision making accuracy between 68 and 96 percent above our accuracy had we simply guessed mastery status at random, depending on which classification method is used for the total item pools.

Another way of determining efficiency is coefficient kappa (Cohen, 1960). Kappa indicates the proportional reduction of error beyond that expected by chance alone (based on obtained marginal distributions). In other words, it is not necessary to assume that there about half masters and nonmasters. As can be seen in Table 1, kappa's ranged from .68 to .96. Although the proportions of mastery and nonmastery decisions are not split 50-50, the proportional reduction of error is nonetheless about the same as indicated above.

DISCUSSION

Mastery test classifications based on item response theory (IRT) appear to be more accurate than those based on the sequential probability test (SPRT), according to Monte Carlo simulations by Kingsbury and Weiss (1983). On the other hand, the IRT approach is less practical than the SPRT approach. The trade-off therefore seems to be one of practicality vs. accuracy. The SPRT was not compared to the IRT approach in this study because the sample size of examinees was not large enough to obtain reasonably accurate estimates of item parameters, according to recommendations by Hambleton and Cook (1983). The major question addressed in this study was: How well do SPRT decisions predict decisions that are reached on the basis of results from a relatively large and heterogeneous item pool, where item parameters vary considerably?

Results indicated that the SPRT predicts fairly well if it is used conservatively. In this study decision error rates were set at .025, and the mastery and nonmastery levels were chosen on the basis of a typical grading policy. A score of 85 percent or higher is often considered satisfactory for minimal mastery (e.g., comparable to a grade of B or better), whereas a score of 60 percent or lower is considered nonmastery or failing. Probably the most important finding was that, on the two major methods of total test score classifications, only <u>one</u> mastery/nonmastery reversal was observed in 158 cases. In that particular case, the student missed the first four questions randomly administered, resulting in an SPRT nonmastery decision at that point. However, the total test decision for this person was mastery in all three comparison methods. There were <u>no</u> cases where the SPRT predicted mastery, but the total test decision was nonmastery. Depending on which total test classification method was used, the agreement between SPRT decisions and the criterion ranged from .84 to .98 over all cases observed on two different mastery tests, when expected agreement was .95. The average test length for SPRT decisions was about 20 items, though there was considerable variance in SPRT test lengths.

Disagreements tended to occur when the SPRT predicted either mastery or nonmastery, but the total test outcome was no decision. More no-decision disagreements occurred when the classification method for the total test was to determine the mid-point between the mastery and nonmastery levels and then require that the obtained score confidence interval not include the mid-point to render a decision. When no confidence interval is used, SPRT decisions did agree very highly with total test decisions—i.e., almost all SPRT decisions were in the right direction, but some of the obtained scores were not far away enough from one hypothesis or the other in order to reject one of them with sufficient statistical power.

Based on the results of this study, the SPRT appears to be a practical alternative to adaptive mastery testing, where the goal is to render a decision on mastery of a particular educational objective, with as short a test as possible and without sacrificing too much accuracy. It is important to note that these results would be expected only if the SPRT is used rather conservatively. In a true mastery learning context where students have multiple opportunities to retake a test if they have not mastered a particular objective, the consequences of occasional incorrect mastery decisions by the SPRT would seem to be outweighed by the substantial savings in test administration time, particularly when demand for access to computers is high relative to the number of computers or terminals available. The SPRT would also appear to be especially useful for diagnostic testing on a number of objectives (tested one by one, drawing from separate item pools for each objective), since nonmastery decisions tend to be reached very rapidly when a student is clearly ignorant with respect to the knowledge necessary to master a given objective.

Limitations of the Study

As with any study, replications in a variety of contexts with a variety of examinees are needed. It could be that since students were not selected at random, some unkown factor might have affected the results of this study. If similar results obtain in other settings, then it is more likely that the findings are generalizable.

