Measurement and Research Department Reports 2008-2

The Sequential Probability Ratio Test in Educational Testing

Theo J.H.M. Eggen

Measurement and Research Department Reports 2008-2

The Sequential Probability Ratio Test in Educational Testing

Theo J.H.M. Eggen

Cito
Arnhem, 2008

Abstract

The sequential probability ratio test (SPRT) is a sequential statistical test developed by Wald (1949). In this chapter a general description of the procedure is given. Then the application of the SPRT in computerized adaptive testing (CAT) is elaborated. Considered is the basic problem of whether a student has mastered a certain criterion or not. Next the extension to the three category problems is described. Finally the problem of adaptive item selection in combination with the SPRT is discussed.

The Sequential Probability Ratio Test in Educational Testing

The sequential probability ratio test (SPRT) was developed more than 60 years ago by Wald (1947) for quality control problems. It is a statistical procedure in which a choice is made between two simple hypotheses and was initially used to determine whether the majority of products (e.g., 80%) in a production process meet specifications or if this was true in fewer cases (e.g., 50%). The statistical properties of the SPRT have since been well established. Various extensions of the original testing procedure have been proposed in the literature and their statistical optimality properties have been studied extensively (see, e.g., Ghosh & Sen, 1991). In this paper we will not go into detail regarding the statistical properties of the SPRT, but will focus on to some successful applications of it in educational testing.

In education, practitioners make use of test results for many different purposes, but from an educational measurement point of view, it generally suffices to distinguish between two main aims of testing: the precise estimation of a person's ability in a certain domain or the classification of a person in one of a limited number of proficiency classes. For the latter purpose, the SPRT has often been very successfully applied. In what is called sequential mastery testing (Lewis & Sheenan, 1990), Ferguson (1969) used a basic application of the SPRT to decide whether a student is a master or non-master in a certain domain. Since then, starting with Reckase (1983), many algorithms for computerized adaptive testing (CAT) have been developed (e.g., Eggen,1999) using the SPRT methodology as their basis.

The present paper discusses the use of SPRT in CAT. After describing the SPRT and the basic elements of CAT, the way the SPRT is used in CAT will be treated. The situation considered is how, on the basis of a test, can be decided whether or not a certain criterion or standard is met. Lastly, this one cutting-point situation is extended to more than one, together with some other extensions.

The sequential probability ratio test

In sequential testing it is not only the observations X, that are random variables, but also the number of observations, K. Inspired by the Neyman-Pearson lemma (1933), which provides a method of constructing a most powerful statistical test for deciding between two simple hypotheses, Wald (1947) proposed the SPRT. In his treatment of the problem, Wald (1947) considered random variables X with two possible values: $x = 1$ if a product meets the criteria and $x = 0$ otherwise. Next, two statistical hypotheses are formulated:

The null hypothesis, $\mathrm{H0}: p = p_0$,and the alternative $\mathrm{H1}: p = p_1$ in which p is the unknown proportion of all products meeting the criteria.

If we denote a series of k observations by $\underline{X}_k = (X_1,....,X_k)$ and the probability distribution of X_i with $\mathrm{P}(X_i = x_i; p) = p^{x_i}(1-p)^{1-x_i}$,

then the probability of these k observations is given by

$P_{0k} = \prod_{i=1}^{k} p_0{}^{x_i}(1-p_0)^{1-x_i}$ if H0 is true and $P_{1k} = \prod_{i=1}^{k} p_1{}^{x_i}(1-p_1)^{1-x_i}$ if H1 is true.

The SPRT then chooses two constants A and B with $A < B$ and, after making every observation, computes the ratio of the probabilities P_{0k} / P_{1k} and the decision is made as follows:

1. if $P_{1k} / P_{0k} \geq A$, then reject H0;

2. if $P_{1k} / P_{0k} \leq B$, then accept H0;

3. if $B < P_{1k} / P_{0k} < A$ (the critical inequality of the procedure) then take another observation.