Admittedly, one of the most troublesome parts of this study was to find a method of classifying total test scores in a manner that would render a fair but unbiased comparison with SPRT classifications. Three methods were chosen and they each have their weaknesses. The Neyman-Pearson double test is somewhat novel and was in the opinion of the author the most fair and unbiased method of comparison. One criticism that could be levied is that the same observed score is used to test two different "null" hypotheses. Because the "contrasts" are nonorthogonal, alpha may be inflated. This is analogous to the problem in ANOVA when an F test is significant, where nonorthogonal, multiple contrasts are made.

A further criticism might concern independence of observations. If we believe that this assumption is violated, then we should not be using either the SPRT or the Neyman-Pearson decision model. We would hope, however, that the assumption of local independence would hold (which is also required for IRT); and we try to minimize the problem by selecting test questions at random without replacement, by not giving feedback on correctness of answers during the test, and by not allowing students to change previous answers.

The choice of method of determining confidence intervals for both the Neyman-Pearson double test and the mid-point with the .95 confidence interval might be questioned. A normal distribution of errors was assumed. Thus, z scores were used to form a confidence interval around an obtained score by using the standard error of measurement, which is in turn dependent on the reliability of a test and the standard deviation of the group of examinees studied. Alternative sampling distributions that could have been used are the binomial and beta distributions. However, the central portions of these three distributions are very similar for the number of items in the pools studied, and choosing either of the latter two would most likely not affect the overall results and conclusions of the study.

Perhaps the greatest limitation here is the assumption of the SPRT which is apparently violated when item parameters vary. That criticism was addressed earlier, and a counter-argument was put forth: As long as the probabilities of selecting an item that a master or nonmaster would answer correctly on a given administration of a test remain invariant,

respectively, then the assumption is not really violated. Rather, the danger in using the SPRT is that it may end a test too soon, before enough items representative of the universe have been administered. To minimize this problem, the SPRT should therefore be used conservatively--i.e., with small alpha and beta levels, zones of indifference which are not too broad, and with nonmastery levels that are above a proportion correct that might be obtained by guessing.

Whether or not one accepts the counter-argument, the results from the present study indicate that the SPRT remains fairly robust as a decision model if used conservatively--at least when item pools are not too small and total test reliabilities are high.

Though not a limitation of the SPRT per se, there is a broad philosophical or perhaps attitudinal difficulty in accepting it as a decision model. Most practitioners are accustomed to a single cut-off in making mastery decisions, and may tend to resist the requirement that a zone of indifference must be specified--i.e., both a mastery and nonmastery level. On the other hand, when a single cut-off is used, two composite hypotheses are implied. It is known in statistics that there is no uniformly most powerful and unbiased test of composite hypotheses (c.f., Hays, 1972). Such tests will be less powerful when obtained scores are closer to the cut-off level. For this reason, construction of a confidence interval around an obtained score is often recommended. If this is done, then there will be a range of obtained scores for which no decision can be reached, since their confidence intervals include the cut-off. In effect, a zone of indifference is created which is conceptually not different from that required by the SPRT. However, the SPRT requires the decision maker to specify the zone of indifference a priori, whereas the confidence interval method is typically used a posteriori.

Finally, if test items are poor, then poor decisions will most likely result, regardless of the decision methodology used. Using the SPRT does not excuse us from attempting to develop good test items, perform item analyses when possible, throw out or revise poor items, etc.

Some Unanswered Questions

One question that has been raised is, "Does the predictive validity of the SPRT change as a function of choice of mastery, nonmastery, alpha

and beta levels?" Although the theoretical answers to the question are predictable from the nature of the SPRT decision formulas, it is one which can be empirically tested, and is currently under study. A further question is, "Does the predictive validity of the SPRT change as a function of the degree of heterogeneity of item pools?" This, too, is currently under study.

Another obvious question is, "How do the SPRT and AMT approaches compare empirically?" A future study is planned when enough students are tested to obtain good estimates of item parameters in the IRT model.

A question which may be less obvious concerns the psychological effect that adaptive or shortened tests may have on students—e.g., complaints such as, "This isn't fair—I would have done a lot better if I had taken the whole test. She got to answer 23 questions but I only got to answer 6. She passed and I didn't." It may be that students (and teachers) do not want to use efficient testing methods, even if proven to be generally reliable and accurate, particularly when the consequences of passing or failing are perceived as significant (e.g., course grades, admission to a program, etc.).

As a final comment, the use of the SPRT in mastery testing as described here is intended primarily for making instructional decisions in mastery learning contexts. The SPRT would generally not be a good choice for a decision model for achievement tests where it is important to be able to rank individuals along a continuum with high accuracy.

ACKNOWLEDGEMENTS

This study was supported in part by a grant from the Spencer Foundation.

I would like to thank Dr. Joanne Peng for her thoughtful comments about and criticisms of this study, especially for her suggestion that a double statistical test is comparable to the sequential probability ratio test. Comments on earlier drafts of this paper by Lawson Hughes, Jim Knowlton and Lewis Polsgrove are also appreciated.

Nancy Tyan, Sara Hindman, Sharon Goh and Bob Eckert spent numerous hours assisting in the development and pilot testing of item pools used in this study. Hing-Kwan Luk assisted with item analyses. Their contributions are greatly appreciated.

The STEEL Computer-based Testing System used for test administration and record keeping in this study is currently under development and evaluation at the Center for Innovation in Teaching the Handicapped, School of Education, Indiana University, and is supported in part by a grant from the Office of Special Education, U.S.O.E., Washington, DC. The testing system is written in the Dimension Authoring Language, and currently will run on Digital Equipment Corporation VAX mini- and microcomputers with VT240 or GIGI terminals with color graphics capabilities. Inquiries about the testing system should be addressed to the author.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Ferguson, R. Computer-assisted criterion-referenced measurement. Pittsburgh: University of Pittsburgh Learning Research and Development Center, 1969.

Frick, T. The STEEL computer-based critrion-referenced testing system. (Software version 1.0). Bloomington: Center for Innovation in Teaching the Handicapped (CITH), School of Education, Indiana University, 1985.

Hambleton, R. & Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement. 14(2), 1977, 75-96.

Hambleton, R. & Cook, L. Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), New horizons in testing. New York: Academic Press, 1983, 31-50.

Hays, W. Statistics for the social sciences. New York: Holt, Rinehart & Winston, 1973.

Kingsbury, G. & Weiss, D. A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. Weiss (Ed.), New horizons in testing. New York: Academic Press, 1983, 257-283.

Lord, F. Small n justifies Rasch model. In D. Weiss (Ed.), New horizons in testing. New York: Academic Press, 1983, 52-62.

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

Novick, M. & Lewis, C. Prescribing test length for criterion-referenced measurement, 1. Posttests. (ACT Technical Bulletin No. 18). Iowa City, IA: American College Testing Program, 1974.

McArthur, D. & Chou, C.-P. Interpreting the results of diagnostic testing: Some statistics for testing in real time. Los Angeles: University of California Center for the Study of Evaluation, 1984.

Reckase, M. Some decision procedures for use with tailored testing. Paper presented at the Meeting of Computer-Assisted Testing, Minneapolis, June, 1979.

Reckase, M. The use of the sequential probability ratio test in making grade classifications in conjunction with tailored testing. Columbia: University of Missouri Tailored Testing Lab, 1981.

Reckase, M. A procedure for decision making using tailored testing. In D. Weiss (Ed.), New horizons in testing. New York: Academic Press, 1983, 238-256.

Wald, A. Sequentia' analysis. New York: Wiley, 1947.

Weiss, D. & Kingsbury, G. Application of computerized adaptive testing to educational problems. Journal of Educational Measurement. 21(4), 1984, 361-375.