Intuitively, the procedure is: if the outcomes have much larger probability under H1 than under H0, i.e., the likelihood ratio is large, then reject H0; if the ratio is small, accept H0; and if the ratio has values within the critical interval, no decision is taken and sampling is continued. The constants A and B are dependent on the size of the acceptable decision errors.
In practice, the log likelihood ratio is evaluated. This ratio is equal to the sum over i of the terms

$$Z_i = \ln\left[\frac{p_1^{x_i}(1-p_1)^{1-x_i}}{p_0^{x_i}(1-p_0)^{1-x_i}}\right]. \qquad (1)$$

If the acceptable decision errors are specified by

$$\mathrm{P}(\text{reject H0} \,|\, \text{H0 is true}) \leq \alpha \text{ and } \mathrm{P}(\text{accept H0} \,|\, \text{H1 is true}) \leq \beta \ (\alpha, \beta \text{ small constants}), \qquad (2)$$

then the SPRT procedure is

1. if $\ln B < \sum_{i=1}^{k} Z_i < \ln A$: take another observation

2. if $\sum_{i=1}^{k} Z_i \geq \ln A$: reject H0 (3)

3. if $\sum_{i=1}^{k} Z_i \leq \ln B$: accept H0.

Wald (1947) has shown that the decision error rates are met if

$$A = \frac{1-\beta}{\alpha} \text{ and } B = \frac{\beta}{1-\alpha}. \qquad (4)$$

Furthermore, this SPRT procedure will then stop, with probability 1, with a decision in a finite number of observations.

The SPRT was initially developed for situations in which there is a random sample of a variable with a discrete or continuous distribution with one parameter variable and two simple hypotheses on the value of that parameter. But the theory is generalized in various directions. For our purposes two more general situations are important:

1. Although the SPRT stops with a finite number of observations with probability 1, in educational measurement, it is absolutely necessary to define a maximum test length $k_{\max}$. The procedure is then called the Truncated SPRT (TSPRT).
2. The observations are not random draws from the same distribution, but are independent variables from not necessarily the same distribution.

In the first application by Ferguson (1969), a maximum length was already specified, but the answers on the items were assumed to come from the same binomial distribution, which implies that all items have the same difficulty. In the CAT application, described next, this is not the case.

Computerized adaptive testing

In computerized adaptive tests (CATs), the construction and administration of the test is computerized and individualized. A different test is constructed for every test taker by selecting items from an item bank tailored to the ability of the test taker as demonstrated by the responses given thus far. CATs assume the availability of an item bank, which is calibrated with an item response model. Confining ourselves to item banks with items which are dichotomously scored, logistic item response models are commonly used. In item response theory (IRT), a relation is specified between the non-observable ability θ that is to be estimated and the probability of correctly answering item i. The exact relationship is determined by the parameters of the items. A commonly used IRT model is the two-parameter logistic model (2PL):

$$p_i(\theta) = \mathrm{P}(X_i = 1 \mid \theta) = \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))}, \tag{5}$$

in which b_i is the location or difficulty parameter and a_i the discrimination parameter.

In CATs the parameters of the IRT model are always assumed to be estimated with such precision that they can be considered to be known. In a CAT, the likelihood function of a

student's ability, θ, plays a central role in the inference on the student. Given the scores on k items $x_i, i = 1,...,k$, this function is given by

$$L(\theta, \underline{x}_k) = L(\theta; x_1, ..., x_k) = \prod_{i=1}^{k} p_i(\theta)^{x_i} (1 - p_i(\theta))^{1-x_i}, \quad (6)$$

which states the probability of getting the observed scores on the items as a function of θ.

In CATs where the main aim is the efficient estimation of the ability θ of an examinee, this likelihood function (6) is the basis for estimating the ability of an examinee as well as for the selection of items. The maximum likelihood (ML) estimate of the ability after administering k items follows from the maximization (6) with respect to θ. Because of less bias, the weighted maximum likelihood (WML) method proposed by Warm (1989) is a good alternative for ML. WML follows, in the case of the 2PL model (5) from

$$\hat{\theta} = \max_{\theta} \left[\left(\sum_{i=1}^{k} \mathrm{I}_i(\theta) \right)^{1/2} . L(\theta; \underline{x}_k) \right] . \quad (7)$$

In (7), $\mathrm{I}_i(\theta)$ is the Fisher information function of item i, which is defined as the (statistical) expectation of the squared relative change of the likelihood function

$\mathrm{I}_i(\theta) = E\left(\frac{\partial L(\theta; x_i) / \partial \theta}{L(\theta; x_i)} \right)^2$, which, in the 2PL model is given by

$$\mathrm{I}_i(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta)) = \frac{a_i^2 \exp(a_i(\theta - \beta_i))}{(1 + \exp(a_i(\theta - \beta_i)))^2} .$$

This information function is commonly used for item selection in CATs: an item is selected if it gives maximum information at the current ability estimate. This method ensures that each examinee is administered items which fit his ability and, consequently, his ability is estimated efficiently.

The SPRT in CAT

When classification in one of two categories is the purpose of testing in CAT, SPRT can be applied as follows. On the latent ability scale, a decision or cutting point θ_0 is given which distinguishes between, for example, a master and non-master, or between an examinee who passes and an examinee

who fails an exam. A small region on both sides of this point, a so-called indifference zone, is selected. The widths of these regions are δ_1 and δ_2. The indifference interval expresses the

fact that, owing to measurement errors, making the right decision about examinees very near the cutting point can never be guaranteed. One could also
say that the interval expresses the indifference of an examiner to the classification of the examinees who are very near to the cutting point. Next, the statistical hypotheses are formulated:

H0: $\theta \le \theta_0 - \delta_1 = \theta_1$ against H1: $\theta \ge \theta_0 + \delta_2 = \theta_2$. (8)

If the acceptable decision error rates are specified as in (2), the test meeting these decision error rates uses as the test statistic, as mentioned above, the ratio of the likelihood function under H1 and H0:

$$LR_k(\theta_2;\theta_1) = \frac{L(\theta_2;x_1,...,x_k)}{L(\theta_1;x_1,...,x_k)} \tag{9}$$

and involves the following procedure:

If	Decision	
$\beta/(1-\alpha) < LR_k(\theta_2;\theta_1) < (1-\beta)/\alpha$	administer another item	
$LR_k(\theta_2;\theta_1) \le \beta/(1-\alpha)$	accept H0	(10)
$LR_k(\theta_2;\theta_1) \ge (1-\beta)/\alpha$	Reject H0	

It can easily be shown (Eggen & Straetmans, 2000) that if the 2-PL model (5) is used, the critical inequality of this test can be written as

$$\frac{\ln\left(\frac{\beta}{(1-\alpha)}\right) - C}{\theta_2 - \theta_1} < \sum_{i=1}^{k} a_i x_i < \frac{\ln\left(\frac{(1-\beta)}{\alpha}\right) - C}{\theta_2 - \theta_1} \ . \tag{11}$$

In this

$$C = \sum_{i=1}^{k} \ln\left(\frac{1+\exp(a_i(\theta_1 - \beta_i))}{1+\exp(a_i(\theta_2 - \beta_i))}\right) = \sum_{i=1}^{k} \ln\left(\frac{1-p_i(\theta_2)}{1-p_i(\theta_1)}\right) \tag{12}$$

which depends only on the parameters of the items on θ_1 and θ_2, which are all known constants in the procedure, This makes clear that the application of the SPRT is easy because the observed weighted score is compared to known constants.

Example

The following example of a simulation study with an operational item bank (Eggen, 1999) illustrates the performance of the SPRT in CAT. This item bank contains 250 mathematics items which are used in adult education. Most of the items have an open-ended short answer format, but all the items are scored dichotomously. The items were shown to fit the one-dimensional 2-PL model. The scale was fixed by restrictions on the item parameters. The mean item difficulty is 0, and the geometric mean of the discrimination parameters is 3.09. On this scale, the distribution of the ability in the population was estimated to be normal with a mean of .294 and a standard deviation of .522.

The simulations were conducted as follows. An ability of a simulee was randomly drawn from $N(0.294;0.522)$. Three relatively easy starting items were selected and subsequent items were selected with the criterion of maximum Fisher information at the current ability estimate. The simulee's response to an item was generated according to the IRT model and this procedure was repeated for N =5000 simulees.

For varying acceptable decision error rates and widths of the indifference zone (δ is respectively 0.2 , 0.3 and 0.4 times the standard deviation of θ), the performance of the procedure was evaluated with the mean number of items required to make a decision $\overline{k}$ and the classification accuracy expressed in the percentages of correct decisions (%cor).

The cutting point on the ability scale in the simulations was $\theta_0 = 0.1$, and the maximum test length was $K_{\max} = 25$. The SPRT adaptive testing procedures were conducted for three different error rates and three different widths of the indifference zone.

As a benchmark, the SPRT procedure was compared to a CAT procedure based on statistical estimation. The procedure proposed by Weiss and Kingsbury (1984), but using the Warm estimate of the ability (7), was conducted. After each item is administered, an estimate is made of the examinee's ability $\hat{\theta}_k$ and of its standard error $\mathrm{se}(\hat{\theta}_k)$. Next, a confidence interval $(\hat{\theta}_k - \gamma.\mathrm{se}(\hat{\theta}_k), \hat{\theta}_k + \gamma.\mathrm{se}(\hat{\theta}_k))$ for the examinee's true ability is constructed, in which γ is a constant that is determined by the required accuracy. The procedure delivers another item as long as the cutting point $\theta_0 = 0.1$ is within the interval. If not, the appropriate decision is made. In the comparison, the value for γ was chosen such that about the same accuracy is reached as with the acceptable decision errors in the SPRTs. The results are in Table 1

Table 1: Mean number of required items ($\bar{k}$) of percentage of correct decisions (%cor) in a problem with one cutting point $\theta_0 = 0.1$.