APPENDIX

Item analyses were performed on two tests: 1) the DA... test--on knowledge of the syntax and structure of the Dimension Authoring Language (n = 53); and 2) the COM Test--on knowledge of how computers functionally work (n = 105). Classical item analyses were first performed. A one-parameter (Rasch) model was also used to estimate item difficulty levels. Two- or three-parameter models were not used due to relatively small sample sizes. In the tables below the following notation is used:

p_{i+} = proportion of examinees who answered item i correctly.

r_{it} = correlation of scores on item i with total test scores.

b_i = difficulty level estimated by the Rasch model for item i.

$S.E._i$ = standard error of estimate of difficulty for item i.

DAL TEST

Item	p_{i+}	r_{it}	b_i	$S.E._i$	Item	p_{i+}	r_{it}	b_i	$S.E._i$
1	.89	.51	-1.89	.49	50	.60	.74	.74	.35
2	.77	.46	- .73	.39	51	.51	.72	1.02	.35
3	.66	.51	.03	.36	52	.60	.71	.41	.35
4	.89	.33	-1.89	.49	53	.58	.53	.53	.35
5	.77	.57	- .79	.39	54	.85	.51	-1.47	.44
6	.57	.41	.65	.35	55	.79	.39	- .95	.40
7	.53	.68	.90	.35	56	.60	.61	.41	.35
8	.64	.62	.16	.36	57	.83	.57	-1.28	.42
9	.42	.61	1.65	.36	58	.77	.47	- .79	.39
10	.70	.34	- .23	.37	59	.62	.48	.29	.36
11	.72	.43	- .36	.37	60	.91	.41	-2.15	.52
12	.79	.65	- .95	.40	61	.68	.62	- .10	.36
13	.91	.50	-2.15	.52	62	.72	.31	- .36	.37
14	.60	.54	.41	.35	63	.68	.63	- .10	.36
15	.42	.72	1.65	.36	64	.66	.77	.03	.36
16	.23	.53	3.10	.41	65	.60	.72	.41	.35
17	.55	.72	.78	.35	66	.91	.49	-2.15	.52
18	.87	.34	-1.67	.46	67	.72	.61	- .36	.37
19	.55	.51	.78	.35	68	.74	.50	- .50	.38
20	.36	.60	2.05	.37	69	.94	.26	-2.81	.65
21	.45	.73	1.40	.36	70	.58	.55	.53	.35
22	.73	.51	- .50	.38	72	.47	.27	1.27	.35
23	.68	.44	- .10	.36	73	.53	.80	.90	.35
24	.66	.75	.03	.36	74	.38	.74	1.91	.37
25	.81	.35	-1.11	.41	75	.55	.69	.78	.35
26	.68	.57	- .10	.36	76	.51	.69	1.03	.35
27	.57	.57	.66	.35	77	.68	.39	- .10	.36
28	.91	.48	-2.15	.52	78	.57	.71	.66	.35
29	.81	.47	-1.11	.41	79	.64	.45	.16	.36
30	.83	.43	-1.28	.42	80	.79	.56	- .95	.40
31	.57	.28	.66	.35	81	.81	.56	-1.11	.41
32	.89	.31	-1.89	.49	82	.47	.50	1.27	.35
33	.81	.35	-1.11	.41	83	.62	.62	.29	.36
34	.68	.32	- .10	.36	84	.79	.23	- .95	.40
35	.81	.44	-1.11	.41	85	.60	.62	.41	.35
36	.91	.41	-2.15	.52	86	.53	.69	.90	.35
37	.45	.65	1.40	.36	87	.40	.67	1.78	.36
38	.72	.49	- .36	.37	88	.40	.70	1.78	.36
39	.45	.47	1.40	.36	89	.51	.70	1.03	.35
40	.85	.56	-1.47	.44	90	.49	.79	1.15	.35
41	.89	.56	-1.90	.49	91	.52	.80	.90	.35
42	.85	.51	-1.47	.44	92	.57	.60	.66	.35
43	.47	.67	1.27	.35	93	.43	.71	1.52	.36
44	.57	.51	.66	.35	94	.55	.50	.78	.35
45	.87	.36	-1.67	.46	95	.66	.52	.03	.36
46	.64	.43	.16	.36	96	.64	.56	.16	.36
47	.70	.73	- .23	.37	97	.55	.46	.78	.35
48	.49	.69	1.15	.35	98	.28	.53	2.62	.39
49	.81	.30	-1.11	.41					