	SPRT						Estimation	
	$\delta = 0.11$		$\delta = 0.16$		$\delta = 0.21$			
Error rate	$\bar{k}$	%cor	$\bar{k}$	%cor	$\bar{k}$	%cor	$\bar{k}$	%cor
$\alpha = \beta = 0.05$	23.24	95.76	15.37	95.16	11.37	95.20	15.41	94.58
$\alpha = \beta = 0.075$	20.35	95.32	13.54	95.02	9.89	94.70	13.40	94.56
$\alpha = \beta = 0.1$	18.72	95.60	12.51	94.90	9.10	93.94	12.97	94.46

The results show that applying the SPRT instead of a traditional estimation procedure possibly improves the performance of the CAT. A striking result is that the chosen acceptable decision error rates and also the chosen width of the indifference zone hardly influence the percentages of correct decisions, but have a major influence on the average number of items needed for taking this decision. It is clear that the number of items needed is larger with lower allowed error rates and with wider indifference zones. The small differences in the percentage correct decisions are always in the expected direction.

Some extensions of the application of SPRT

The SPRT procedure described above can also be applied to other IRT models and used for polytomously scored items (Allen Lau & Wang, 1998). Next, three other extensions of the SPRT application will be addressed.

The SPRT in a three-category classification problem

The Cat application of the SPRT is readily generalized to the case in which there are more than two decision categories. Following Eggen (1999), the generalization to three categories will be described. In this case, there are two cutting points, θ_1 and θ_2, by which three levels of ability are distinguished. An indifference zone is identified around each cutting point:
This is sketched schematically in Figure 1.

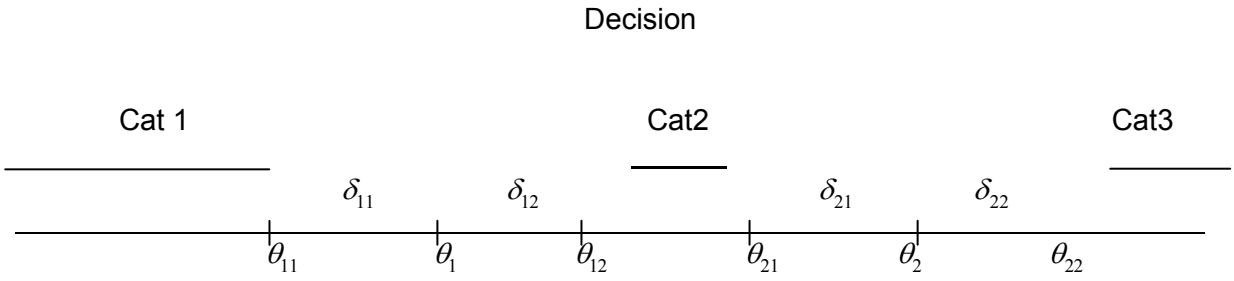

Figure 1. Schematic representation of the classification problem with three categories.

Two pairs of hypotheses are formulated:

H0_1: $\theta \le \theta_1 - \delta_{11} = \theta_{11}$ against H1_1: $\theta \ge \theta_1 + \delta_{12} = \theta_{12}$

H0_2: $\theta \le \theta_2 - \delta_{21} = \theta_{21}$ against H1_2: $\theta \ge \theta_2 + \delta_{22} = \theta_{22}$

The SPRT described in (10) is applied for each pair of hypotheses. The specification of the acceptable decision errors are α_1, β_1 and α_2, β_2, as defined in (2). Next, the two SPRTs are combined in one procedure. The decisions to assign a person to a certain category are given in Table 2 .

Table 2. Decisions based on combination of two SPRTs

	Decision on test 1 H0_1: $\theta \le \theta_{11}$ against H1_1: $\theta \ge \theta_{12}$	
Decision on test 2 H0_2: $\theta \le \theta_{21}$ against H1_2: $\theta \ge \theta_{22}$	Accept H0_1	Reject H0_1
Accept H0_2	Category 1	Category 2
Reject H0_2		Category 3

This combination procedure of the SPRTs originates from Sobel and Wald (1949). It can be shown that, by using the 2-PL IRT model (5) or any other model belonging to the exponential family, the simultaneous acceptance of the null hypothesis H0_1, and the rejection of H0_2 cannot occur. It is noted that Spray (1993) proposed extensions of the use of the SPRT for classification in three and more categories. Her generalization is based on the combination procedure developed by Armitage (1950) which uses the simultaneous application of three SPRTs for classification into three categories instead of only the two needed in the Sobel and Wald (1949) combination procedure proposed here. It is beyond the scope of this paper to discuss in detail the properties of these two combination procedures of SPRTs.

In the practical applications, the combined procedure operates as follows:

If	Decision
$\sum_{i=1}^{k} a_i x_i < \frac{\ln\left(\frac{\beta_1}{(1-\alpha_1)}\right) - C}{\theta_{12} - \theta_{11}}$	Level 1
$\frac{\ln\left(\frac{\beta_2}{(1-\alpha_2)}\right) - C}{\theta_{22} - \theta_{21}} < \sum_{i=1}^{k} a_i x_i < \frac{\ln\left(\frac{(1-\beta_1)}{\alpha_1}\right) - C}{\theta_{12} - \theta_{11}}$	Level 2
$\sum_{i=1}^{k} a_i x_i > \frac{\ln\left(\frac{(1-\beta_2)}{\alpha_2}\right) - C}{\theta_{22} - \theta_{21}}$	Level 3
In all other cases	Continue testing

In this procedure C is as in (12) with the appropriate corresponding constants filled in. From this it easily can be seen that if the width of the indifference intervals, e.g., $\theta_{22} - \theta_{21} = \delta_{12} + \delta_{22}$, increase a shorter test can probably be used to take a decision.

Item selection

An important part of a CAT algorithm is the item selection procedure, which during testing, determines the choice of the items which are administered. In CATs that use the SPRT, item selection is often based on a criterion which is in fact closely related to statistical estimation. Items are selected that maximize the item Fisher information, which means the item will be chosen that minimizes the expected contribution of an item to the standard error of the ability estimate of an examinee. An alternative is to base item selection procedures on the Kullback-Leibler information (Cover & Thomas, 1991). The Kullback-Leibler information expresses the expected contribution of an item to the discriminatory power between two hypotheses and, in that sense, the K-L information fits the statistical testing algorithm more closely conceptually than Fisher information. Eggen (1999) has reported on the comparison of Fisher-based an Kullback Leibler-based information in CATs in combination with the application of the SPRT. In this context, with hypotheses H0: $\theta = \theta_a$ against H1: $\theta = \theta_b$, the K-L information is given by

$$K(\theta_b \| \theta_a) = E \ln \frac{L(\theta_b; \underline{x}_k)}{L(\theta_a; \underline{x}_k)} = \sum_{i=1}^{k} E \frac{L(\theta_b; x_i)}{L(\theta_a; x_i)} = \sum_{i=1}^{k} K_i(\theta_b \| \theta_a) . \qquad (13)$$

It can be seen to be a measure of the expected distance between the two likelihoods of the hypotheses. In the CAT-SPRT example, it consists of the sum of the contribution of all the items in the test and can be seen to be useful item information index. When the K-L test information, (13), is maximized by selecting items having a maximum contribution, the expected difference between the log likelihoods under both hypotheses is maximized. This is the same as making the likelihood ratio more extreme, which is, in turn, expected to minimize the number of items needed to take a decision because the test statistic is the likelihood ratio. In the case of the 2PL model (5), the K-L information is easily computed as

$$K_i(\theta_b \| \theta_a) = a_i(\theta_b - \theta_a)p_i(\theta_b) + \ln\left(\frac{1-p_i(\theta_b)}{1-p_i(\theta_a)}\right).$$

Example (continued)

The performance of the K-L item selection methods will be shown by an example using the same item bank and simulation design described in the above (Eggen (1999)) . For a classification problem in three categories, the cutting points were $\theta_1 = -0.13$ and $\theta_2 = 0.33$ and the maximum test length was $k_{\max} = 25$. In this three-way classification problem, there are more possibilities for K-L item selection. The first is to select the item which maximizes the K-L information at two fixed points. Possible choices are (see Figure 1):

K2a. $\theta_a = \theta_1 + \delta_{12}$ and $\theta_b = \theta_2 - \delta_{21}$

K2b. $\theta_a = \theta_1$ and $\theta_b = \theta_2$ and

K2c. $\theta_a = \theta_1 - \delta_{11}$ and $\theta_b = \theta_2 + \delta_{22}$.