COM TEST

Item	p_{i+}	r_{it}	b_i	$S.E._i$	Item	p_{i+}	r_{it}	b_i	$S.E._i$
1	.65	.53	1.05	.24	44	.92	.31	-1.18	.39
2	.78	.49	.26	.26	45	.89	.43	- .78	.34
3	.98	.30	-2.71	.73	46	.72	.51	.71	.25
4	.87	.38	- .47	.31	47	.78	.35	.33	.26
5	.76	.64	.40	.26	48	.84	.41	- .11	.29
6	.87	.26	- .47	.31	49	.82	.49	- .03	.28
7	.77	.22	.33	.26	50	.69	.68	.88	.24
8	.91	.26	- .90	.36	51	.72	.49	.71	.25
9	.85	.44	- .28	.30	52	.81	.27	.12	.27
10	.74	.58	.52	.25	53	.97	.30	-2.28	.60
11	.89	.49	- .78	.34	54	.73	.47	.59	.25
12	.89	.35	- .78	.34	55	.85	.39	- .28	.30
13	.93	.26	-1.33	.41	56	.81	.33	.12	.27
14	.70	.22	.82	.24	57	.63	.41	1.21	.23
15	.89	.23	- .78	.34	58	.56	.45	1.57	.23
16	.88	.29	- .56	.32	59	.84	.45	- .20	.29
17	.88	.48	- .67	.33	60	.80	.42	.19	.27
18	.85	.52	- .20	.29		.91	.29	.90	.36
19	.87	.59	- .37	.31	62	.94	.28	.33	.41
20	.65	.33	1.10	.23	63	.96	.23	-1.72	.48
21	.79	.10	.19	.27	64	.88	.28	- .56	.32
22	.77	.41	.40	.26	65	.64	.48	1.10	.23
23	.92	.26	-1.03	.37	66	.82	.62	- .03	.28
24	.86	.63	- .28	.30	67	.56	.41	1.62	.23
25	.88	.47	- .56	.32	68	.66	.33	.99	.24
26	.82	.59	- .03	.28	69	.63	.55	1.26	.23
27	.81	.51	.04	.28	70	.51	.56	1.81	.22
28	.93	.57	-1.33	.41	71	.74	.45	.52	.25
29	.50	.39	1.91	.22	72	.73	.31	.58	.25
30	.81	.53	.04	.28	73	.24	.29	3.31	.26
31	.90	.43	- .90	.36	74	.88	.29	- .67	.33
32	.80	.45	.12	.27	75	.91	-.18	- .90	.36
33	.67	.39	.99	.24	76	.79	.57	.26	.26
34	.83	.14	- .11	.29	77	.64	.04	1.10	.23
35	.43	.26	2.26	.23	78	.66	.48	.99	.24
36	.90	.48	- .90	.36	79	.83	.33	- .11	.29
37	.83	.63	- .11	.29	80	.82	.50	.04	.28
38	.81	.69	.12	.27	81	.84	.32	- .20	.29
39	.98	.10	-2.71	.73	82	.75	.28	.46	.26
40	.94	.43	-1.51	.44	83	.50	.39	1.91	.22
41	.89	.28	- .78	.34	84	.84	.21	- .11	.29
42	.92	.50	-1.18	.39	85	.72	.38	.71	.25
43	.87	.33	- .47	.31					