In each case the items will be selected with maximum information to distinguish between two hypotheses. This may cause a problem because a decision in one of three categories is needed. One way to deal with this problem is to look for the nearest cutting point and to select the items with maximum K-L information around this cutting point (K3). The nearest cutting point is determined without estimation by comparison of the score with the midpoints of the critical intervals of the tests. As a benchmark, in the comparison the same method for finding the cutting point is combined with selecting the item maximum Fisher information at this point (F3).

The results of the comparison when the error rates are all 0.05 or 0.1 and with $\delta = 0.13$ are presented in Table 3.

Table 3. Mean number of required items ($\bar{k}$) of percentage of correct decisions (%cor) in a problem with two cutting points $\theta_1 = -0.13$ and $\theta_2 = 0.33$.

	Error rates			
	0.05		0.1	
Selection	$\bar{k}$	%cor	$\bar{k}$	%cor
K2-a	18.7	89.5	16.3	88.1
K2-b	18.4	88.4	16.3	88.6
K2-c	18.7	87.9	16.4	88.6
K-3	16.8	90.1	14.2	89.4
F	16.8	89.6	14.3	88.5

A comparison of the selection methods shows that the differences between them are consistent over the different error rates. Furthermore, it is seen that varying the exact fixed points for which the K-L information is computed has no impact on the performance of the adaptive test. All three methods designed for distinguishing between two points, K2-a, K2-b, and K2-c, perform about the same. But the performance of these “two points” methods is clearly worse than the methods in which during testing first the “best” cutting point is selected and then the item with maximum information. It is seen that, in the latter case, there are no big differences in the performance of the adaptive tests when either the Fisher or Kullback-Leibner information is used.

Stochastic curtailment of the TSPRT

Finkelman (2004, 2008) recently introduced the application of stochastic curtailment to enhance the performance of the Truncuated SPRT in educational testing. The idea of this method is to stop testing sooner without losing accuracy. The method of curtailment determines whether further testing will possibly change a classification decision which would be taken if testing were stopped directly. Stochastic curtailment (Lan, Simon & Halperin, 1982) also extends the observation to the case in which a change in decision is possible but unlikely. In an example, Finkelman (2004) showed that in the case of one cutting point, a reduction of 20% in the number of items needed, while keeping the same accuracy, can be reached by applying stochastic curtailment.

Conclusion

The sequential probability ratio test has been shown to be a very useful statistical procedure. In this paper the applications in the context of educational testing were explored. If classification in a limited number of categories is the main goal of computerized adaptive testing, the (combination of more) SPRT gives very efficient algorithms.

References

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, B, 12*, 137-144.

Allen Lau, C. & Wang T. (1998). *Comparing and Combining Dichotomous and Polytomous Items with SPRT Procedure in Computerized Classification Testing*. Paper AERA, San Diego, April 1998.

Cover, T.M. & Thomas, J.A. (1991). *Elements of information theory*. New York: Wiley.

Eggen, T.J.H.M. (1999). Item selection with adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Eggen, T.J.H.M & Straetmans, G.J.J.M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 66*, 713-734.

Ferguson, R.L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction.* Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh PA.

Finkelman, M . (2004). *Statistical issues in computerized adaptive testing. Unpublished doctoral dissertation,* Standford University, California.

Finkelman, M . (2008). On Using Stochastic Curtailment to Shorten the SPRT in Sequential Mastery Testing. *Journal of Educational and Behavioral Statistics.*

Ghosh, B.K. & Sen, P.K. (1991). *Handbook of Sequential Analysis.* Marcel Dekker, Inc: New York.

Lan, K.K.G., Simon, R., & Halperin, M. (1982). Stochastically Curtailed Tests in Long-Term Clinical Trials. *Communications in Statistics- Sequential Analysis, 1*, 207-219.

Lewis, C. & Sheenan, K. (1990). Using Baysian decion theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 376-386.

Neyman, J & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*, 289-337.

Reckase, M.D. (1983). A procedure for decision making using tailored testing. In: D.J. Weiss (Ed.), *New horizons in testing* (pp. 237-255). New York: Academic Press.

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics, 20*, 502-522.

Spray, J.A. (1993). *Multiple-category classification using a sequential probability ratio test*. (Research report 93-7). Iowa City: American College Testing.

Wald, A. (1947). *Sequential Analysis.* Wiley: New York.

Warm, T.A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika,54,* 427-450.